UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO DA UNIVERSIDADE DE SÃO PAULO (FFCLRP-USP)
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Amr Galal Abdelraheem Ibrahim

Discovery of Transcription Start, Transcription Termination and Transcript Processing Sites in *Halobacterium salinarum* NRC-1 using dRNA-seq

Ribeirão Preto
2023

UNIVERSIDADE DE SÃO PAULO
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo
(FFCLRP-USP)
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Discovery of Transcription Start, Transcription Termination and Transcript Processing Sites in *Halobacterium salinarum* NRC-1 using dRNA-seq

The presented thesis was in: Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo (FFCLRP-USP), with support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Bioinformatics

Candidate: Amr Galal Abdelraheem Ibrahim
Supervisor: Prof Dr Ricardo Z. N. Vêncio
(FFCLRP-USP)

Ribeirão Preto
2023

[a ser elaborada pela biblioteca da Unidade]

Name: Amr Galal Abd El-Raheem Ibrahim

Title: Discovery of Transcription Start, Transcription Termination and Transcript Processing Sites in *Halobacterium salinarum* NRC-1 using dRNA-seq.

The thesis presented for the degree of Doctorate in Sciences
1-Faculty of Philosophy, Sciences and Letters at Ribeirão Preto;
Department of Computation and Mathematics (DCM - FFCLRP)
(Departamento de Computação e Matemática da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Universidade de São Paulo)
2- Institute of Mathematics and Statistics (IME) University of São Paulo
(Instituto de Matemática e Estatística da Universidade de São Paulo)

Approved in: 28th June, 2023.

Examination Committee

Committee Member	**Role**	**Institution**	**Judgment**
Ricardo Zorzetto Nicoliello Vêncio	President	FFCLRP - USP	Doesn't Vote
Angela Kaysel Cruz	Titled	FMRP - USP	**Approved**
André Fujita	Titled	IME - USP	**Approved**
Rodrigo da Silva Galhardo	Titled	ICB - USP	**Approved**

1-FFCLRP - USP:
Faculty of Philosophy, Sciences and Letters at Ribeirão Preto - University of São Paulo
(Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Universidade de São Paulo).

2-FMRP - USP:
Faculty of Medicine at Ribeirão Preto - University of São Paulo.
(Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo).

3-IME - USP:
Institute of Mathematics and Statistics - University of São Paulo
(Instituto de Matemática e Estatística - Universidade de São Paulo).

4- ICB - USP:
Institute of Biomedical Sciences - University of São Paulo.
(Instituto de Ciências Biomédicas da Universidade de São Paulo).

Dedication

I dedicate this thesis project to:

My beloved parents who are motivating me to continue this PhD while they are not in this life anymore; the members of my family who are not longer of this world but their memories still my life. My family, my relatives and, my wife the symbol of love and unlimited giving. All friends who always encourage and support. All colleagues. To everyone who I met and touch my heart. To those are/were part of my daily life.

Acknowledgements

In the Name of ALLAH, the Most Merciful, the Most Compassionate All praise and thanks are for Allah, the lord of all creations ; and Blessings and Peace of Allah be upon Mohammed His servant and last messenger for humanity.

First and foremost, I’m grateful to ALLAH for giving the strength to conduct this work as I could never have achieved the doctorate without His help and bless.

I would like to express my appreciation and my sincere thanks to my supervisor Prof. Dr. Ricardo Z. N. Vêncio who is kindly has contributed by his time and efforts as well as offering the constructive criticism for this work to be done; really I am forever indebted to my supervisor for his support, encouragements and patience all the time during the journey of obtaining the doctorate degree. With a great pleasure I would like to acknowledge prof.(a) Dr.(a) Tie Koide for her helps, assistance and her contribution from the beginning of this project by providing many feedback.

My sincere thanks extended to all the academic staff and the administrative officers (both past and present) at the Bioinformatics Postgraduate Inter-units Program in particular the staff at The Institute of Mathematics and Statistics of the University of São Paulo (IME-USP) and the staff at The Department of Computation and Mathematics (DCM-FFCLRP). My unlimited thanks also extended to the secretariat of the Post-Graduate Program in Bioinformatics -the former staff, and all the assistance staff for their assistances during this doctoral period.

I am extremely grateful for both the past and the present colleagues at the Labora- tory for Biological Information Processing – LabPIB- for their helps, conversations and encouragement. I am extremely grateful for both past and present colleagues at the Microbial Systems Biology Lab -LaBiSisMi- for their helps, conversations and encouragement.

I would like to express my unlimited acknowledge to the Coordination of Improvement of Higher Education Personnel (CAPES), Ministry of Education, Brazil; for fun- ding this project.

I would like to thank all Arabic, forigners, and Brazilian friend a for their conversation and support with special thanks for those are/were part of my daily life.

I thank a lot my dear wife who became a part of my life and helped with a light of hope after long period of struggles and hard time.

I thank everyone who helped me during this doctoral period with criticism, conversations and suggestions. Also, my special thanks extended for those who helped me to grow up as bioinformatician-bioiformaticist. Last but not least, my deepest thanks go to everyone who took part in the journey of making this thesis real. Their names are too numerous to mention, but many of them inspired me to continue learning and sharing with others.

So, many thanks for you all.

Amr Galal Abdelraheem Ibrahkm

Table of Contents

Table of Figures

Index of Tables

Abstract

Differential RNA sequencing (dRNA-seq) has been proven to be a valuable tool for studying bacterial and archaeal genomes by identifying transcription start sites (TSS). With the help of statistical analysis, researchers can analyze dRNA-seq reads from treated and untreated libraries based on the activity of the 5'-monophosphate dependent terminator RNA exonuclease (TEX). Our study focuses on *Halobacterium salinarum* NRC-1, a type of extremophilic archaeon that is commonly found in highly saline environments and is a model organism for studying molecular biology and genetics in extreme environments.

The objective of our study is to identify and map Alternative Transcription Start Sites (alTSS), Transcript Process Sites (TPS), and Transcription Termination Sites (TTS) using the dRNA-seq data in *H. salinarum* NRC-1. We modified the TSSAR tool to detect TSSs from dRNA-seq data, assuming that sequencing reads start at the exact position during transcription and follow a Gamma-Poisson distribution. We annotated alTSSs into four types based on the number of dRNA-seq library reads and differences between two main TSS locations. Alternative TSSs have lower RNA reads than primary ones and can have upstream open reading frames, leading to changes in gene regulation output, 5'UTR isoform, and gene transcription pausing. Mapping alTSSs allowed us to explain changes in cell response to growth conditions and gene expression across different growth stages.

Our findings revealed a significant number of falsely annotated internal transcription start sites (iTSSs) that were redefined as alternative TSSs in previous genome annotations. These alternative TSSs produced different protein isoforms depending on the length of the amino acid chain and the open reading frame. Additionally, alternative TSSs were identified not only in *H. salinarum* but also in other organisms, suggesting a crucial role in regulating gene expression across various species.

Furthermore, we conducted a re-analysis of dRNA-seq data, focusing on non-primary transcripts (monophosphorylated RNAs) instead of the traditional method of enriching for primary transcripts (triphosphorylated RNAs) to identify genome-wide transcript processing sites (TPS) in *H. salinarum* NRC-1. We also applied this approach to *Haloferax volcanii* for comparative analysis.

Lastly, we used dRNA-seq data to identify hairpin structures and mapped them onto the genome, providing insights into the potential role of transcription termination sites (TTS).

Keywords: Alternative TSS; dRNA-seq; *Halobacterium salinarum* NRC-1; UTRs; Caloi-seq; Transcription Processing sites; RNA processing sites; post-transcriptional regulation; gene expression; TTS.

Resumo

O sequenciamento diferencial de RNA (dRNA-seq) tem se mostrado uma ferramenta valiosa para o estudo de genomas bacterianos e arqueanos por meio da identificação de sítios de início de transcrição (TSS, do inglês Transcription Start Sites). Com o auxílio da análise estatística, os pesquisadores podem analisar as leituras de dRNA-seq de bibliotecas tratadas e não tratadas com base na atividade da exonucleasse de RNA terminador dependente de 5'-monofosfato (TEX). Nosso estudo se concentra em *Halobacterium salinarum* NRC-1, um tipo de arqueia extremófila comumente encontrada em ambientes altamente salinos e é um organismo modelo para estudar a biologia molecular e genética em ambientes extremos.

O objetivo de nosso estudo é identificar e mapear sítios alternativos de início de transcrição (alTSS, do inglês Alternative Transcription Start Sites), sítios de processamento de transcrição (TPS, do inglês Transcript Proces Sites) e sítios de terminação de transcrição (TTS, do inglês Transcription Termination Sites) usando os dados dRNA-seq em *H. salinarum* NRC-1. Modificamos a ferramenta TSSAR para detectar TSSs a partir de dados dRNA-seq, assumindo que as leituras de sequenciamento começam na posição exata durante a transcrição e seguem uma distribuição Gamma-Poisson. Anotamos os alTSSs em quatro tipos com base no número de leituras da biblioteca dRNA-seq e diferenças entre duas localizações principais de TSS. Os TSSs alternativos possuem leituras de RNA mais baixas do que as primárias e podem ter fase de leitura aberta upstream, levando a mudanças na produção de regulação gênica, isoformas 5'UTR e pausas na transcrição gênica. O mapeamento de alTSSs nos permitiu explicar as mudanças na resposta celular às condições de crescimento e expressão gênica em diferentes estágios de crescimento.

Nossas descobertas revelaram um número significativo de sítios de início de transcrição internos (iTSSs, do inglês Internal Transcription Start Sites) erroneamente anotados que foram redefinidos como TSSs alternativos nas anotações genômicas anteriores. Esses TSSs alternativos produzem diferentes isoformas de proteína dependendo do comprimento da cadeia de aminoácidos e do quadro de leitura aberto. Além disso, foram identificados TSSs alternativos não apenas em *H. salinarum,* mas também em outros organismos, sugerindo um papel crucial na regulação da expressão gênica em várias espécies.

Ademais, executamos uma reanálise dos dados de dRNA-seq, com ênfase em transcritos não primários (RNAs monofosforilados) ao invés do método tradicional de enriquecimento em transcritos primários (RNAs trifosforilados) para identificar sítios de processamento de transcrição (TPS, do inglês Transcript Proces Sites em todo o genoma em *H. salinarum* NRC-1. Também aplicamos essa abordagem ao *Haloferax volcanii* para análise comparativa.

Por fim, utilizamos dados de dRNA-seq para identificar estruturas em *hairpin* e mapeá-las no genoma, fornecendo informações sobre o potencial papel dos locais de terminação da transcrição (TTS).

Palavras-chave: TSS alternativo; dRNA-seq; *Halobacterium salinarum* NRC-1; UTRs; Caloi-seq; Sítios de Processamento da Transcrição; Sítios de Processamento de RNA; regulação pós-transcricional; expressão gênica; TTS.

Introduction

1. Introduction

1.1 Background

For a long while, living organisms were divided into two kingdoms: Animalia and Plantae. During the nineteenth century, new classification started to add new kingdoms: Bacteria, Protista, Fungi, Plantae, and Animalia. Nowadays the living organisms are divided into two main divisions: the Prokaryote and the Eukaryote (Woese et al., 1990). The Prokaryotes are currently divided into two domains, Bacteria and Archaea, as totally different from one another as either is from the Eukaryotes (Castelle & Banfield, 2018). None of those seems to be ancestral to the other and each shares features with the other yet as having distinctive characteristics of its own.

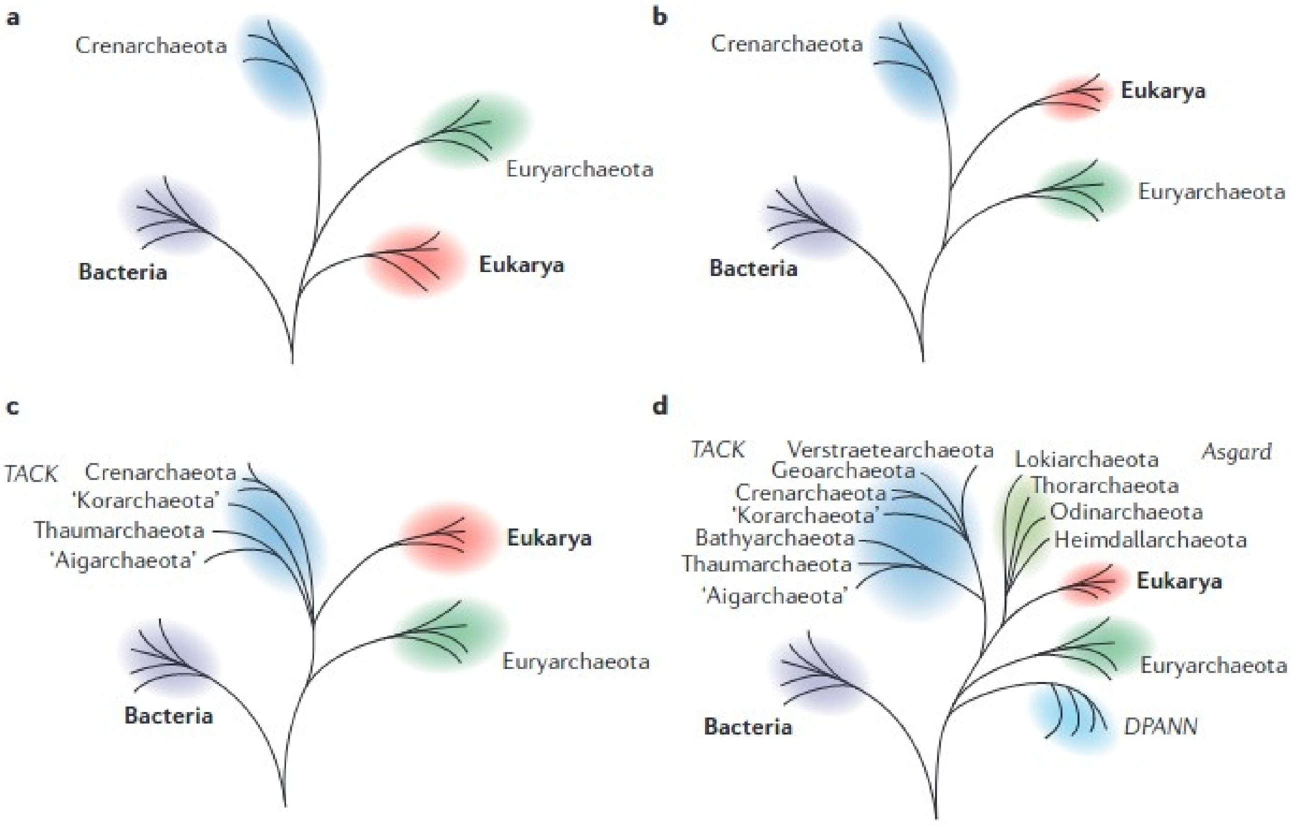


Figure (1-1): *Schematic representation showing the relationships between eukaryotes and archaea understood at different historical times, from the classical model (a) to the most recently proposed model (d). Schematic shows the bacteria and eukaryotes in light purple and red, whereas green and blue both represent archaeal lineages. (a) In the three-domain tree of life model there are two sub-domain branches that make the archaea domain. (b) Several studies illustrated molecular relationships between eukaryotes and Crenarchaeota. (c) Eukaryotes were found to branch within, or as sister to TACK superphylum: Thaumarchaeota (now Nitrososphaerota), Aigarchaeota, Crenarchaeota (now Thermoproteota), and Korarchaeota. (d) Involving members phylogenetic analysis of the Asgard superphylum strongly suggested that eukaryotes had their ancestry or sister groups from Asgard archaea. Rank-level names for domains and are bolded. Proposed phyla names are enclosed in quotation marks. The DPANN group is named after: Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota and are represented as a monophyletic lineage. [Taken from (Eme et al., 2017) , fig 1].*

Starting during the nineties, many additional works have been done to resolve connections inside the Eukaryote. Moreover, an intense debate is ongoing recently suggesting that a two domain may be better suited to accommodate recent findings and revised bioinformatics analysis, shown in Figure (1-1). In this PhD work we take a conservative position and still consider a classical three domain model. Archaea was discovered in the late seventies (Woese, 1994). The studies of Prokaryotes showed different DNA sequences, mainly those related to ribosomes, and found that there have been two clearly completely different groups in their traits.

Archaeal transcription is more commonly summarized as a simplified version of the eukaryotic machinery (Gehring et al., 2016). A summary of major differentiating features among Bacteria, Archaea, and Eukarya can be seen in Table (1-1).

Table (1-1): Comparison summary of major features among the three domains: Bacteria, Archaea, and Eukarya. Note that for many features only particular representatives within a domain show the property. The table is taken from (Pun et al. 2005 - Table 2).

Characteristic	Bacteria	Archaea	Eukarya
Prokaryotic cell structure	Yes	Yes	No
DNA present in covalently closed and circular form	Yes	Yes	No
Histone proteins present	No	Yes	Yes
Membrane-enclosed nucleus	Absent	Absent	Present
Cell wall (Muramic acid)	Present	Absent	Absent
Membrane lipids	Ester-linked	Ether-linked	Ester-linked
Ribosomes	70S	70S	80S
Initiator tRNA	Formylmethionine	Methionine	Methionine
Introns in most genes	No	No	Yes
Operons	Yes	Yes	No
Capping and poly-A tailing of mRNA	No	No	Yes
Plasmids	Yes	Yes	Rare
Ribosome sensitivity to diphtheria toxin	One (4 subunits)	Several (8-12 subunits each)	Three (12-14 subunits each)
RNA polymerases	No	Yes	Yes
Transcription factors required	No	Yes	Yes
Promoter structure	-10 and -35 sequences	TATA box	TATA box
Sensitivity to chloramphenicol, streptomycin, and kanamycin	Yes	No	No
Methanogenesis	No	Yes	No
Reduction of S^0 to H_2S or Fe^{3+} Fe^{2+}	Yes	Yes	No
Nitrification	Yes	No	No
Denitrification	Yes	Yes	No
Nitrogen fixation	Yes	Yes	No
Chlorophyll-based photosynthesis	Yes	No	Yes (in chloroplasts)
Chemolithotrophy (Fe, S, H_2)	Yes	Yes	No
Gas vesicles	Yes	Yes	No
Synthesis of carbon storage granules composed of poly-β-hydroxyalkanoates	Yes	Yes	No
Growth above 80°C	Yes	Yes	No

Many types of bacteria can live in high temperature conditions, but still are different from archaea in genetic structure (Woese, 1994). Archaea could be distinguished from Bacteria by analysis of the chemical composition of their cell wall since the archaeal cell wall does not contain peptidoglycan. Archaea aren't restricted to extreme environments: techniques such as environmental genome-shotgun sequencing have indicated that Archaea are also found in the open ocean for example (Allers & Mevarech, 2005). It is clear that Archaea is a very important biological group of organisms for different studies but they have been highlighted by their extremophile reputation since they are capable of surviving in various extreme growth environments (Whitehead et al., 2009; Williams et al. 2017).

Some of these species are tolerant to high salt concentration. For example, species belonging to the genus Halobacterium provide insights to the study of mechanisms that address the molecular response and adaptation to survive in extreme environments. The extremophile archaea used in this work as model organism is the ***Halobacterium salinarum***.

Halobacterium is a genus that belongs to the Archaean domain which consists of several species; phylum Euryarchaeota; class Halobacteria; order Halobacteriales; family Halobacteriaceae and genus Halobacterium (Allers & Mevarech, 2005; DasSarma & DasSarma, 2008). The particular strain *Halobacterium salinarum* NRC-1 has been used intensively as a model organism in the field of system biology (DasSarma et al., 2006). Its phylogenetic context is given in Figure (1-2).

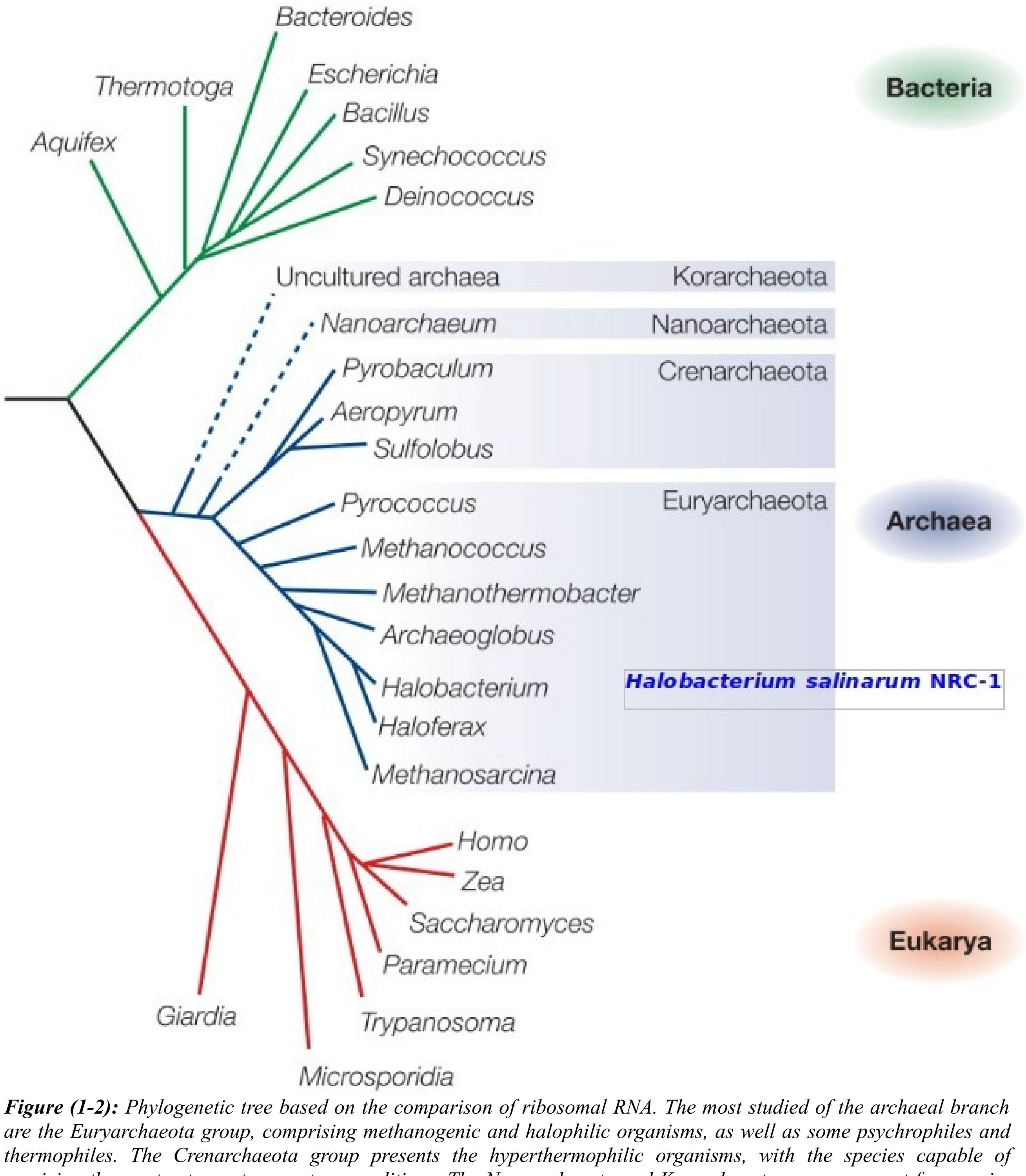


Figure (1-2): *Phylogenetic tree based on the comparison of ribosomal RNA. The most studied of the archaeal branch are the Euryarchaeota group, comprising methanogenic and halophilic organisms, as well as some psychrophiles and thermophiles. The Crenarchaeota group presents the hyperthermophilic organisms, with the species capable of surviving the most extreme temperature conditions. The Nanoarchaeota and Korarchaeota groups present few species and the positioning of their phyla in the phylogenetic tree is still uncertain (dashed lines). The organism most studied in this PhD dissertation, Halobacterium salinarum strain NRC-1, is highlighted in the Halobacterium branch. [Modified from (Allers & Mevarech, 2005), Box1].*

Halobacterium salinarum NRC-1 is a photosynthesizing archaeon that relies on neither chlorophyll nor bacteriochlorophyll (DasSarma et al., 2006). It shows no turgor pressure and uses the "salt-in" strategy to achieve osmotic balance. *Halobacterium salinarum* is an obligate halophile, shown in Figure (1-3), which grows optimally around 4.3 mol/l NaCl concentrations and gets lysed at low salt concentrations. Its intracellular concentrations of K$^+$ and Na$^+$ were measured approximately equal to 4 mol/l and 1.4 mol/l, respectively, with Cl$^-$ just 10% higher than the growth medium (Ingoldsby et al., 2005). Biological and biotechnological achievements were undertaken using *Halobacterium salinarum* as a model, from the structural elucidation of bacteriorhodopsin (Henderson et al., 1990) to vaccine improvement (DasSarma et al., 2014).

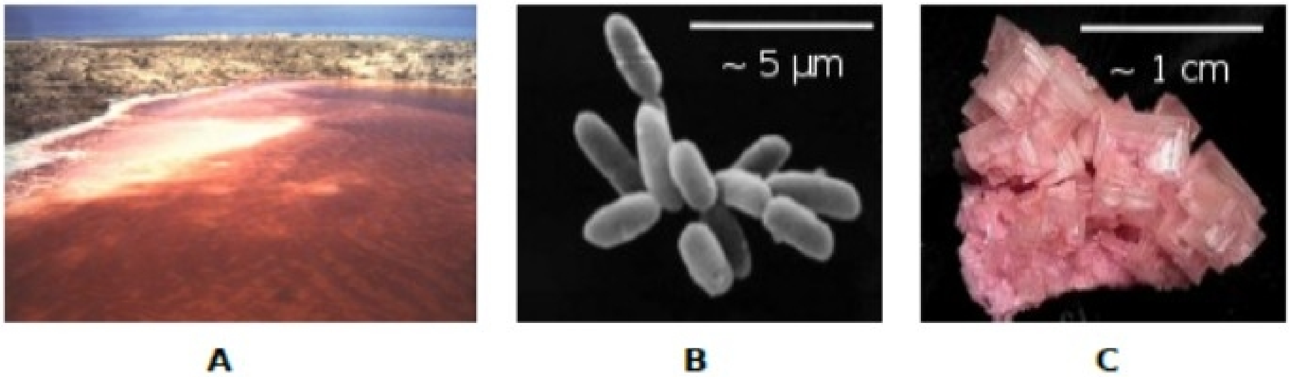


Figure (1-3): *The archeal organism H. salinarum NRC-1. (A) Salt lake with Haloarchaea. The red color is due to gas vesicles, the Bacterioruberin biological pigment and the Bacteriorhodopsin protein production present in the membrane of Halobacteria and responsible for turning sunlight into chemical energy. (B) Electron microscopy of H. salinarum NRC-1. (C) Halite containing H. salinarum, found at Lake Searles, California. Taken from ((DasSarma et al., 2010)& https://microbewiki.kenyon.edu/index.php/Halobacterium; http://www.webmineral.com/data/Halite.shtml).*

The sequenced genome of *Halobacterium salinarum* NRC-1 consists of 2,571,010 bp (base pairs) and contains more than 91 insertion sequences. These sequences are organized into a large circular chromosome (2,014,239 bp) and 2 related plasmids pNRC100 (191,346 bp) and pNRC200 (365,425 bp) (Ng et al., 2000). The two plasmids contain some essential genes so sometimes they are also referred to as minichromosomes.

Halobacterium salinarum NRC-1 has many origins of replications for the three replicons, like the Eukaryotes, while the genomic replicons are circular (Coker et al., 2009). Although replication starts from both oriC1 and oriC2, and the two origins of replication are dividing the whole genome into four replichores, the reference genome sequence starts at the sequence near oriC1 which is more conserved (Makarova et al., 2015; Zhang & Zhang, 2003).

Genome annotation in *Halobacterium salinarum* NRC-1 had different steps: genome sequencing (Ng et al., 2000), transcription factor binding sites (TFBS), gene expression data (Koide et al., 2009), using the RefSeq data on the GenBank NCBI (Pruitt et al., 2007), Archaeal genome browser on UCSC (Chan et al., 2012), and using Halolex portal to get proteome information (Pfeiffer et al., 2008). Then, using Gaggle Genome Browsers (Bare et al., 2010) and EGRINs (Environment And Gene Regulatory Influence Network) (Brooks et al., 2014), new functional annotations were added. Meanwhile, NCBI is continuously developing its RefSeq pipeline for prokaryotic data (Haft et al., 2018; O'Leary et al., 2016). The common feature of all annotations of *Halobacterium salinarum* NRC-1 efforts were based on the available annotation data of NRC-1's closely related strain *Halobacterium salinarum* R1 (Pfeiffer et al., 2008). There is a vast increase in comparative genomic data of Halobacterium species in recent years (139 genomes) due to the lower cost of next generation sequencing and readily available tools for assembly and annotation (Gaba et al., 2020).

Concerning archaea, our own research group also contributed to the literature discovering novel RNAs: transcripts were identified near the 3' end of transposases (Gomes-Filho et al., 2015); intragenic RNAs (intraRNAs) were identified by sequencing transcripts interacting in some organisms with the Hfq RNA chaperone (Lorenzetti et al., 2023). The different types of RNAs are produced due to different positions of promoters and RNA processing during or after transcription, as depicted in Figure (1-4) reproduced from (Ten-Caten et al., 2018).

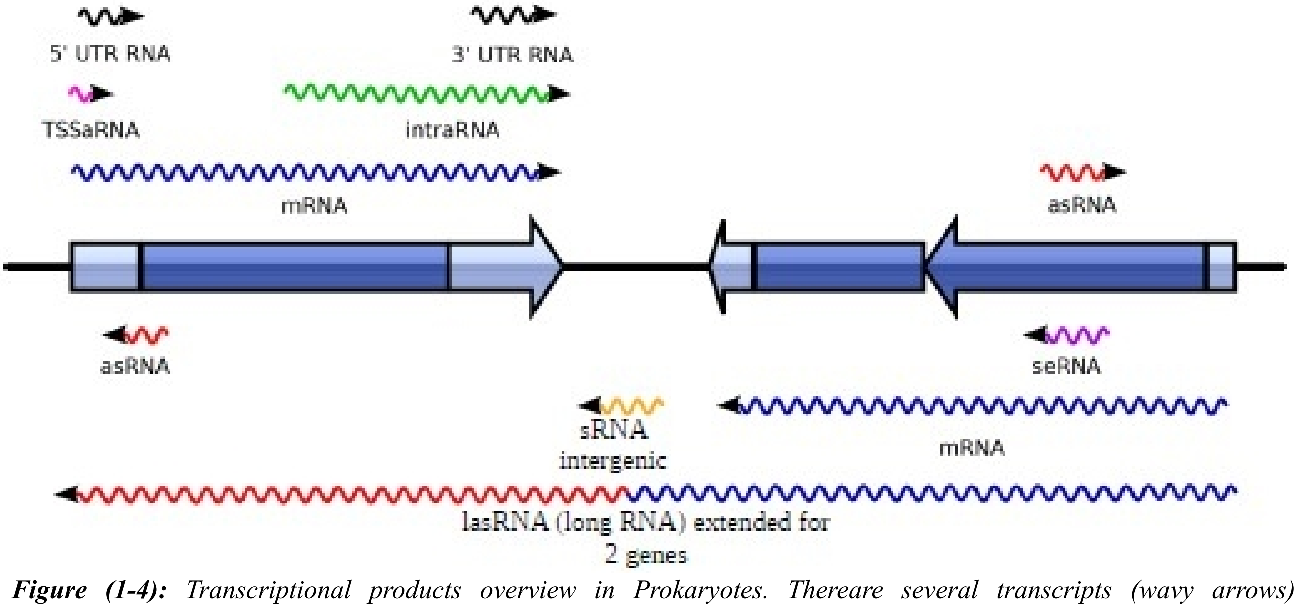


Figure (1-4): *Transcriptional products overview in Prokaryotes. Thereare several transcripts (wavy arrows) superimposed on genes (blue arrows) in the genome of Prokaryotes. The light blue region in the genes represents the 5' and 3'UTRs. TSSaRNA, transcripts associated with early transcription; intraRNA (green), intragenic RNAs (yellow); asRNA (red), antisense transcripts; seRNA (violet), sense transcripts; lasRNAs, long asRNAs (blue and red linked), which have a coding part and a non-coding part capable of regulating the mRNA on the opposite strand. The direction of the arrows indicates the orientation of genes and transcripts. Taken from doctoral thesis (Ten-Caten, 2017).*

RNA transcription is conceptually divided in three stages: 1) Initiation, 2) Elongation, and 3) Termination, as shown in Figure (1-5). The following base after the promoter is called the **Transcription Start Site (TSS)** where the transcription begins and from where on the DNA information is carried in a RNA molecule. The **Transcript Termination Sites (TTS)** are more elusive since are a non necessarily precise region in the DNA where the information to terminate the transcription is kept in an indirect way since it is the actual RNA folding/sequence that defines it.

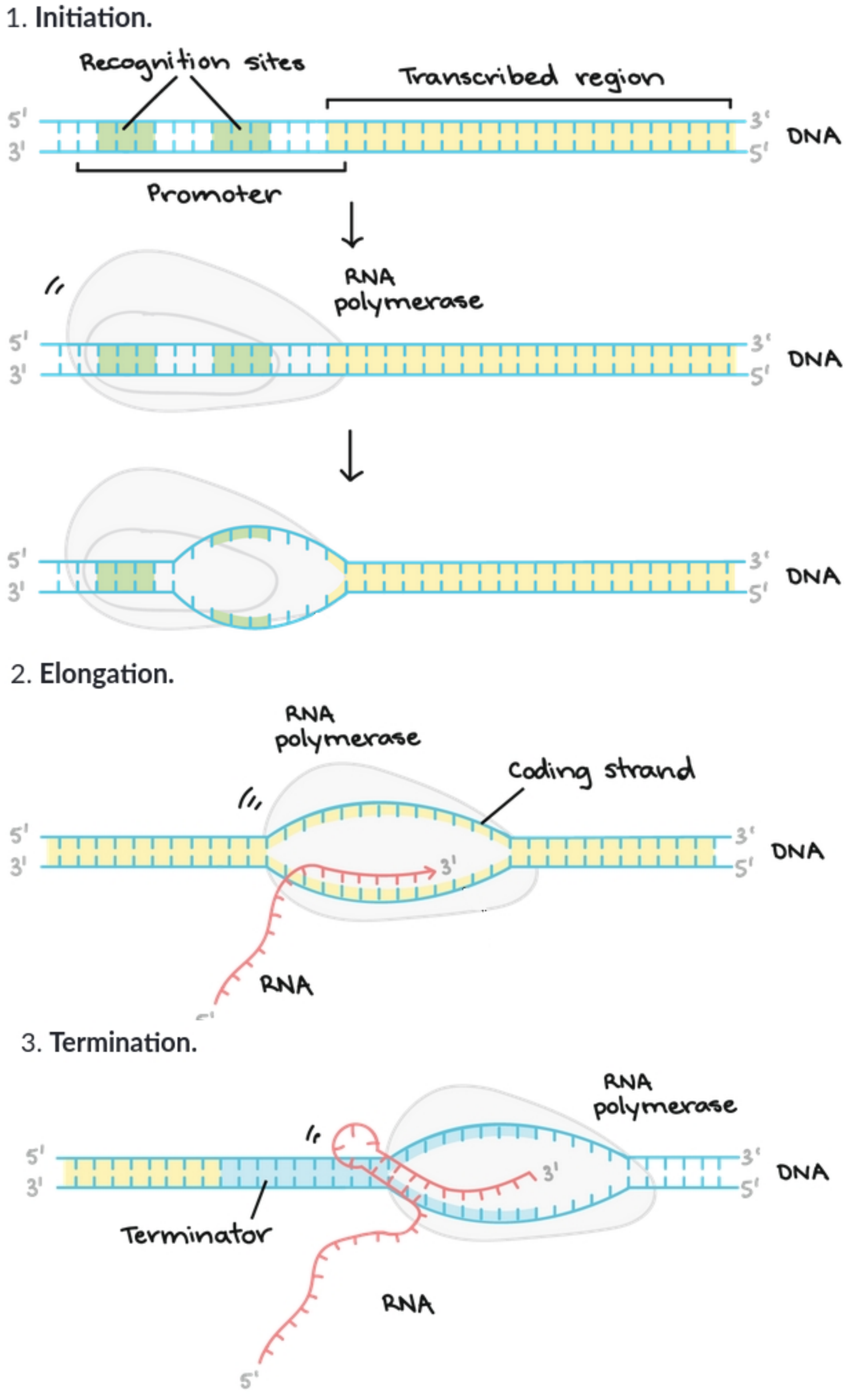


Figure (1-5): *Simplified overview of RNA transcription process. A gene's transcription occurs in three stages: initiation, elongation, and termination. (1) Near the start of a gene, RNA polymerase binds to a DNA region known as the promoter. (2) RNA polymerase uses one DNA strand, known as the complementary sequence, as a template. (3)*

Terminator sequences indicate that the RNA transcript is finished. Once they have been transcribed, they cause the RNA polymerase to release the transcript. Adapted from (https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription).

Genome annotation in *Halobacterium salinarum* NRC-1 had many revisions and improvements since when started by genome sequencing (Ng et al., 2000). Many microarray experiments and transcription factor binding sites (TFBS) mapping efforts improved the localization of TSS. It was also possible to identify operons which work as operons in certain conditions, Figure (1-6), but have internal TSS, thus "breaking" the operon, in other conditions (Koide et al., 2009).

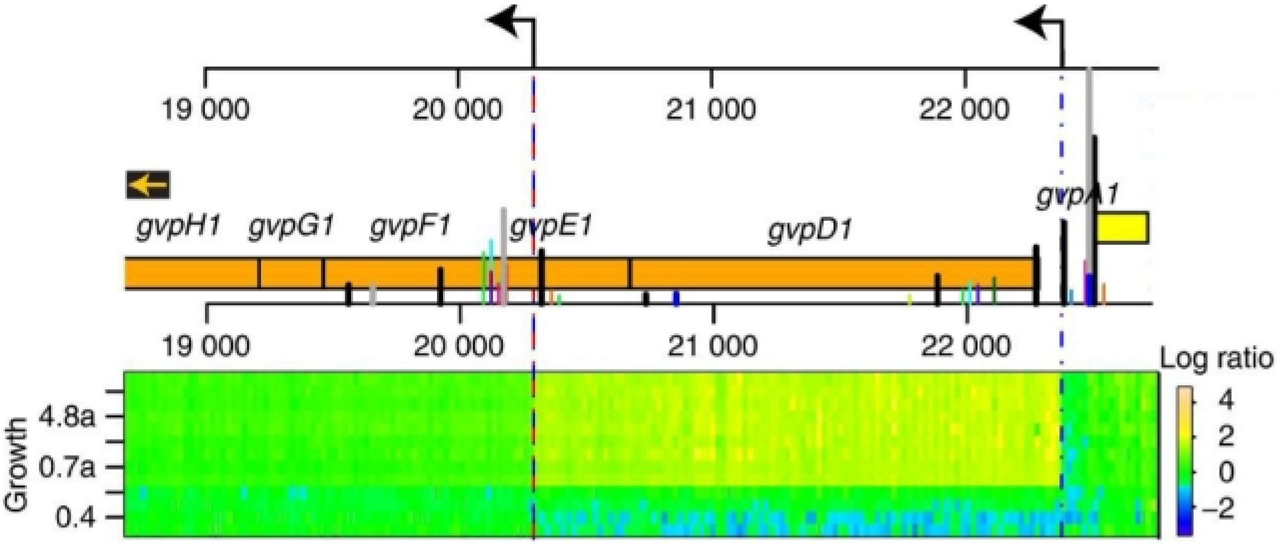


Figure (1-6): *Example of H. salinarum operon (orange boxes, reverse strand, 5' to 3' direction is right to left) that can, depending on the environmental condition (growth curve moment) express independently "breaking" the operon. The TSS are indicated as black arrows. Taken from (Koide et al., 2009).*

The 5' untranslated region (5'UTR) is the region in mRNA that is located directly upstream to the start codon. Genomically, goes from the TSS to the first base of the start codon. This region plays an important role in the transcription regulation by different mechanisms in viruses, Prokaryotes, and Eukaryotes. For example, the 5'UTR may contain a sequence where the ribosome binding starts the translation for mRNA which is called the ribosomal binding site (RBS) and it is known in bacteria as the Shine Dalgarno sequence (SD sequence). This sequence is complementary to the 3'-end in the small ribosomal subunit of rRNA. SD sequence is different from organism to another and for *Halobacterium salinarum* is: **GGAGGUCA** (Pfeifer, 2015). However, this SD sequence could be unnecessary for translation initiation in 5'UTRs (Kramer et al., 2014; Pfeifer, 2015) or may use some yet unknown sequence motif.

In this PhD thesis, we examined and identified the start and termination sites (TSS and TTS, respectively) of the *Halobacterium salinarum* NRC-1 genome. To do this we studied differential RNA sequencing (dRNA-seq) data as will be explained in the following sub-section.

Another important genomic position, or site, that is explored in this PhD work which is different in nature from the previous mentioned two (TSS and TTS) is the **Transcript Processing Site (TPS)**. Although always represented as a coordinate of the (or mapped into the) genome, this site is a position on a RNA molecule. A TPS is a proxy of a blunt cleavage or a gradual degrading process that happens on RNA molecules and it is detected by dRNA-seq, but is represented in its respective genomic (thus DNA) location. It is a RNA focused phenomena that is represented in a DNA-based genomic coordinate (or site).

RNA processing is a complex process involving various RNA molecules and enzymes that work together to regulate gene expression. The expression levels of RNAs are determined by their rates of transcription and degradation, which are precisely monitored by a set of enzymes responsible for RNA maturation and degradation (Arraiano et al., 2010). Ribonucleases (RNases) are a group of enzymes that catalyze the exo- or endo-nucleolytic disruption of RNA phosphodiester bonds. The majority of known RNases are protein enzymes, but in some cases, the catalytic part is an RNA molecule (dos Santos et al., 2018)

Endoribonucleases (endo-RNases) and exoribonucleases (exo-RNases) are the two main classes of RNases responsible for initiating RNA maturation and degradation, playing crucial roles in post-transcriptional regulation. The coordination between both systems results in the precise processing of RNA substrates (Linder & Jankowsky, 2011). There is a significant overlap between RNases involved in RNA degradation and those involved in maturation, indicating the importance of fine-tuning RNA processing. However, differences exist between those that carry out RNA degradation or maturation. Currently, about fifteen superfamilies of ribonucleases are known, which can be grouped according to their catalytic activities, providing insight into the complex mechanisms of RNA processingp (Perwez & Kushner, 2006).

These findings provide insight into the complex mechanisms of RNA processing and highlight the importance of RNases in regulating gene expression (Arraiano et al., 2010).

1.2 Differential RNA sequencing (dRNA-seq)

Large-scale sequencing technologies became one of the best tools for distinguishing and characterizing genomes and their expression at the transcriptional level. Such technologies are principally aimed toward increasing information acquisition capability and reducing the prices concerned during this acquisition. Among these relatively new techniques, RNA sequencing has gained appreciable importance in the field study of transcriptomics (Mortazavi et al., 2008). This system relies on RNA-seq sequencing methodologies developed in recent decades, and represents a model for sequence breaking, permitting the generation of a massive quantity of information. Capable of performance-enhancing a brand new generation of sequencing, usually knows as "next-generation" (in spite of the obvious flaw that there are always be a next technological generation relative to the current one) is minimizing the manual bench-based intervention steps and therefore showing high ability of sequencing parallelization (Metzker, 2010).

Some studies published in an attempt to focus sequencing efforts on specific groups of RNAs, which may lead to a decrease in signal complexity and facilitate processing steps. An example is a study published by (Sharma et al., 2010) where the researchers were able to globally map the transcription starts TSS of *Helicobacter pylori*, an important human pathogen, by comparing a library enriched with primary RNAs with a control library. Enrichment of the sample with primary RNAs is performed with the enzyme Terminator 5 'Phosphate-Dependent Exonuclease (TEX). This enzyme is capable of degrading all mRNAs at the 5' end, RNAs have undergone some processing step in that region and are not triphosphate anymore. Sequencing a library of primary RNAs and a control library the researchers were able to determine the transcriptional beginnings by identifying the positions with a higher number of counts in the library enriched for primary RNAs when compared to the control library (Sharma, Hoffmann, Darfeuille, Reignier, Findeiß, et al., 2010; Takahashi et al., 2012; Ten-Caten et al., 2018).

This method was originally named **dRNA-seq**, standing for "differential RNA sequencing" and was then widely used to a number of prokaryote organisms after its invention. This method is central in the present PhD work. Different information can be gained from dRNA-seq data as shown in Figure (1-7). Some of this information has many studies and others still very less studies (Sharma & Vogel, 2014). The large-scale identification of transcription starts is an important tool to obtain a better characterization of pervasive transcription.

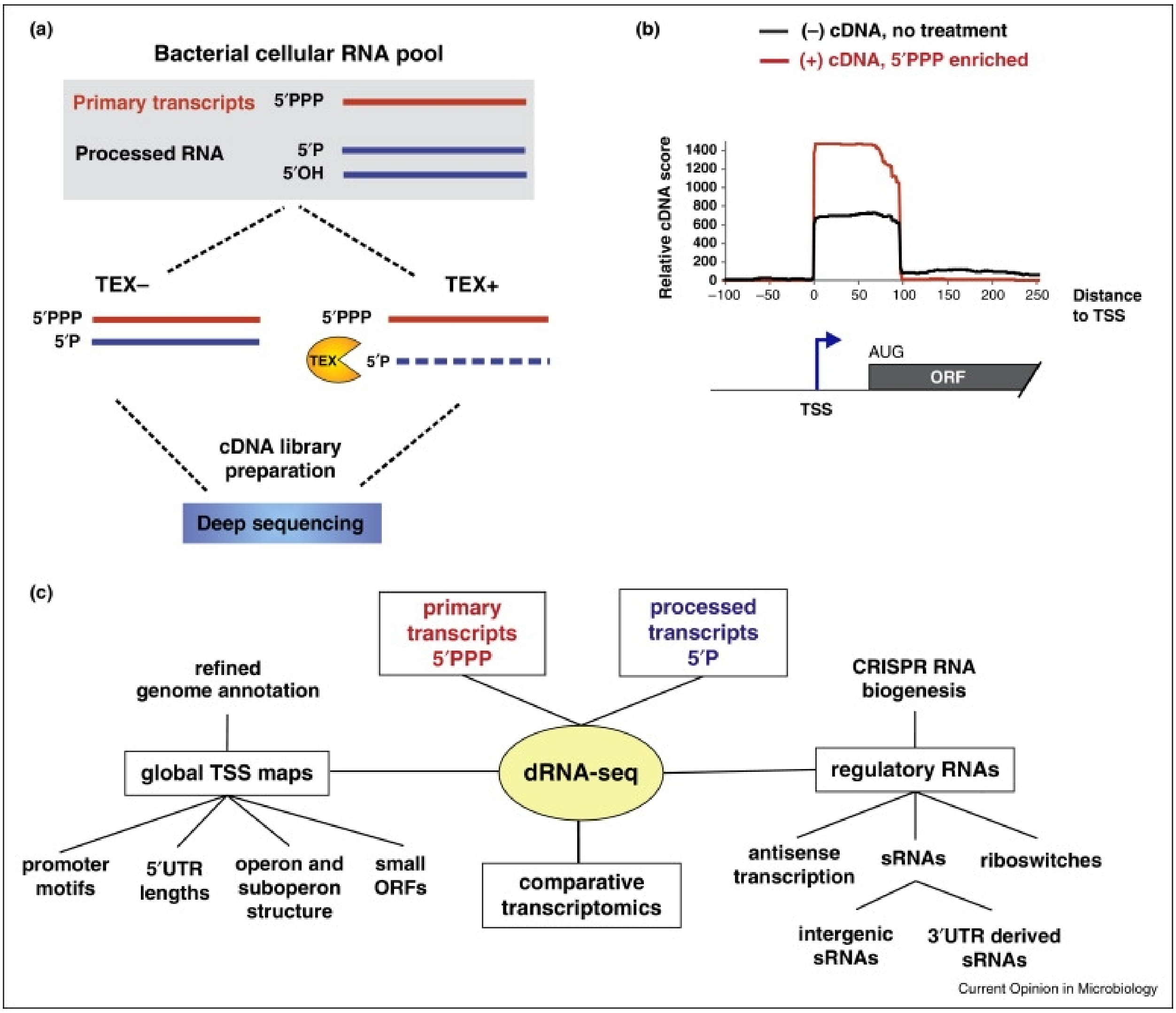


Figure (1-7): *(a) Preparing dRNA-seq data for TSS identification. TEX exonuclease enzyme is used to digest RNA that has a 5´-monophosphate (5'P) or other modifications (5'OH) at their ends, as in processed RNA. TEX does not digest RNAs that have a 5´-triphosphate (5'PPP), as in primary RNAs). (b) RNA-seq is deployed on treated (TEX+, red) and untreated (TEX-, black) libraries and TEX+ > TEX- signals indicate TSS positions. (c) Summary of information that can be gained from dRNA-seq. Extracted from (Sharma and Vogel 2014).*

The analysis of data from the dRNA-seq has allowed the identification of TSS throughout the length of the analyzed prokaryotic genomes, providing information about transcription starts of genes already known, identification of TSS within gene regions, antisense to coding regions and also in no known annotated intergenic regions, as summarized in Figure (1-7) (Sharma & Vogel, 2014). The identification of the primary transcription also aids in the characterization of the promoter region, besides helping to identify genes that present more than one promoter, generating transcripts with different regions 5'URTs and that can be differentially regulated (Sharma & Vogel, 2014).

Transcription start sites may be identified by different methods, some methods are depending on biological experiments and organisms. In Eukaryotes for example, cap analysis of gene expression (CAGE), an oligo-capping in 5′-End sequencing method, was extremely popular before the next-gen platforms arrived (Hashimoto et al., 2004). Variations such as robust analysis of 5'-transcript ends (5'-RATE) were also developed (Gowda et al., 2006). However, these methods were not applicable to Prokaryotes. Genomes of Prokaryotes are relatively simple, and most of their RNAs do not have a 5’ cap to protect it. Thus dRNA-seq was a breakthrough and became the main source of prokaryotic data for TSS identification for these unicellular organisms (Sharma et al., 2010). Transcriptional initiations located upstream to annotated regions, commonly referred to as primary TSS (pTSS), provide important information not only on the early transcription of genes but also on the presence of 5’UTR. In addition to the most understood primary TSS, there are internal TSS (iTSS) located inside the CDS and transcribed in the same orientation of the CDS. Antisense TSS (asTSS) also located inside the CDS but transcribed in the opposite orientation of the CDS, provide the location of antisense RNAs. Finally, orphan TSS (oTSS) located outside the CDS and without known role, show robust ncRNAs definitively transcribed. All these classes are illustrated in Figure (1-8). Some TSSs can have more than one class for different genes at the same time (Amman et al., 2014).

Algorithms were developed for quantifying enrichment from dRNA-seq information for the aim of identification of transcript 5′ ends from its signal on sequencing readouts, illustrated in Figure (1-8) panel b.

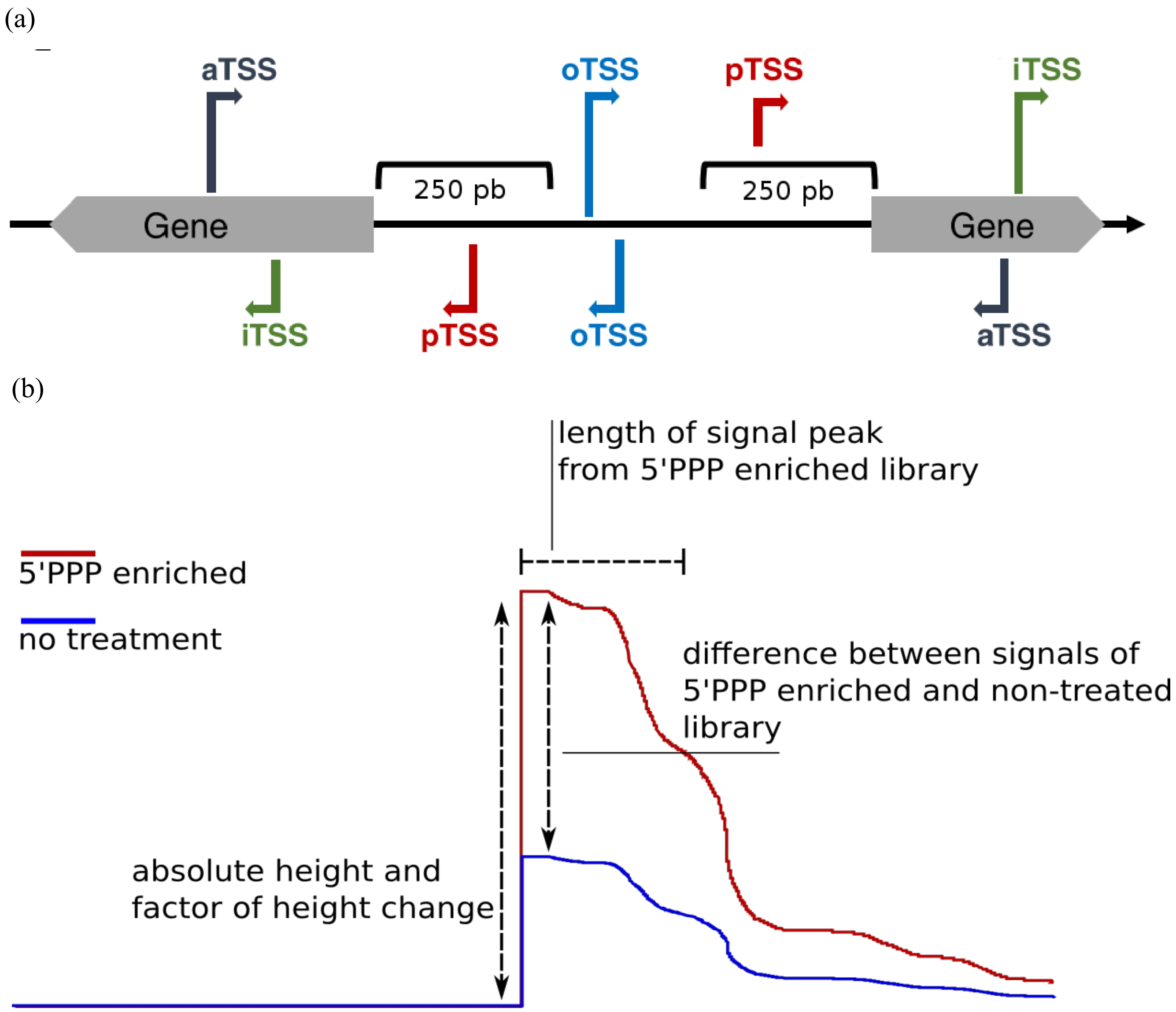


Figure (1-8): *TSS classes and signal. (a) Transcript start site classes are: primary TSS (pTSS), internal TSS (iTSS), antisense TSS (asTSS), orphan TSS (oTSS). The distance to a CDS wich turns a pTSS into an oTSS is arbitrarily chosen 250 base pairs. Adapted from (Amman et al., 2014). (b) Illustration of an ideal 5'PPP enriched and controlled libraries showing the difference between both libraries as a main idea to detect TSS using automated tools.*

The TSSpredator algorithmic rule represented in (Dugar et al., 2013) and (Jorjani & Zavolan, 2014) is one of the first tools for filtering candidate nucleotides supported by arbitrarily pre-selected thresholds of normalized enrichment ratios across experimental replicates. The standardization method corrects for systematic variations in enrichment potency from one experiment to a difference by globally rescaling enrichment ratios. However, since no statistical modeling of the enrichment is performed, there's no way to assess confidence as a main idea to detect TSS using automated tools within the annotations, which can be too conservative or too lenient counting on the experimental variation of the enrichment. The TSSpredator algorithmic rule evaluates posterior enrichment chances in native genomic regions, mistreating info of standardized enrichment ratios across replicate experiments. The posterior chances are also unreal although there's substantial variation in enrichment potency across experiments.

The TSSAR algorithmic rule (Amman et al., 2014) is currently one of the best automated tool for determinate the statistical enrichment of reads at every genomic position. TSSAR fits the information to the Skellam model distribution (the difference in zero-inflated Poisson distributions of read counts between unenriched TEX- and enriched TEX+ datasets). The model should be applied in raw aligned reads to account for variation in transcript abundance, and in some regions, the reads work poorly or can't be modeled. Therefore, some sites are also incorrectly predicted and a few true sites are also incomprehensible owing to inappropriate model parameters.

ToNER tool (Promworn et al., 2017) has been specifically developed for analysis of Cappable-seq data (Ettwiller et al., 2016). The candidate enriched positions were known to recruit a threshold of normalized browse counts from the enriched Cappable-seq dataset.

In theory, the tools developed for dRNA-seq analysis can be applied for Cappable-seq data; but, a lot of pronounced bias in browse count distribution in Cappable-seq enriched knowledge (owing to the bigger enrichment efficiency) compared with dRNA-seq might have an effect on algorithmic rule performance.

ANNOgesic tool (Yu et al., 2018) is a python pipeline that offers adaptive parameter-optimizations. Additionally, numerous visualizations and statistics help the user to quickly evaluate feature predictions resulting from an ANNOgesic analysis. The tool was heavily tested with several RNA-Seq datasets from bacterial as well as archaeal samples with yearly development.

HOMER (Hypergeometric Optimization of Motif EnRichment) is a suite of tools (Heinz et al., 2010) for Motif Discovery and next-gen sequencing analysis. It is a collection of command line programs for Unix-style operating systems written in Perl and C++. HOMER was primarily written as a *de novo* motif discovery algorithm and is well suited for finding 8-20 bp motifs in large-scale genomics data. HOMER contains many useful tools for analyzing ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C and numerous other types of functional genomics sequencing datasets. While the tool is supporting different processes, and needs special compilation for archaeal genomes.

Many other tools, scripts and in house methods available over the internet, also have been created to get the accurate TSS depending on the RNA-seq data, some of them are for Prokaryotes and some for Eukaryotes.

When the fist wave of excitement and discoveries from large-scale mapping of Prokaryote TSS passed, a second moment started, where this PhD thesis is embedded, when additional biological information began to be extracted from the same dRNA-seq data. After TSS where mapped, 5’UTR analyses scaled-up to encompassed all genes, and antisense genes were established; biological phenomena such as the existence of alternative transcripts also began to be found in large scale for Prokaryotes. Comparative Transcriptomics was now a possibility comparing gene content not only at DNA level but also at RNA level.

High-throughput sequencing systems produce massive quantities of biological data for several years. Raw sequencing reads are mapped, assembled, annotated, reviewed, and ultimately accumulated into sequence databases as records enabling the very productive field of Comparative Genomics. Now, comparative transcriptome analysis of different species provides valuable genomic resources and serves to uncover common and conserved sequences of genome and gene evolution in species (Won et al., 2017).

Comparative Transcriptomics of several prokaryotic organisms had shown that transcriptional maps may differ even within the closest species. These comparisons disclosed stunning complexity in the transcriptomes of bacteria and archaea including diverse long regulatory 5′UTRs, non-coding RNAs, alternative operon structures, internal promoters including abundant cis-antisense transcription (Cohen et al., 2016).

The first transcriptome maps were based on manual curation of multiple putative TSSs, the methods are primarily based on dRNA-seq libraries (Sharma et al., 2010). Additional methods that analyze TSSs integrate comparative signals from closely related species, while other methods attempt to describe transcript boundaries (both TSS and transcript termination) based on statistically significant local differences in RNA-seq coverage (Cohen et al., 2016).

Alternative splicing has been investigated in Eukaryotes as a major mechanism for the enhancement of transcriptome and proteome diversity. Prokaryotes haven't the same interest in respect to the simplicity of their gene structure and it is sometimes yet still believed that there is no such combinatorial diversity. Within mRNAs biogenesis, transcription is representing the primary layer of this phenomenon. Alternative transcription initiation ends up in the formation of transcripts differing within their initial CDS or in the length of the 5’UTR. As an alternative, transcripts could share a similar CDS region, however, a distinct 5′UTR may be subject to differential the translational regulation through short upstream ORFs (uORFs) concerned in translational management or within the production of biologically relevant peptides (de Klerk & ’t Hoen, 2015). If instead of larger than usual 5’UTR the utilization of an alternative transcription start site happens downstream to the usual start codon, it may also result in transcripts with different ORFs and diversifies the repertoire of encoded molecules giving rise to protein isoforms with alternative N-terminal. The isoforms of alternative transcripts regulate vital biological processes in Eukaryotes and are related to diseases, aging, as well as cancer. For different genes, many transcripts yield alternative and various proteins with distinct protein interactions, RNA-protein interaction, gene timing, expression location, control aging, cell development speed, control post-transcription, affecting on gene transcription pausing, subcellular localization, stability, DNA-binding properties, lipid-binding properties or protein activity (Reyes & Huber, 2018). It is reasonable to conceive that the same kind of impact would be present in a Prokaryote version of such phenomena, highlighting the importance of Alternative Transcript Start Sites (alTSS), a central concept in this PhD work.

Even though they are untranslated regions, fractions of the 5′UTR sometimes are coded into protein products as upstream ORFs (uORFs) (Araujo et al., 2012; Calvo et al., 2009; Kumar et al., 2015). Some segments of 5'UTR sequences encode to small proteins using start and stop codons. The new small proteins may be playing different roles in gene regulation as well as the interaction of these fractions with the main coding sequences in positive or negative (Chung et al., 2006; Kumar et al., 2015). According to the switching in TSS position-wise the lengths of 5'UTRs are changing with a rearrange of their sequences in a way that create some uORF and change the main way of regulation. The regulation of some coding sequences (CDS) are performed by more than one uORF can regulate in same time (Somers et al., 2013), or regulate in different growth time/conditions switching between them depending on TSS/alTSS switching, illustrated in Figure (1-9).

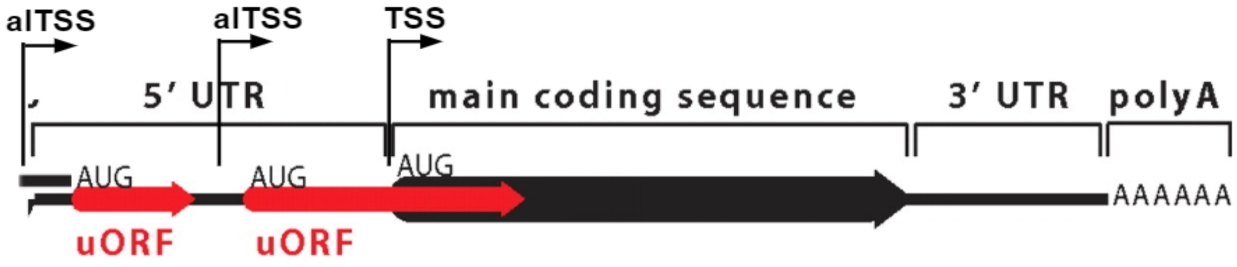


Figure (1-9): *uORF in a gene sequence and the possibility to have other in addition to the overlapped uORF with CDS. Alternative alTSSs could exist also and the switches in their location cause differences in the 5'UTR length. Taken and adapted from (Calvo et al. 2009).*

Secondary TSS (sTSS) is an alternative method for gene regulation instead of the primary TSS (pTSS). Upstream the CDS and have a lower number of transcriptomic reads are the specific features of secondary TSS as a type of alTSSs. A secondary TSS is giving a different isoform of the 5'UTR. Many studies investigated it using both terminology: secondary TSS or alternative TSS (Down & Hubbard, 2016; Dugar et al., 2013; Jorjani & Zavolan, 2014; Karlsson et al., 2017; Li et al., 2015; P. Zhang et al., 2017). The latter, alTSS, is preferred in this PhD work.

1.3 Rationale

The rationale behind the present PhD thesis was to keep exploring a dRNA-seq dataset generated by our group and not yet fully explored. The first work from our group using this dataset explored the iTSS, associated with intraRNAs and their capacity to translate protein isoforms (Ten-Caten et al., 2018). All information regarding alternative TSS (alTSS) potentially available on that dataset was purposely left aside and planed to be addressed by this PhD work. Moreover, alternative uses for the same kind of data were also left unexplored such as secondary structure-based transcript termination sites (TTS) and transcript processing sites (TPS).

Besides these aforementioned scientific open questions, a handful of technical questions/problems were also hindering the group's usage of this dataset. During the original TSS prediction (Ten-Caten, 2017) some points where unclear: (1) different number of TSS in different growth moments, shown in Figure (1-10); and (2) TSS positions were changing every time even running the same bioinformatics protocol. Therefore, along with biological questions to advance *H. salinarum* gene regulatory understanding, bioinformatics technical challenges where addressed as prerequisite to original scientific contributions.

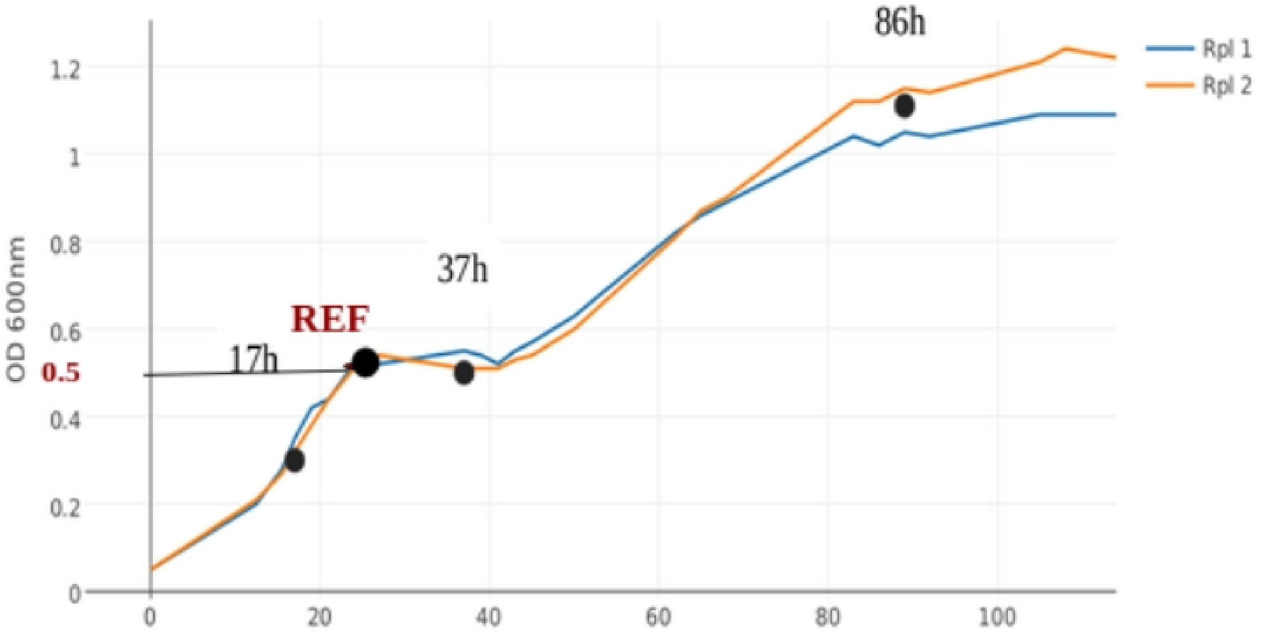


Figure (1-10): *Halobacterium salinarum NRC-1 growth curve and dRNA-seq sampling points (black dots). 2 replicates of the 3 respective points were used. An additional Reference (REF) condition was also sampled in replicate but is not from the growth curve (roughly qualitatively equivalent to an O.D. 0.5 point). Taken and adapted of PhD thesis (Ten-Caten, 2017). This temporal information was deliberately ignored in published manuscript (Ten-Caten, 2018).*

Objectives

2. Objectives

Our research group generated dRNA-seq data for *H. salinarum* NRC-1 as result of a PhD thesis in the University of São Paulo Bioinformatics Grad Program (Ten-Caten, 2017) but only a fraction of the scientific questions it can answer were actually addressed at that moment. In the present PhD thesis the dataset was revisited and the following original biological questions were addressed as objectives for this PhD:

1. Map transcript processing sites (TPS) into the *H. salinarum* NRC-1 genome;
2. Detect alternative transcription start sites (alTSS) in the *H. salinarum* NRC-1 genome;
3. Identify the change within the usage of alTSSs influenced by different growth conditions;
4. Detect transcription termination sites (TTS) in the *H. salinarum* NRC-1 genome.

Materials and Methods

3. Materials and Methods

In the following section we describe the methods used to accomplish the objectives of the present PhD work and also the used materials and datasets along with their sources.

3.1 dRNA-seq libraries of *Halobacterium salinarum* NRC-1 and other organisms

The dRNA-seq data used throughout this PhD thesis was not generated here but rather by another former PhD student from the USP Bioinformatics Grad Program, Dr. Felipe Ten-Caten within his research project. The preliminary analysis was actually defended by Dr. Ten-Caten in his PhD defense, published along with his thesis and is referred here as (Ten-Caten, 2017). The second and improved dataset, published in a peer-review prestigious journal, is referred as (Ten-Caten et al., 2018). In the following there is a precise description of how the datasets used here was constructed by Dr Ten-Caten.

The RNAs were extracted from *Halobacterium salinarum* NRC-1 growing in CM medium (250g / l NaCl, 20g / l MgSO4, 2g / l KCl, 3g / l Sodium Citrate, 10g / l Bacteriological Peptone Oxoid) in the condition as described in (Ten-Caten, 2017). When the culture reached optical density (OD) around 0.5 (half of the first log growth phase = Reference sample), 2 ml aliquots were collected in Eppendorf tubes and centrifuged at 14,000 rpm for 1 minute. After centrifugation, the supernatant was discarded and the pellets were stored at -80°C until extraction of the RNAs was performed. For a collection of samples along the growth curve, the NRC-1 lineage was grown in CM medium in 200 ml volume at 37°C under constant agitation (125 rpm) and luminosity. The pellets were collected as previously presented at three distinct growth times, at 17 hours, 37 hours, and 86 hours, as shown in Figure (1-10). All experiments were performed with biological replication and prepared by (Ten-Caten et al., 2018). RNA extraction and PCR reactions, for the four samples, used the same protocol of (Ten-Caten et al., 2018; Zaramela et al., 2014). The RNA samples from each replicate were divided into two aliquots. One of them was treated with 1U of the Terminator 5'-Phosphate-Dependent Exonuclease enzyme (TEX) while the second aliquot was incubated only in the presence of the buffer with no TEX enzyme, then the processed RNA the samples were treated with the TAP (Tobacco Acid Pyrophosphatase) enzyme (Ten-Caten et al., 2018). These two experimental conditions are named throughout this PhD dissertation text as TEX+ and TEX- to refer to samples treated or not treated with TEX, respectively.

The libraries of dRNA-seq were prepared using the method in (Sharma et al., 2010; Ten-Caten et al., 2018). The transcriptome samples were defined in 4 representative samples: 17h, 37h, 86h, as aforementioned, and reference growth condition, named REF throughout this text. For strand specific dRNA-seq library preparation, TruSeq Small RNA Sample Preparation (Illumina) was used. Sequencing was performed using MiSeq Reagent v2 300 cycles. The raw sequencing data is available at the NCBI SRA database under the accession ID: SRP137801.

Reads obtained were processed using a bioinformatics pipeline assembled from publicly available tools, specific for each point in the workflow, by our own group with the input of several PhD students and projects, including the preset one, but lead mainly by Dr Alan Lorenzetti as part of his PhD thesis at the USP Bioinformatics Grad Program (Lorenzetti, 2021). The pipeline, published not as a bioinformatics tool but rather as a mean to achieve the biological question answering in (Ten-Caten et al., 2018) was named Caloi-seq and it is mentioned many times throughout this and other PhD dissertations from our group. This pipeline is openly available at https://github.com/alanlorenzetti/frtc/ and is summarized in the following.

Libraries were downloaded from NCBI Sequence Read Archive (SRA) and converted to FASTQ format using SRAdb v1.40.0 or fastq-dump v2.8.2. Paired-end and single-end libraries were processed using Trimmomatic v0.36 (Bolger et al., 2014) to trim known adapters. Reads were trimmed to the end if the mean Phred of a four nucleotide sliding window was less than 30 and only reads satisfying the minimum length of 20 nucleotides were allowed to move on into the pipeline. Reads filtered in as a pair were aligned to reference genomes using HISAT2 v2.1.0 suppressing alignments resulting in fragments longer than 1000 nucleotides. Orphan sequences that should have pair from paired-end libraries and those coming from single-end runs were aligned using the single-end mode. Multi-mappers were allowed if aligning up to 1000 times and with no spliced, soft-clipped, gapped, discordant and mixed alignments. The output SAM files were converted to BAM using SAMtools v1.3.1 and input in MMR to find the most likely position for each multi-mapper. MMR computes the genome-wide coverage considering only uniquely aligned reads, and then assign a unique position to each multi-mapper based on its potential of reducing the local variance of coverage. Paired-end alignments adjusted by MMR that are too small and align entirely to direct repeatswere removed to avoid uncertainty. Genome-wide coverage was computed for every library by deepTools v2.5.3. We used bedtools v2.2.26 to compute 5′ and 3′ profiles for each library. Data visualization was performed sporadically using IGV v2.4.6. The data were processed for visualization in the Gaggle Genome Browser tool (GGB) (Bare et al., 2010). This GGB browser, cited many times in this text, was used for exploring and gathering biological insights but it was also used to generate many figures presented in this PhD text or even manuscripts published or in preparation.

In addition to the aforementioned data of *Halobacterium salinarum* NRC-1, five other organisms, shown in Table (3-1-1) were used according to the availability of their dRNA-seq in at least one growth condition and technical possibility to use our Caloi-seq pipeline for equal analysis.

Table (3-1-1): Taxonomy and genomic general information about the organisms used.

Organism	Phylum	#Chr.	#Plasmids	#Genes
Halobacterium salinarum NRC-1	*Euryarchaeota*	1	2	2834
Haloferax volcanii DS2	*Euryarchaeota*	1	4	4130
Methanocaldococcus jannaschii DSM 2661	*Euryarchaeota*	1	2	1832
Thermococcus onnurineus NA1	*Euryarchaeota*	1	0	2027
Caulobacter crescentus NA1000	*Proteobacteria*	1	0	3989
Thermus thermophilus HB8	*Deinococcus-Thermus*	1	2	2291

The dRNA-seq data experiments for the selected organisms were retrieved from NCBI's SRA database, as shown in Table (3-1-2):

Table (3-1-2): List of the organisms and their dRNA-seq data runs used.

Organism	**Accessions**	**SRA run**
Halobacterium salinarum NRC-1	NC_002607.1 NC_001869.1 NC_002608.1	SRR6953855 SRR6953856 SRR6953857 SRR6953858 SRR6953859 SRR6953860 SRR6953861 SRR6953862 SRR6953863 SRR6953864 SRR6953865 SRR6953866 SRR6953867 SRR6953868 SRR6953869 SRR6953870 SRR6953871 SRR6953872 SRR6953873 SRR6953874
Haloferax volcanii DS2	NC_013964.1 NC_013965.1 NC_013966.1 NC_013967.1 NC_013968.1	SRR3623113 SRR3623114 SRR3623115 SRR3623116 SRR3623117 SRR3623118
Methanocaldococcus jannaschii DSM 2661	NC_000909.1 NC_001732.1 NC_001733.1	SRR4238017 SRR4238018 SRR4240639 SRR4240641
Thermococcus onnurineus NA1	NC_011529.1	SRR4042633 SRR4042634 SRR4042635 SRR4042636
Caulobacter crescentus NA1000	NC_011916.1	SRR1273068 SRR1273069
Thermus thermophilus HB8	NC_006461.1 NC_006462.1 NC_006463.1	SRR390354 SRR390355 SRR390356 SRR390357 SRR390358 SRR390359

3.2 TSS inspection and tool selection

TSSAR tool (Amman et al., 2014) was recommended as the best tool to identify TSS from dRNA-seq data of prokaryotes (Amman et al., 2014, 2018; Promworn et al., 2017; Ten-Caten et al., 2018; Yu et al., 2018), and many studies used it in both TSSAR-web (http://rna.tbi.univie.ac.at/TSSAR) and TSSAR stand-alone programs, cited at least 62 times until January/2023 (Google Scholar).

To account for the various transcription actions in the genomic sequence, every site is evaluated within the context of its native local transcription level encompassing by a window approach. An arbitrary region within the window could be a mix of transcribed and not transcribed sections. For the first, scan start counts are modeled by a Poisson distributed random variable; whereas the latter is predicted to be zero-mean noise uniformly distributed. To estimate the parameters that describe solely the Poisson part, TSSAR applies a zero-inflated Poisson model regression. All excess zeros believed to originate from untranscribed regions. Finally, the mean (λ) of the remaining sample is calculated, describing the distribution of the transcribed part of the reading window. TSSAR aims for locating positions with a considerably enriched signal within the TEX+ library, considering the expected variability from the reference model TEX-, or in other words statistically test for $\lambda_{(+)} > \lambda_{(-)}$. The actual model implementation uses the difference (D) of such read alignment counts since the random variable of such Poisson-following random variables, $D = N_{(+)} - N_{(-)}$, is well-known and follows a Skellam distribution with mean $\mathbf{E}[D] = \lambda_{(+)} - \lambda_{(-)}$. (https://en.wikipedia.org/wiki/Skellam_distribution). The distribution's form and location measure are characterized by the previously deduced λ parameters related to the total sample and every value is often evaluated on how well it fits the reference model, as usual in model fitting. Given a *p*-value cutoff, background noise threshold positions with a specified number of reads in the TEX treated library (TEX+) and merge a range of consecutive positions, a coordinate on the genome is annotated as TSS. Thus, TSSAR was used to produce the TSS lists for this project and a simplified workflow is shown in Figure (3-2-1).

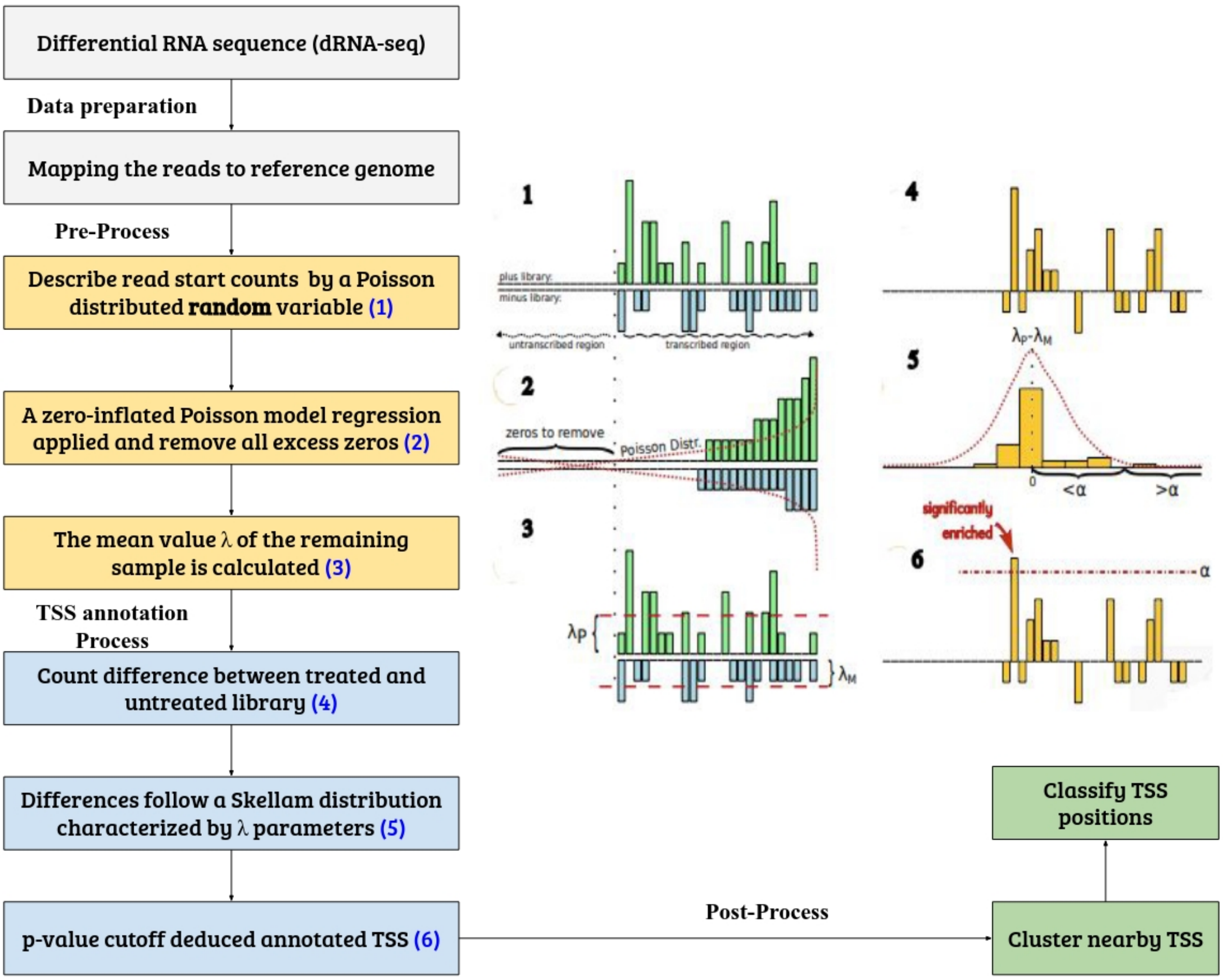

Differential RNA sequence (dRNA-seq)
Data preparation
Mapping the reads to reference genome
Pre-Process
Describe read start counts by a Poisson distributed random variable (1)
A zero-inflated Poisson model regression applied and remove all excess zeros (2)
The mean value λ of the remaining sample is calculated (3)
TSS annotation Process
Count difference between treated and untreated library (4)
Differences follow a Skellam distribution characterized by λ parameters (5)
p-value cutoff deduced annotated TSS (6)
Post-Process
Cluster nearby TSS
Classify TSS positions
1
plus library
minus library
untranscribed region
transcribed region
2
zeros to remove
Poisson Distr.
3
λp
λM
4
5
λP-λM
0
<α
>α
significantly enriched
6
α

Figure (3-2-1): *Illustration of TSSAR Workflow describing the dRNA-seq experiment entrance data, preparation of the data, pre-processing and statistical analysis depending on zero-inflated Poisson model regression shows in panels 1to 6, the TSS annotation process then post process. Statistical graphs are taken and adapted of (Amman et al., 2014).*

Around 100 times of repetition using the dRNA-seq data of *Halobacterium salinarum* NRC-1 treated by Caloi-seq pipeline were rerun in TSSAR-standalone tool, avoiding the different results in each run of TSSAR-web tool coursed by some non user-defined internal random number generator. The results were analyzed to get the intersect TSS of each repetition using fixed post-process parameters, *p*-value cutoff 0.001, merge 10 consecutive TSS, and threshold of 10 reads per position.

We made a small modification to the original code adding an explicit seed to the random number generator. We controlled reproducibility using a setseed(12345) by editing the TSSAR program in R language script parts. All TSS less than *p*-value 0.01 was fixed even if changed the setseed for the other once. This procedure made the tool more stable and useful to get all TSS and alternative positions.

In General, TSSAR has always omitted regions where they couldn't be mapped using the original statistical method (Zero-inflated Poisson), the regions are depending on the used window. Moreover, the TSSAR cluster script *gff_cluster.pl* is depending in sequence on: 1) Lowest *p*-value of clustered positions; 2) the highest reads of clustered positions; 3) the mean position of the clustered reads. Then to classify the final clustered positions using *TSSAR_classification.pl* the code is depending on the 2 files resulting of previous clustering application **_c.gff* and **_smear.table* as positions are clustered in the first step. The smear file contains the chosen positions and their uncertainty of the highest reads. At all, the automated clustering method is not exactly certain and sometimes change the expected class for the next position, so that, creating a new cluster method was necessary, using R-script in simple way *TSS_cluster.R* to depend on the highest read position as the first choice then *p*-value, or using 0 position cluster method to classify the new data using *TSS_classes.R* so can get the exact classes for each position.

Other similar tools were used to non-extensively compare with TSSAR and allow us confidence that TSSAR choice was sound. TSSpredator (Dugar et al., 2013) provides parameters with an easy interface that permits for the set of all necessary parameters. By all aspects of the prediction, the procedure is often made-to-order. However, the program conjointly provides many presets for a fast application of the program. In addition, TSSpredator can even be used via a statement interface for the simple integration into automatic information analysis pipelines. TSSer, an automated tool depends on a binomial distribution to calculate the probability of TSS represented (Jorjani & Zavolan, 2014), then identify the frequency of reads of enriched and unenriched samples (TEX-), calculate the probability that a genomic position has a higher expression in the TEX-treated (TEX+) compared with the untreated sample using Gaussian distribution, calculate the 5′ enrichment factor then assuming that factor follows a normal distribution, calculate the probability that a TSS is enriched across a panel of *k* replicate paired samples, then examine the frequencies of sequenced reads in a region of length $2L$ centered on the putative TSS. ToNER is a tool for identifying nucleotide enrichment signals in feature-enriched RNA-seq data (Promworn et al., 2017). The main idea in this tool is to consider the RNA-seq data from the RNA-enriched library and a separate control RNA-unenriched is pre-processed and use a simple normalized ratio between counts. The tool is designed particularly for Cappable-seq data and is not very accurate with dRNA-seq, according to the authors.

3.3 Alternative transcription start site selection and estimation

The existence of alTSS is related to the sequence reads that should exist in normal conditions, however, TSS differ in their usage in specific conditions. To choose the most abundant primary genuine TSS (gTSS) and its pair alTSS we searched reads where a TSS1 (gTSS) is higher in reads number than a TSS2 (alTSS). Select the most two abundant TSSs depending on the distance of CDS with a window of 5-nt downstream the start codon and 150-nt upstream the CDS. Using TSSAR modified tool, the statistical criteria utilized a *p*-value should be lower than 10^{-6}, with at least 100 reads per position and interval difference portion lower than 95% between both TSSs (i.e. they must have comparable heights), illustrated in Figure (3-3-1).

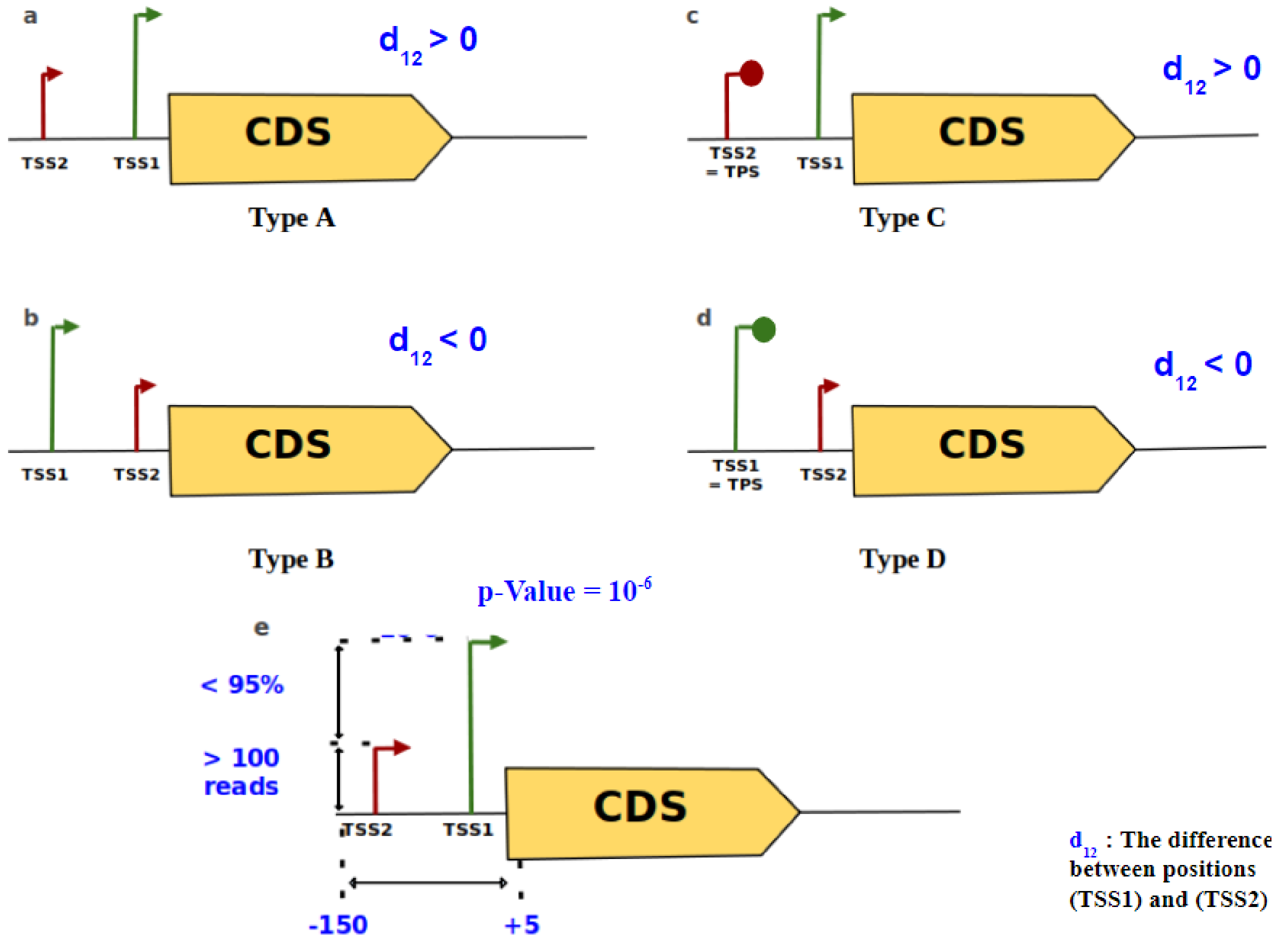


Figure (3-3-1): *Illustrate the selection of gTSS and alTSS, where always the principal gTSS (TSS1) have more reads than alTSS (TSS2). (a) TSSs pairs Type A, where TSS1 is closer to CDS and both TSS1 and TSS2 are primary TSS; (b) TSSs Type B, where TSS1 is farther than TSS2 and both TSS1 and TSS2 are primary TSS; (c) TSSs Type C, where TSS1 is closer than TSS2 and TSS1 is primary TSS, TSS2 is processed TPS; (d) TSSs Type D, where TSS1 is farther and TSS1 is processed TPS, TSS2 is primary TSS; (e) The parametric criteria to choose the most abundant TSS, p-value lower than 10^{-6}, not less than 100 reads per site and difference proportion between both sites <95%. Distance d_{12} is the distance between coordinates of TSS1 and TSS2. d_{12} >0 means TSS1 downstream of TSS2 and d_{12} < 0 TSS1 upstream of TSS2.*

3.4 fake TSS (fTSS) detection

Our growing understanding that dRNA-seq derived conclusions can be affected by RNA secondary structure lead to careful differentiation between TSS and potential false or fake TSS (fTSS). To the best of our knowledge, no tool currently available makes the identification of potential false positives based on structural properties automatically. Therefore, we split these cases using a simple approach introduced by (Ten-Caten et al., 2018) which is filtering out the TSS that is just upstream of regions with high MFE (minimum free energy). The novelty introduced is to give some use to a subset of fTSS: to find putative transcript termination sites (TTS). The MFE cutoff was established considering the overall genome MFE of short tiled subsequences, as shown in Figure (3-4-1).

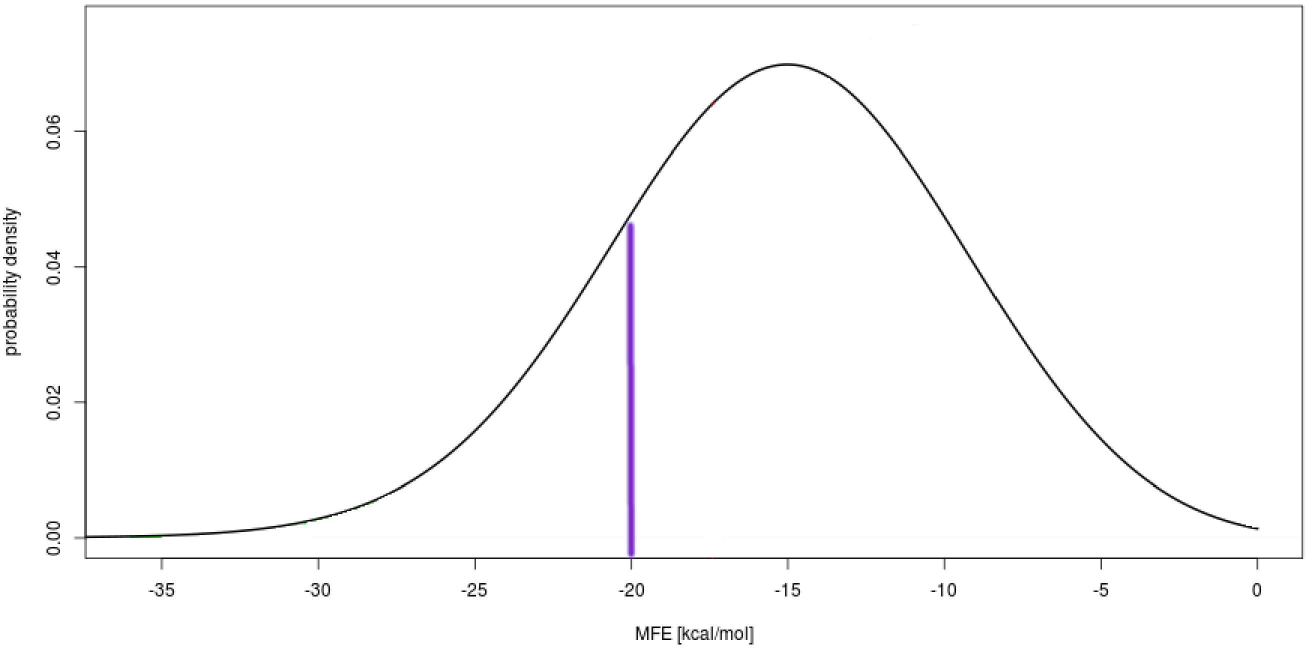


Figure (3-4-1): *Filtering fTSS due to structure forming sequences distribution (kernel smoothed) of minimum free energy (MFE) calculated for 51 nt long sequences tiled with 10 nt offset sliding window along the H. salinarum NRC-1 genome. Set comprising all genome is in black and filtering cutoff in purple. Sequences with MFE below cutoff were filtered out as potential false alTSS.*

3.5 Transcript processing site (TPS) mapping

We used TSSAR-GaVI2 to identify TPS positions deliberately changing the tool's TEX+ input for the TEX- alignment file and *vice versa* in order to highlight statistically significant 5' monophosphorylated depletion signals (TEX- > TEX+). Therefore, TSSAR's "TSS" outputs are actually TPS, as shown in Figure (3-5-1). TSSAR parameters were *p*-value of $p < 10^{-9}$ with a minimum of 10 reads and a distance of "TSS" grouping of at least 5 nt. More details can be found the manuscript published (Ibrahim et al., 2021) reproduced in the Results section. Also, since TSSAR-GaVI2 is a result of this PhD work, it is described in that section and here at Methods section we just cite its usage.

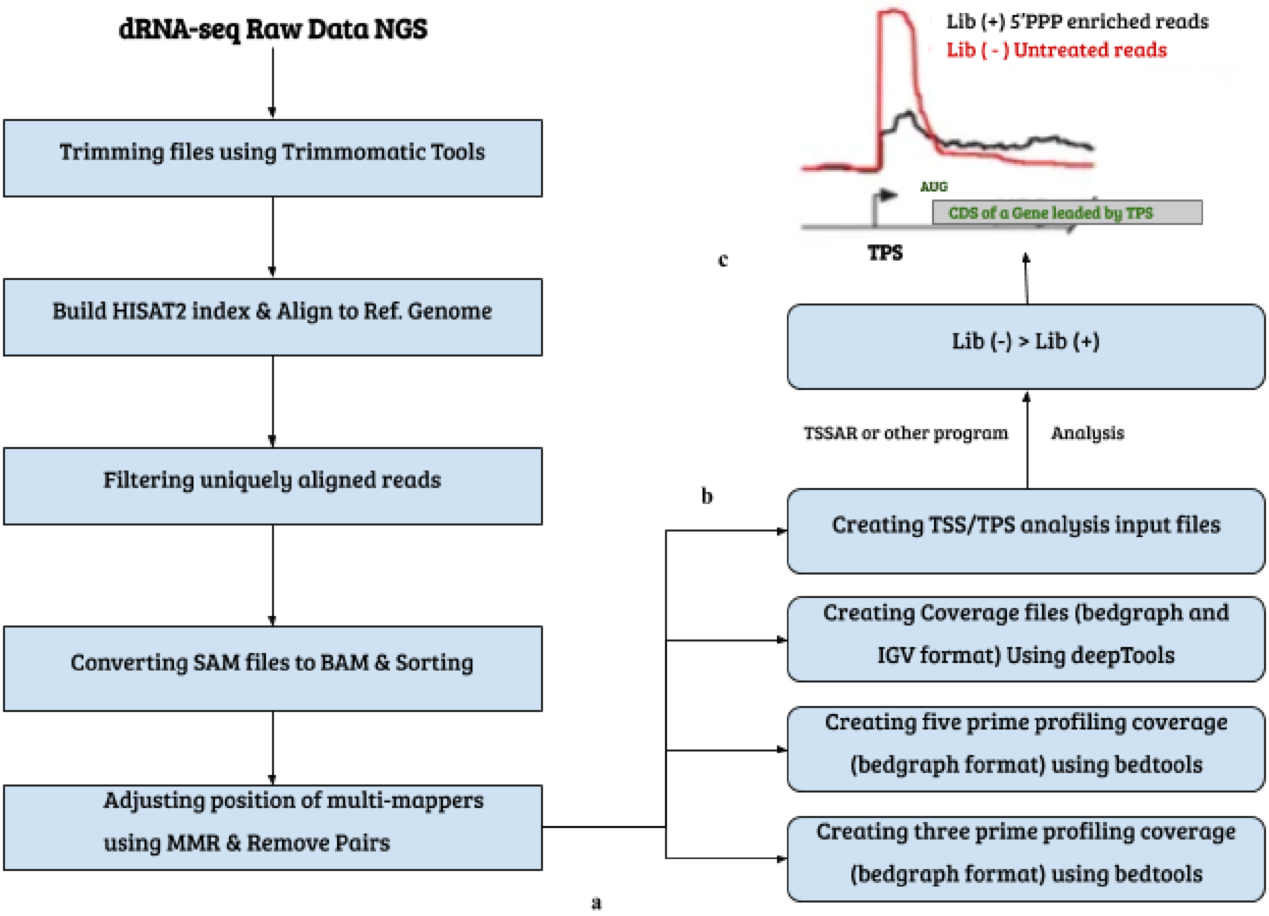

dRNA-seq Raw Data NGS
Trimming files using Trimmomatic Tools
Build HISAT2 index & Align to Ref. Genome
Filtering uniquely aligned reads
Converting SAM files to BAM & Sorting
Adjusting position of multi-mappers using MMR & Remove Pairs
a
b
Creating TSS/TPS analysis input files
Creating Coverage files (bedgraph and IGV format) Using deepTools
Creating five prime profiling coverage (bedgraph format) using bedtools
Creating three prime profiling coverage (bedgraph format) using bedtools
TSSAR or other program
Analysis
Lib (-) > Lib (+)
c
TPS
AUG
CDS of a Gene leaded by TPS
Lib (+) 5'PPP enriched reads
Lib (-) Untreated reads

Figure (3-5-1): *The workflow to annotate the TPS from dRNA-seq data produced by Illumina next generation sequencer. (a) From raw to coverage (Caloi-seq) steps: mapping, filtering and adjusting the reads. (b) The output getting coverage data to visualize and study the raw data, and bam files to use for the TSS or TPS annotating using automated programs. (c) Illustrate of TPS leading the gene transcription, where library (-) of untreated cDNA reads more than library (+) of treated cDNA reads.*

Results and Discussion

4. Results and Discussion

In this section the results of this PhD work are going to be presented along with our interpretation of them. We prefer this format of coupled Result and Discussion to facilitate reading. It starts with a section describing our efforts to re-analyze legacy data from our own research group, establishing the bioinformatics protocol used and non extensively comparing it with some known options. The second sub-section applied the developed protocol to answer our first biological question, the search for alternative TSS in *H. salinarum*. The third sub-section presents the same application but to other organisms. The fourth sub-section deals with false positive TSS and how they can be turned into useful information. The fifth sub-section applies the discussion form the previous section to answer a meaningful biological question: the search for termination start sites. Finally the sixth and last sub-section, the most important of all since it resulted in the publication as a main author in a peer-reviewed journal, deals with mapping transcript processing sites tweaking dRNA-seq data in a way not originally intended by their creators.

4.1 Differential RNA-seq (dRNA-seq) Reanalysis

As mentioned in previous sections, the dRNA-seq data used throughout this PhD thesis was generated by Dr. Felipe Ten-Caten for his PhD thesis. The first analysis is referred as (Ten-Caten, 2017), the publicly available PhD thesis itself. Our own preliminary research at the beginning of this project showed that improvement was necessary. The group kept working with the same dataset, improved the whole analysis bioinformatics pipeline, and end up publishing part of the resulting work focusing only on iTSS and its relationship with alternative protein isoforms (Ten-Caten et al., 2018). The second (and improved) dataset is referred to as (Ten-Caten et al., 2018) and is the start point from improvements yielded by the present PhD project, which are described from here one wards.

TSSAR is the most recommended tool by different studies to perform dRNA-seq analysis (Amman et al., 2014, 2018; Promworn et al., 2017; Yu et al., 2018), and it was used to produce TSS lists for this PhD thesis and (Ten-Caten et al., 2018) alike. However, we noticed that the final list obtained by its online implementation was different when we ran again the same input dataset and parameters relative to (Ten-Caten, 2018). The original TSSAR paper do not mention any stochastic step on its inference procedure but by inspecting the source code and exchanging messages with the authors we could detect a subroutine that indeed made use of a pseudo-random number generator. Therefore we controlled this by changing the source code adding a “seed” definition, with an illustrative result shown in Figure (4-1-1).

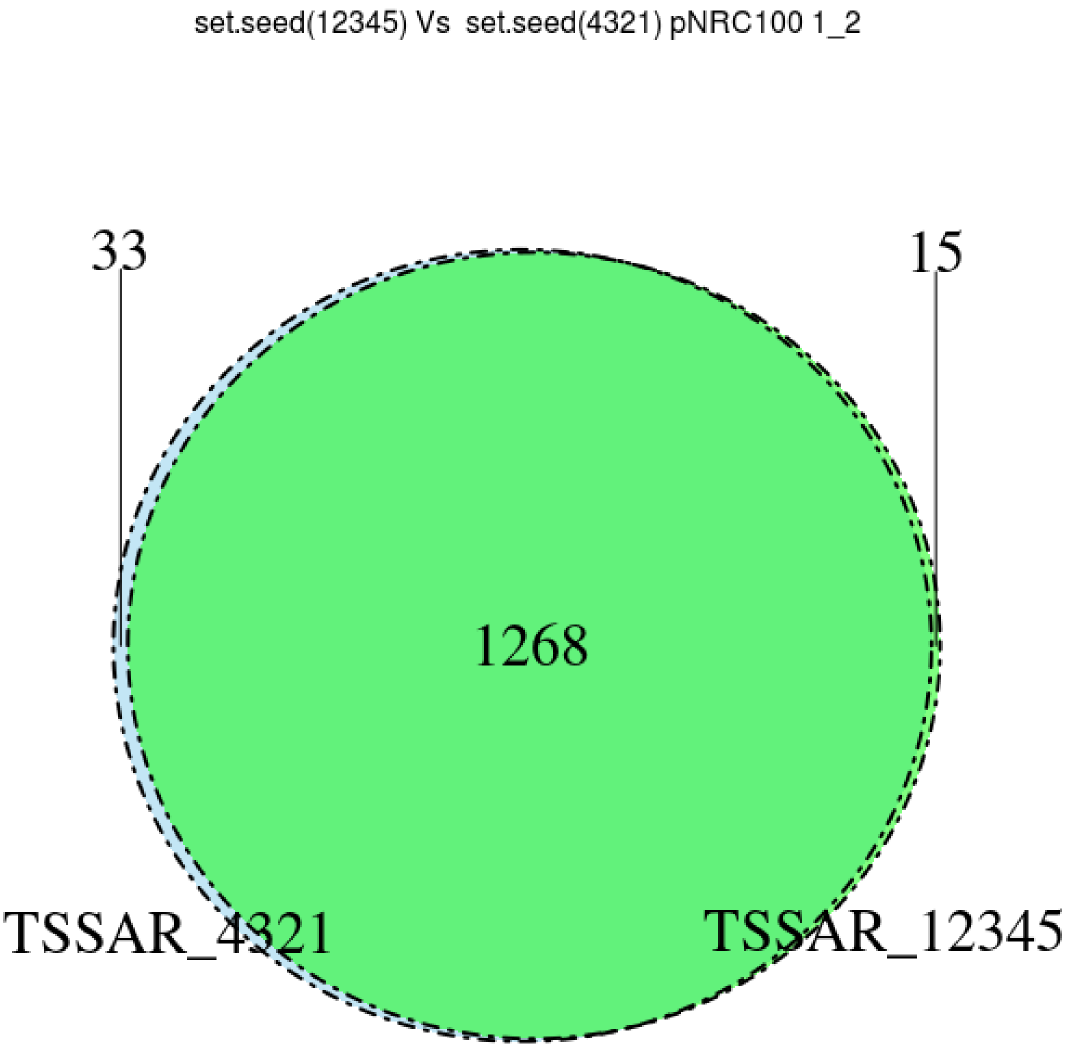


Figure (4-1-1): *TSSAR-standalone using different setseeds and detect TSS differences for p-value 0.1 that out the intersect area. The example used the information of plasmid pNRC100 to choose a minimum p-value cutoff for different setseeds and fix this cutoff to run the program.*

Another relatively simple modification that improved the TSSAR tool was the change of the statistical model counting the aligned reads. The modification of Zero-inflated Poisson to Zero-truncated Negative binomial (Sampford, 1955) statistical method (also called Gamma-Poisson) resulted more accurate results for all *p*-values even if *p*-value was higher than 0.01, running the program became 7X faster than the original TSSAR-standalone for same input data and parameters. The main conceptual difference between the models is that the well known Poisson model has a numerical coincidence between mean and variance (λ), being a single parameter model (Ziegel & Ross, 1998). Gamma-Poisson, on the other hand, is a mixture of a Gamma distribution with a Poisson distribution and does not require the mean to be the same as the variance. This implementation was named in the present PhD work as **TSSAR-GaVI2**.

A high quality manually curated set of clear TSS in *H. salinarum* was defined in order to compare the TSSAR and TSSAR modified, TSSAR-GaVI2. The whole genome was browsed meticulously using Gaggle Genome Browser GGB searching for clear indisputable start sites in the global dRNA-seq data. This yielded 634 test TSS, which 86% were retrieved by TSSAR-GaVI2, as shown in Figure (4-1-2).

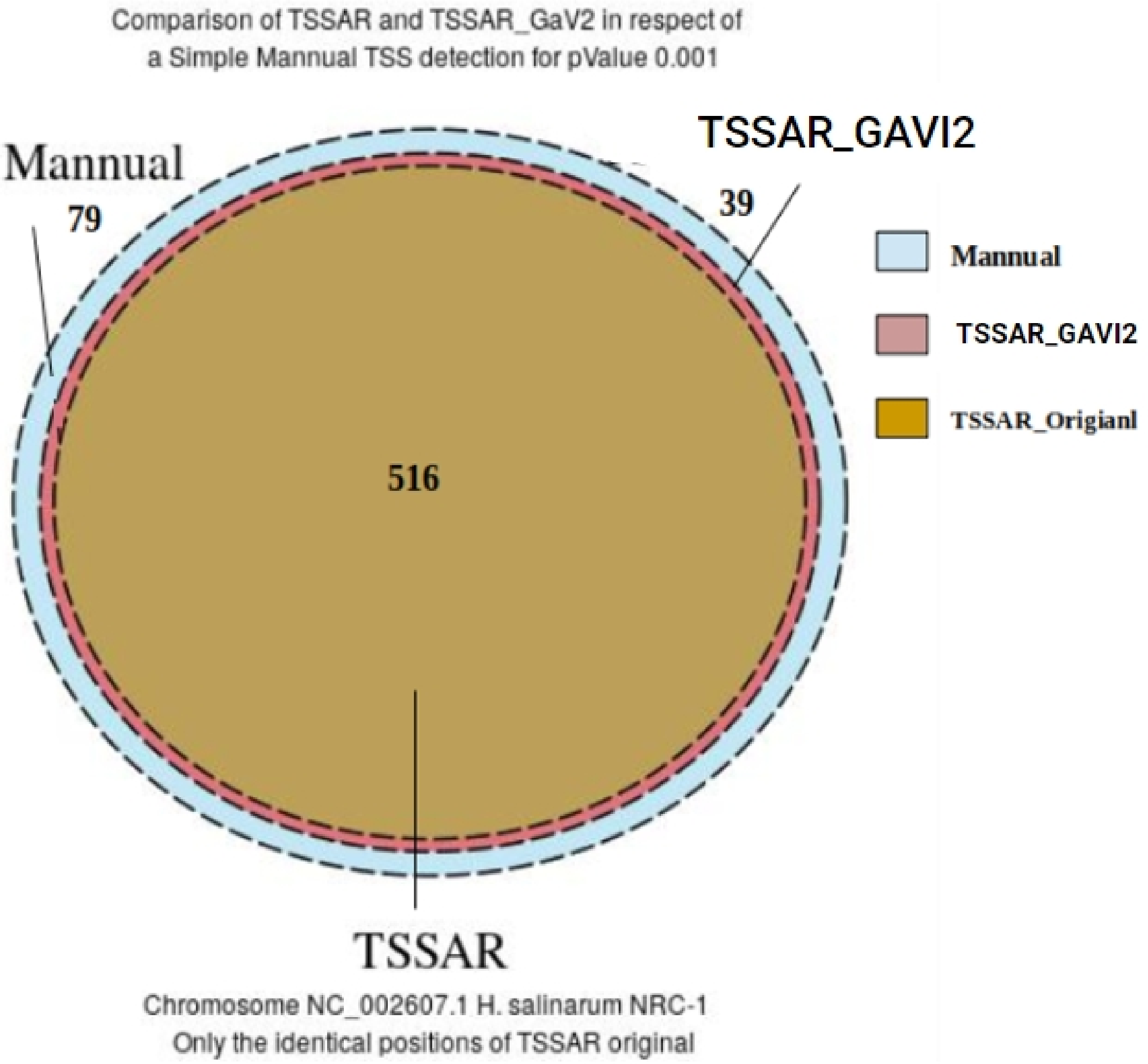


Figure (4-1-2): *A comparison between the results of the original TSSAR after fixing the random variable generator and TSSAR-GaVI2 used the Gamma-Poisson statistical model to inspect the TSS positions in respect to the detection using simple manual dRNA-seq read walk. The reads were filtered using p-Value < 0.001 then chose the TEX+ reads > 5 folds of TEX- reads as a specific filter. Then to create the Venn Diagram used the identical points equaled to manual inspection and removed all the aberrations that existed in TSSAR original run.*

As additional tests to ground the choice of TSSAR as our tool during this PhD and also the improvements made over it, we compared published TSS data form a well known model organism for which there is manually curated start site information: *Escherichia coli* K-12 (Promworn et al., 2017). Since this is a high-quality manually curated dataset we could have good confidence about the overall performance of the selected method, confirming what the literature already concluded before about TSSAR. Figure (4-1-3) shows the comparison between several methods (a combination of the read alignment pipeline they use with the actual TSS finding procedure) for the *E. coli* dataset. Selected data were examined as three groups: 1) raw data downloaded from their sources in NCBI's SRA repository; 2) pre treated with Caloi-seq pipeline, the one we use in our group created for (Ten-Caten et al., 2018); and 3) pre processed with ToNER pipeline (Promworn et al., 2017). Figure (4-1-3) shows only the best performance combination for each method but all combinations are available as online appendix (Appendix 1).

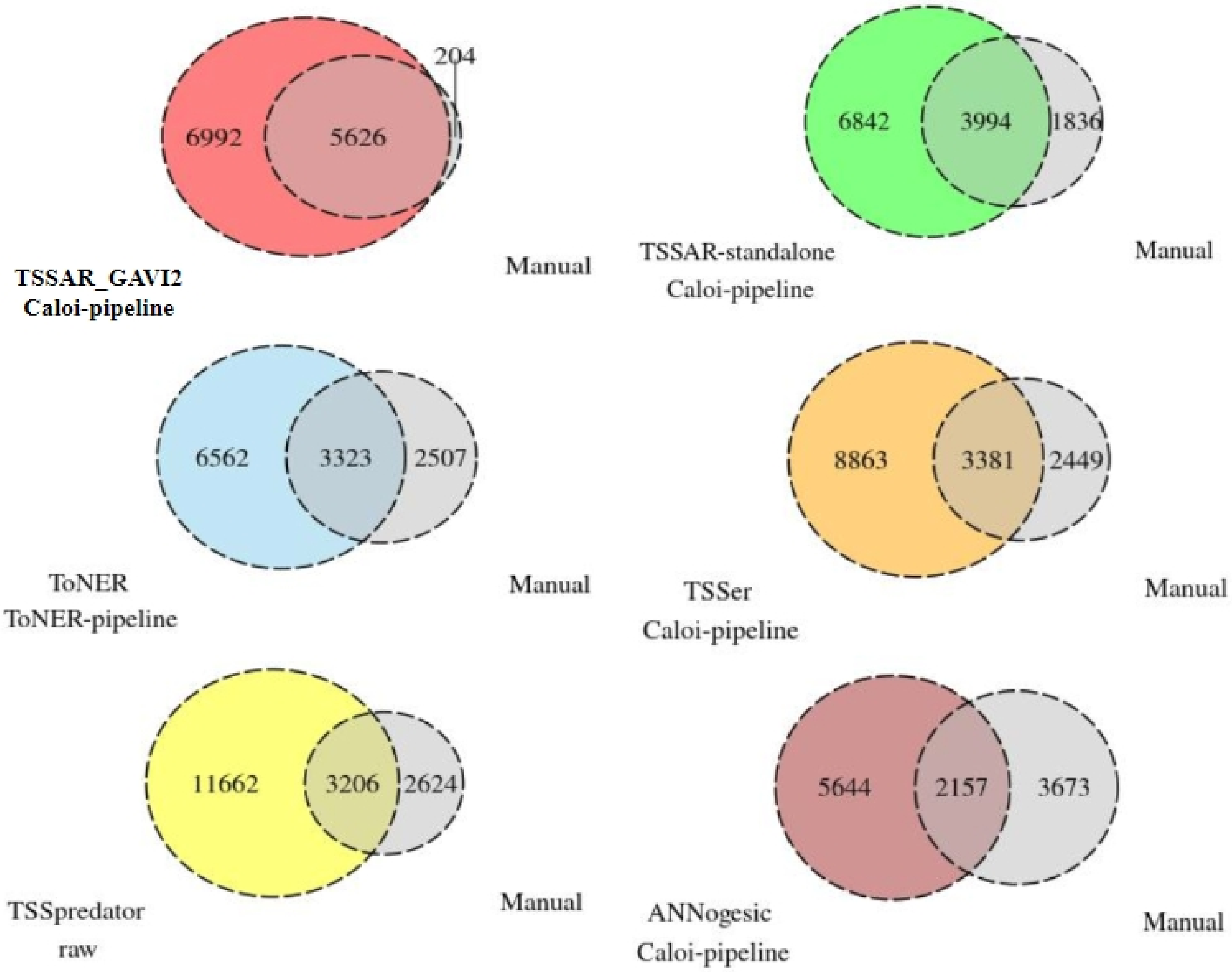


Figure (4-1-3): *A comparison between the results of several tools using a high-quality manually curated dataset of E. coli. In Venn diagrams, Manual-exclusive findings added to the intersections give the 5830 total E. coli TSSs.*

4.2 Alternative Transcription Start Sites (alTSS) in *H. salinarum* NRC-1

We discovered **91** genes with alTSSs using all time-course points collapsed into a single dataset. This collapsing approach is the same used by (Ten-Caten et al., 2018). The approach that considers each time point separately was mainly developed for the present PhD work and is also explored in subsequent analysis, however it is convenient to start with all time points collapsed and thus, consider the 91 genes presenting alTSS.

All candidate alTSS belong to one out of four different sub-categories depending on the signal characteristic: Types A, B, C and D, which definition is detailed in Methods section and summarized in the Figure (3-3-1). In short, Types A and B alTSSs are both primary TSS with the strongest (TSS1) signal closest or farthest from CDS start codon, respectively. Type C have the strongest signal primary (TSS1) closer to the start codon and the weaker signal (TSS2) upstream to it. Conversely, Type D shows evidence of processing for its strongest signal (TSS1) upstream to the weakest primary site (TSS2), which in turn is closer to the start codon. The literature usually name the canonical (i.e. established or official) TSS as “genuine TSS” (gTSS), although this is a definition resembling an outdated concept from when the widespread condition-dependent alternatives were largely unknown. We conform with the tradition nomenclature in spite of the fact that we know that in some environmental scenarios, “alternative” would actually be the norm. In this PhD work we adopt the simple definition that gTSS are those from a given pair with strongest sequencing signal (highest peak, more normalized aligned read counts).

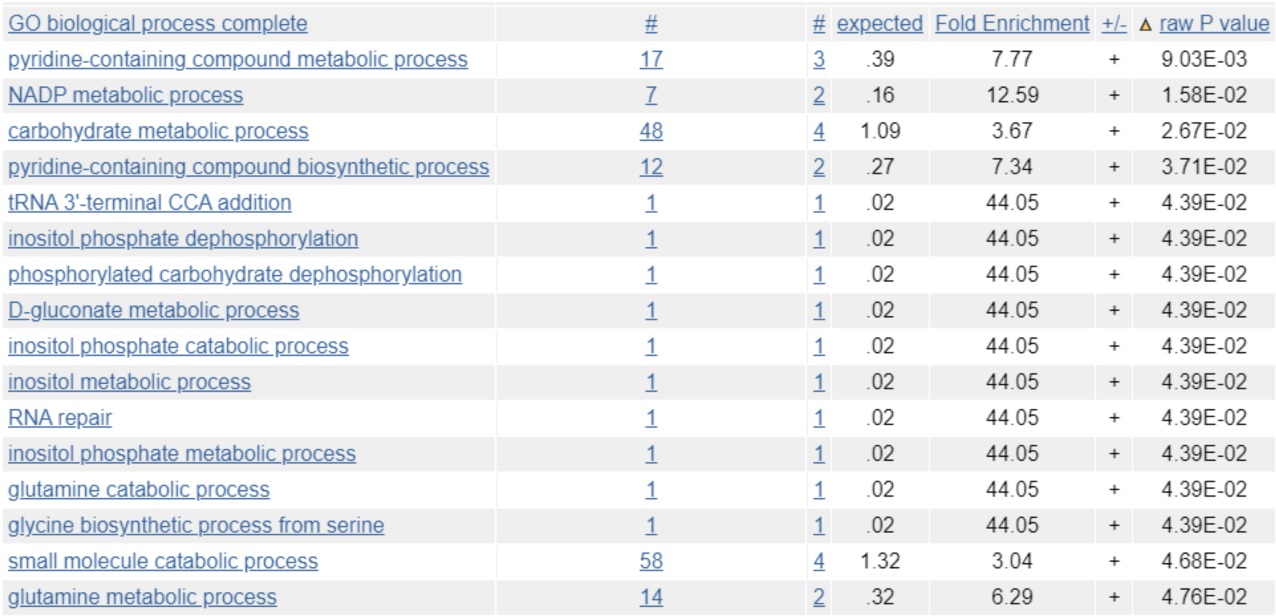
Type A pairs of TSS, one gTSS and other alTSS, were found in 35 genes; Type B were found in 5 genes, Type C were found in 41 genes, and Type D were found in 10 genes in *Halobacterium salinarum* NRC-1 (35 + 5 + 41 + 10 = 91), shown in Table (4-2-1), also can find the full table in Appendix table 2. Important features are registered in the table: the annotation column shows the gTSS/alTSS read counts ratio (or #TSS1/#TSS2) and the gTSS and alTSS positions distance (or TSS1 - TSS2 difference), before and after the pipe mark, respectively. Types B and D always have negative distances because in these situations gTSS is always upstream from alTSS thus at lower genomic coordinate values, as visually depicted by Figure (3-3-1) in the Methods section. As result of our designed filter, alignment signal always have comparable highs between genuine canonical starts and alternative starts: gTSS cannot be bigger than 95% of the smaller alTSS signal (#TSS1/#TSS2 < 0.95).

Table (4-2-1): Arbitrary sample of genes (10 out of 91) that have alTSS in *Halobacterium salinarum* NRC-1 considering all time points from the growth curve into a single collapsed dataset. Rep, is the replicon (chromosome or plasmids); Start and End are the genomic coordinates for the CDS; Strand indicates the DNA *loci* direction; Name shows known aliases for the gene and putative functional annotation; Annotation shows the gTSS-to-alTSS counts ratio and positional distance; Type refers to 1 out of 4 possible gTSS-to-alTSS relative positioning. [Also can find the full table in Appendix 2].

Rep	Start	End	Strand	Name	Annotation	Type
chr	14279	14184	reverse	VNG_RS00060|VNG0016H||MarR family transcriptional regulator|(OE1023R) HTH domain protein	0.85|-95	Type B
chr	67796	67778	reverse	VNG_RS00300|||hypothetical protein|(OE1126R) conserved hypothetical protein	0.93|-18	Type D
chr	84422	84506	reverse	VNG_RS00390|VNG0099G||50S ribosomal protein L16|rpl16 (OE1160R) 50S ribosomal protein L16	0.79|84	Type C
chr	115132	115157	reverse	VNG_RS00560|VNG0137G||tRNA CCA-pyrophosphorylase|cca (OE1222R) tRNA adenylyltransferase, CCA-adding	0.94|25	Type A
chr	144202	144285	reverse	VNG_RS00695|VNG0168H|| hypothetical protein|rpoeps (OE1279R) DNA-directed RNA polymerase epsilon subunit	0.76|83	Type A
chr	175835	175855	forward	VNG_RS00870|VNG0209H|| integrase	0.76|-20	Type D
chr	205708	205780	reverse	VNG_RS01015|VNG0249G|| DUF5059 domain-containing protein|hcpG (OE1391R) DUF5059 domain / halocyanin domain protein HcpG	0.94|72	Type A
chr	208184	208196	reverse	VNG_RS01030|VNG0252C|| geranylgeranylglyceryl/ heptaprenylglyceryl phosphate synthase|pcrB1 (OE1398R) (S)-3-O-geranylgeranylglyceryl phosphate synthase 1	0.76|12	Type C
chr	209438	209292	forward	VNG_RS01040|VNG0255C|| ribonuclease H|rnhA2 (OE1400F) ribonuclease H, type 1	0.64|146	Type A
chr	221649	221634	forward	VNG_RS01110|VNG0277G||FAD-dependent oxidoreductase|OE1426F (OE1426F) probable flavin containing oxidoreductase (homolog to phytoene desaturase / monoamine oxidase)	0.79|15	Type C

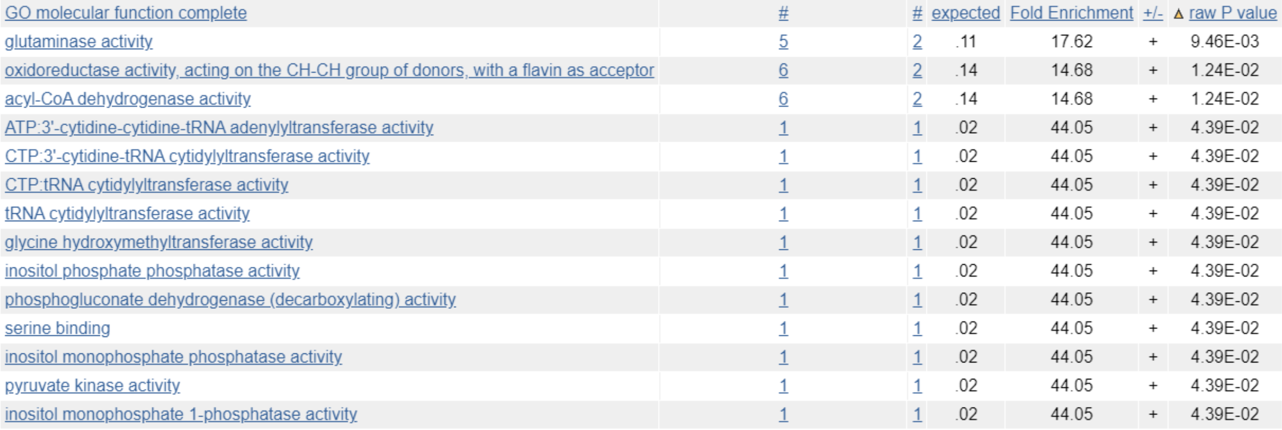
Gene Enrichment Analysis is a powerful tool used to identify biological processes, molecular functions, and cellular components that are significantly enriched in a set of genes of interest. Pantherdb.org is an online resource that provides comprehensive gene annotation and analysis tools for several organisms including *H. salinarum*. The analysis is typically performed using Fisher's exact test to determine the significance of an enrichment. In this case, the results table was created with no multiple corrections, which means that caution should be exercised when interpreting the results. In *H. salinarum* out of 91 genes with alTSSs, a list of 55 genes were recognized and used for enrichment analysis against the reference list of 2423 genes. The following tables, Table (4-2-2), Table (4-2-3) and Table (4-2-3), are showing the GO Biological Process, GO Molecular Function, and GO Cellular Component categories. In Appendix 13, the links for full analysis test were created using the three GO options.

Table (4-2-2): Gene enrichment GO Biological Process using Fisher's exact test with no multiple correction, 55 genes out of 91 submitted genes were used against 2423 genes from the website database.



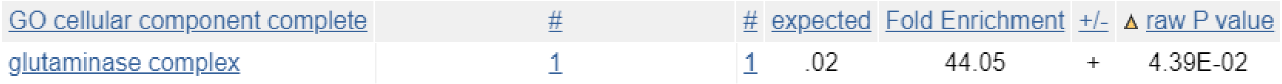
GO biological process complete	#	#	expected	Fold Enrichment	+/-	▲ raw P value
pyridine-containing compound metabolic process	17	3	.39	7.77	+	9.03E-03
NADP metabolic process	7	2	.16	12.59	+	1.58E-02
carbohydrate metabolic process	48	4	1.09	3.67	+	2.67E-02
pyridine-containing compound biosynthetic process	12	2	.27	7.34	+	3.71E-02
tRNA 3'-terminal CCA addition	1	1	.02	44.05	+	4.39E-02
inositol phosphate dephosphorylation	1	1	.02	44.05	+	4.39E-02
phosphorylated carbohydrate dephosphorylation	1	1	.02	44.05	+	4.39E-02
D-gluconate metabolic process	1	1	.02	44.05	+	4.39E-02
inositol phosphate catabolic process	1	1	.02	44.05	+	4.39E-02
inositol metabolic process	1	1	.02	44.05	+	4.39E-02
RNA repair	1	1	.02	44.05	+	4.39E-02
inositol phosphate metabolic process	1	1	.02	44.05	+	4.39E-02
glutamine catabolic process	1	1	.02	44.05	+	4.39E-02
glycine biosynthetic process from serine	1	1	.02	44.05	+	4.39E-02
small molecule catabolic process	58	4	1.32	3.04	+	4.68E-02
glutamine metabolic process	14	2	.32	6.29	+	4.76E-02

Table (4-2-3): Gene enrichment GO Molecular Function using Fisher's exact test with no multiple correction, 55 genes out of 91 submitted genes were used against 2423 genes from the website database.



GO molecular function complete	#	#	expected	Fold Enrichment	+/-	▲ raw P value
glutaminase activity	5	2	.11	17.62	+	9.46E-03
oxidoreductase activity, acting on the CH-CH group of donors, with a flavin as acceptor	6	2	.14	14.68	+	1.24E-02
acyl-CoA dehydrogenase activity	6	2	.14	14.68	+	1.24E-02
ATP:3'-cytidine-cytidine-tRNA adenylyltransferase activity	1	1	.02	44.05	+	4.39E-02
CTP:3'-cytidine-tRNA cytidylyltransferase activity	1	1	.02	44.05	+	4.39E-02
CTP:tRNA cytidylyltransferase activity	1	1	.02	44.05	+	4.39E-02
tRNA cytidylyltransferase activity	1	1	.02	44.05	+	4.39E-02
glycine hydroxymethyltransferase activity	1	1	.02	44.05	+	4.39E-02
inositol phosphate phosphatase activity	1	1	.02	44.05	+	4.39E-02
phosphogluconate dehydrogenase (decarboxylating) activity	1	1	.02	44.05	+	4.39E-02
serine binding	1	1	.02	44.05	+	4.39E-02
inositol monophosphate phosphatase activity	1	1	.02	44.05	+	4.39E-02
pyruvate kinase activity	1	1	.02	44.05	+	4.39E-02
inositol monophosphate 1-phosphatase activity	1	1	.02	44.05	+	4.39E-02

Table (4-2-4): Gene enrichment GO Celullar Componets using Fisher's exact test with no multiple correction, 55 genes out of 91 submitted genes were used against 2423 genes from the website database.



GO cellular component complete	#	#	expected	Fold Enrichment	+/-	▲ raw P value
glutaminase complex	1	1	.02	44.05	+	4.39E-02

One of the main contributions of this PhD work is to consider the time-course experiment derived from the growth curve as separate data points. Therefore we performed the same analysis, the search for the four types of gTSS-alTSS pairs, in each time point dataset separately. A list of **292** genes with alTSSs, using the same parametric setup used before, was created for separate libraries (17h, 37h, 86h, REF2014 and REF2015). The Table (4-2-2) shows an arbitrary sample of the complete list, that can find in Appendix 3.

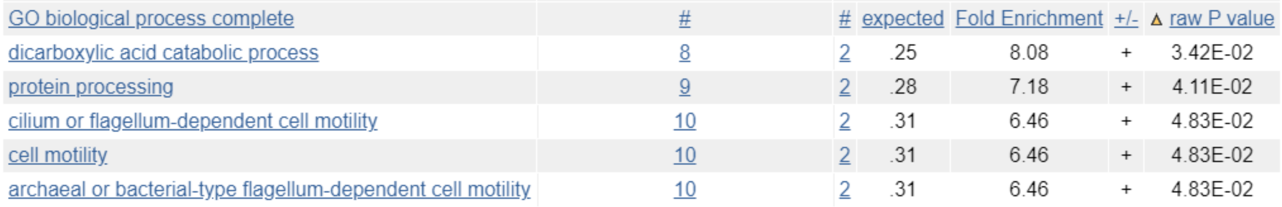
Table (4-2-5): Arbitrary sample of genes (10 out of 292) that have alTSS in *Halobacterium salinarum* NRC-1 considering time points from the growth curve separately: 17h, 37h, 86h, reference condition replicate 1 (REF2014) and replicate 2 (REF2015). Rep, is the replicon (chromosome or plasmids); Start and End are the genomic coordinates for the CDS; Strand indicates the DNA *loci* direction; Name shows known aliases for the gene and putative functional annotation; Annotation shows the gTSS-to-alTSS counts ratio and positional distance; Type refers to 1 out of 4 possible gTSS-to-alTSS relative positioning. [Also can find the full table in Appendix 3].

Rep	**Start**	**End**	**Strand**	**Name**	**Annotation**	**Type**	**Time**
chr	14279	14184	reverse	VNG_RS00060|VNG0016H||MarR family transcriptional regulator|(OE1023R) HTH domain protein	0.85|-95	Type B	REF2014
chr	14279	14184	reverse	VNG_RS00060|VNG0016H||MarR family transcriptional regulator|(OE1023R) HTH domain protein	0.85|-95	Type B	17h
chr	14279	14184	reverse	VNG_RS00060|VNG0016H||MarR family transcriptional regulator|(OE1023R) HTH domain protein	0.85|-95	Type B	37h
chr	14279	14184	reverse	VNG_RS00060|VNG0016H||MarR family transcriptional regulator|(OE1023R) HTH domain protein	0.85|-95	Type B	86h
chr	23271	23335	forward	VNG_RS00110|VNG0028C||transposase|(OE1045F) homolog to ISH16-type transposase	0.93|-64	Type D	37h
chr	23271	23335	forward	VNG_RS00110|VNG0028C||transposase|(OE1045F) homolog to ISH16-type transposase	0.93|-64	Type B	86h
chr	31838	31742	forward	VNG_RS00135|VNG0037H||type II toxin-antitoxin system HicB family antitoxin|(OE1060F)	0.94|96	Type A	REF2015

				conserved hypothetical protein			
chr	35931	36077	reverse	VNG_RS00160|VNG0042G||transposase|(OE1070R) IS1341-type transposase ISH39	0.59|146	Type A	17h
chr	41880	41794	forward	VNG_RS13205|VNG0049H,VNG_0049H||FkbM family methyltransferase|(OE1079F) FkbM family methyltransferase	0.72|86	Type A	17h
chr	41880	41794	forward	VNG_RS13205|VNG0049H,VNG_0049H||FkbM family methyltransferase|(OE1079F) FkbM family methyltransferase	0.72|86	Type A	37h

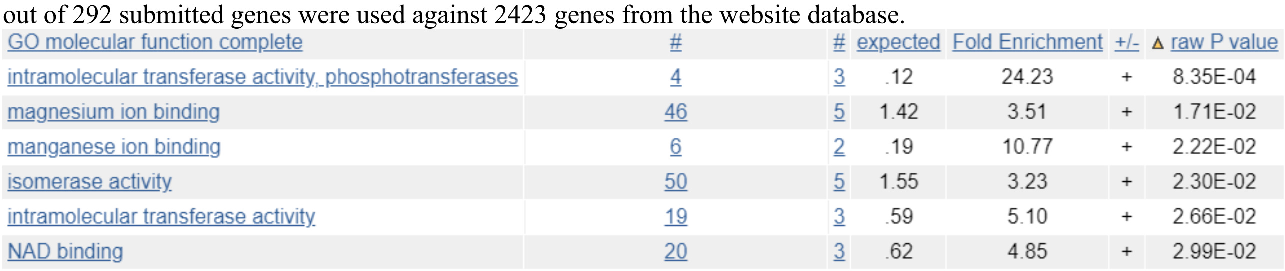
Also for that 292 genes a list of 75 genes were used for enrichment analysis against the reference list of 2423 genes by the same tool Pantherdb.org. The following Tables (4-2-6) and Table (4-2-7) are showing the GO Biological Process, and GO Molecular Function. The GO Cellular Component analysis has no statistically significant results. In Appendix 13, the links for full analysis test were created using the three GO options.

Table (4.2.6): Gene enrichment GO Biological Process using Fisher's exact test with no multiple correction, 75 genes out of 292 submitted genes were used against 2423 genes from the website database.



GO biological process complete	#	#	expected	Fold Enrichment	+/-	▲ raw P value
dicarboxylic acid catabolic process	8	2	.25	8.08	+	3.42E-02
protein processing	9	2	.28	7.18	+	4.11E-02
cilium or flagellum-dependent cell motility	10	2	.31	6.46	+	4.83E-02
cell motility	10	2	.31	6.46	+	4.83E-02
archaeal or bacterial-type flagellum-dependent cell motility	10	2	.31	6.46	+	4.83E-02

Table (4.2.7): Gene enrichment GO Molecular Function using Fisher's exact test with no multiple correction, 75 genes out of 292 submitted genes were used against 2423 genes from the website database.



GO molecular function complete	#	#	expected	Fold Enrichment	+/-	▲ raw P value
intramolecular transferase activity, phosphotransferases	4	3	.12	24.23	+	8.35E-04
magnesium ion binding	46	5	1.42	3.51	+	1.71E-02
manganese ion binding	6	2	.19	10.77	+	2.22E-02
isomerase activity	50	5	1.55	3.23	+	2.30E-02
intramolecular transferase activity	19	3	.59	5.10	+	2.66E-02
NAD binding	20	3	.62	4.85	+	2.99E-02

There are only few genes pointed using the time-course approach that are also pointed by the collapsing time-points approach, and vice-versa: 90 genes are exclusive to the time-course approach, 74 genes to the collapsing approach and 17 genes were found in both. This highlights the importance of the additional analysis approach we implemented in this PhD thesis.

The switching in TSS and their proportion caused naturally by growth-curve moments and could be affected by growth conditions or the change in their level in the media. Multiple TSS per gene and growth moment TSS indicate a progressive transcriptome in *Halobacterium*, therefore implementing a new way for exploring and revealing the 5'UTR lengths, features and role for regulation in addition to the interaction between regulation factors. An example of 5'UTR changes is shown in Figure (4-2-1) as a screenshot of *H. salinarum* NRC-1 Atlas (https://halodata.systemsbiology.net/) using the mapped data of dRNA-seq reads, genome and identified TSS signals.

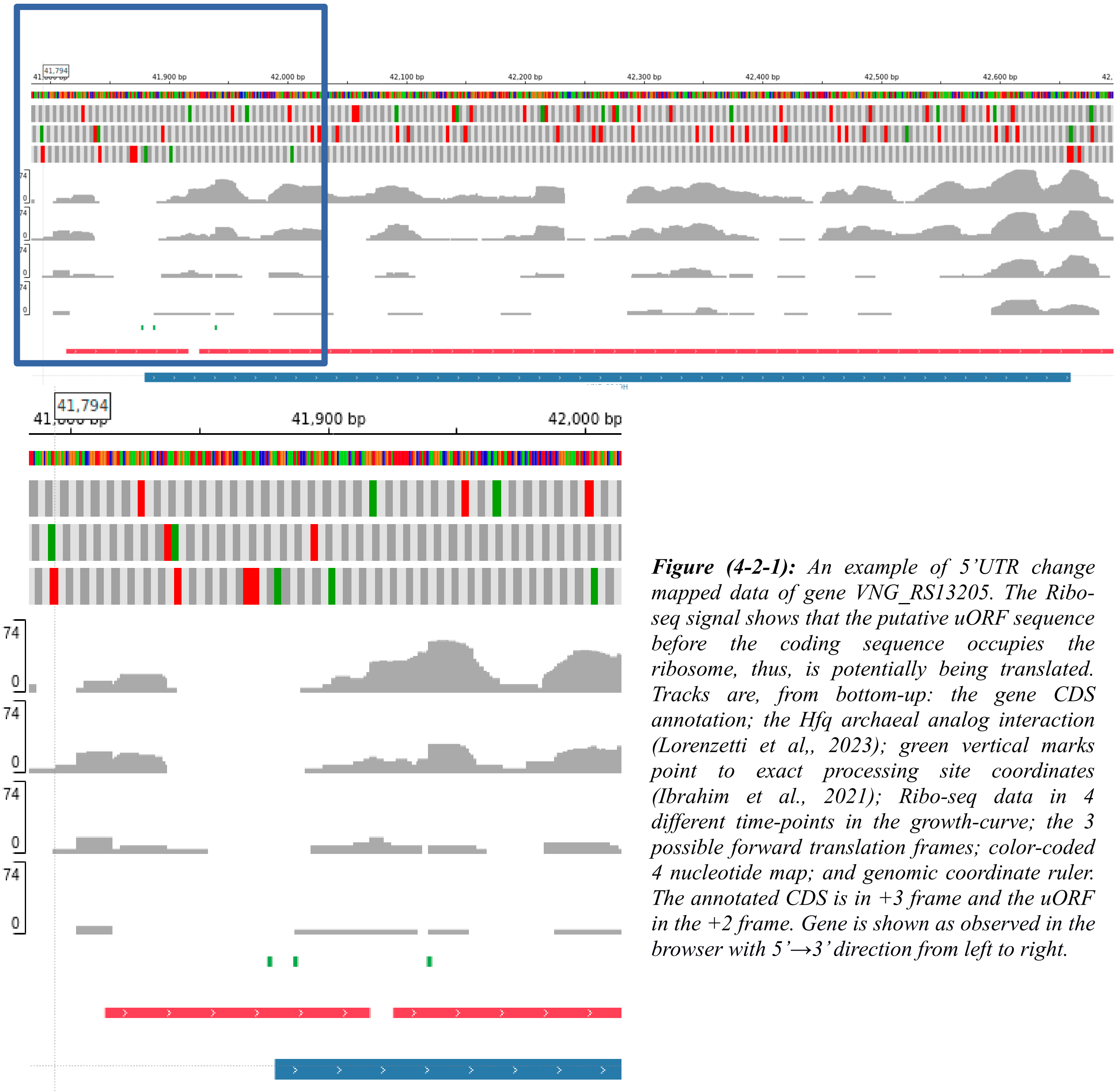


Figure (4-2-1): *An example of 5’UTR change mapped data of gene VNG_RS13205. The Ribo-seq signal shows that the putative uORF sequence before the coding sequence occupies the ribosome, thus, is potentially being translated. Tracks are, from bottom-up: the gene CDS annotation; the Hfq archaeal analog interaction (Lorenzetti et al,, 2023); green vertical marks point to exact processing site coordinates (Ibrahim et al., 2021); Ribo-seq data in 4 different time-points in the growth-curve; the 3 possible forward translation frames; color-coded 4 nucleotide map; and genomic coordinate ruler. The annotated CDS is in +3 frame and the uORF in the +2 frame. Gene is shown as observed in the browser with 5’→3’ direction from left to right.*

From all the genes for which we detected alTSS, the most interesting one is at *loci* VNG_RS09715, named *pst1* (alternatively *yqgG* or OE4485R/VNG2486G) encoding for a ABC-type transport system periplasmic substrate-binding protein which substrate is phosphate (Pi). This gene was studied before in the R1 strain of *H. salinarum* and it was shown by classical laboratory methods that it has two possible mRNA isoforms (Furtwängler et al., 2010), shown in Figure (4-2-2) identical to the published one.

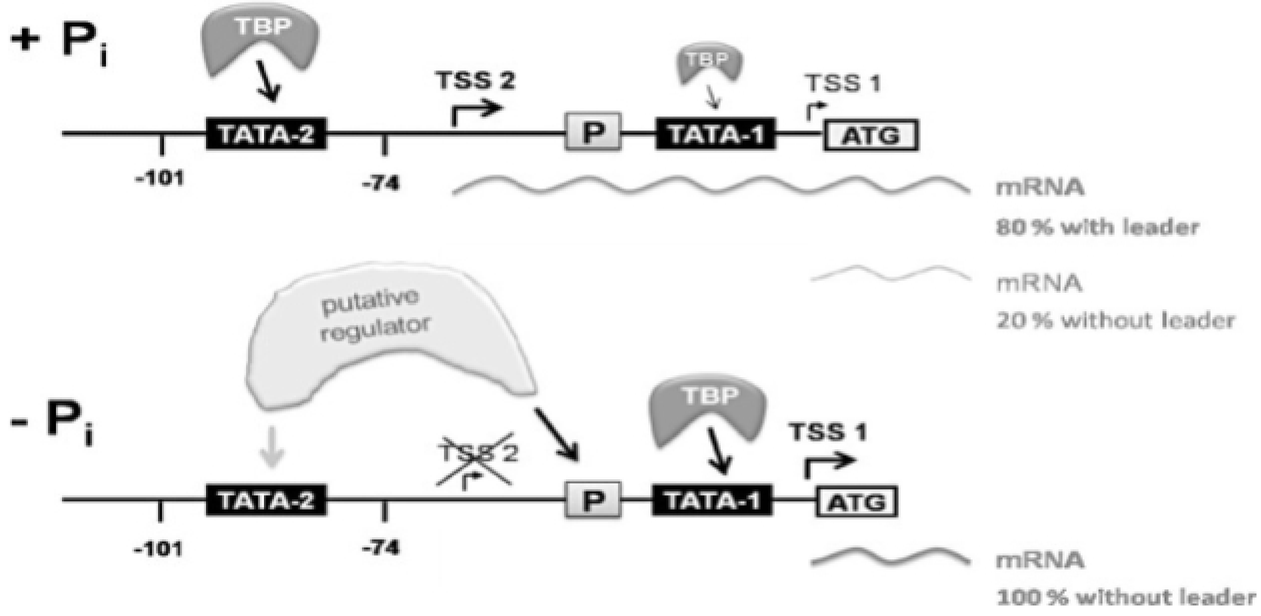


Figure (4-2-2): *Model for transcription regulation of pst1 operon. Under phosphate-saturated conditions (+Pi), TBP (TATA-binding protein) binds both promoters and the alTSS (TSS2) offering a 5'UTR includes a new binding site and transcription start. Under phosphate-limited conditions (−Pi) transcription initiates only from TSS-1. Taken from (Furtwangler et al., 2010). Gene is shown with 5'→3' direction from left to right.*

Our analysis recapitulates this observation made in two "extreme" lab conditions of lack and excess of Pi and generalize it to more close to natural growth-curve conditions, as shown in Figure (4-2-3). Switching in TSS usage depending on phosphate uptake via Pst operons (*pst1* and *pst2*) in *H. salinarum* R1 was shown before (Furtwängler et al., 2010). These ten-years-old results were obtained in a closely related strain R1 which differs from NRC-1 practically only on plasmids. It was shown that translation efficiency is markedly different between 5'UTR containing and leaderless mRNA isoforms. Our results show that there is change in TSS usage over time, as show in Figure (4-2-4), and not only with Pi concentration, although it probably reflects Pi abundance over the growth-curve.

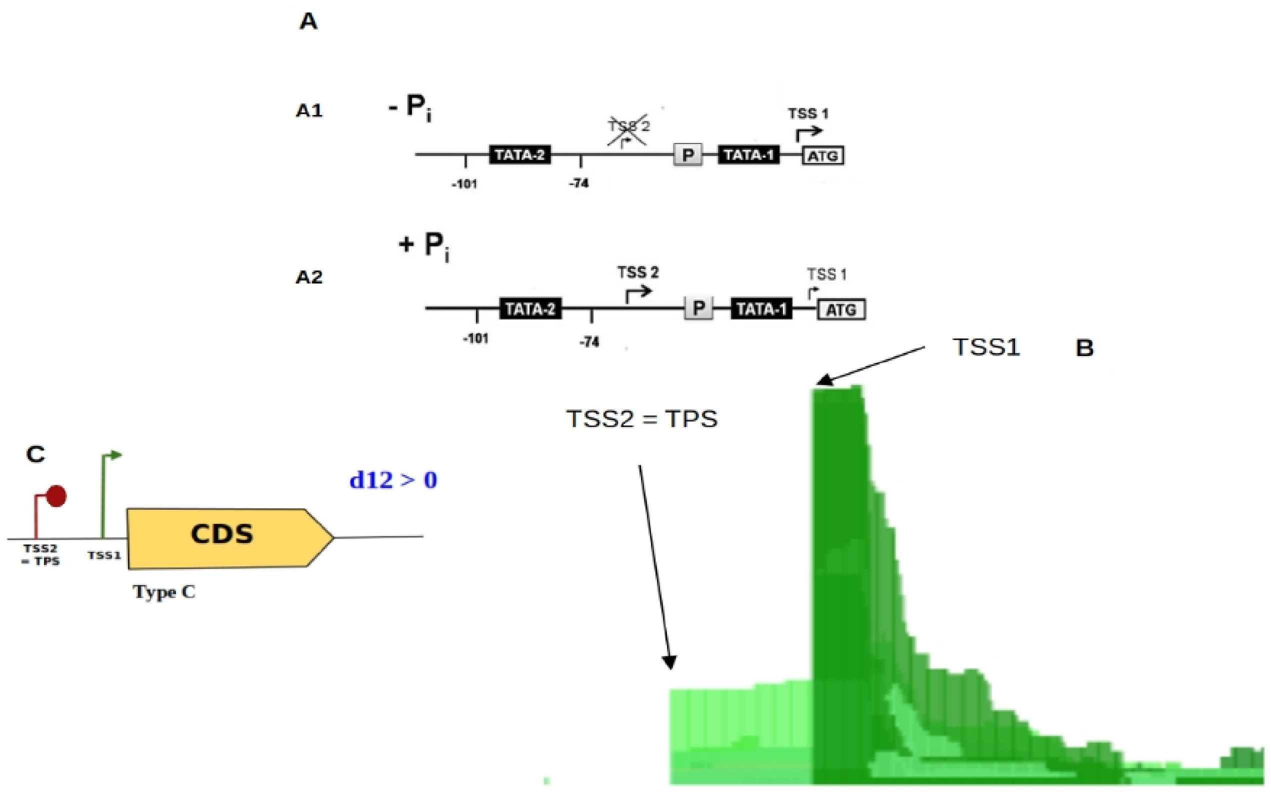


Figure (4-2-3): *Explanation of why the RNA transcription in the Phosphate uptake regulator gene (pst1) has two different TSS discovered in the study (Furtwangler et al., 2010). A: RNA transcription regulation of pst1 (Figure was taken and edited from Furtwangler et al., 2010), A1: RNA transcription under phosphate-limited conditions (−Pi) transcription initiates only from TSS-1. A2: RNA transcription under phosphate-saturated conditions (+Pi), the alTSS (TSS2) offering an mRNA includes a new binding site and transcription start. B: A screenshot of the Halobacterium salinarum NRC-1 genome browser where the Phosphate uptake regulator gene (pst1) shows the dRNA-seq reads for the same gene, dark green bars are the TEX+ library and light green bars are TEX- aligned reads. C: The illustration of {gTSS, alTSS} pairs Type C as explained in Figure (3-3-1). Gene is shown as observed in the browser with 5'→3' direction from left to right.*

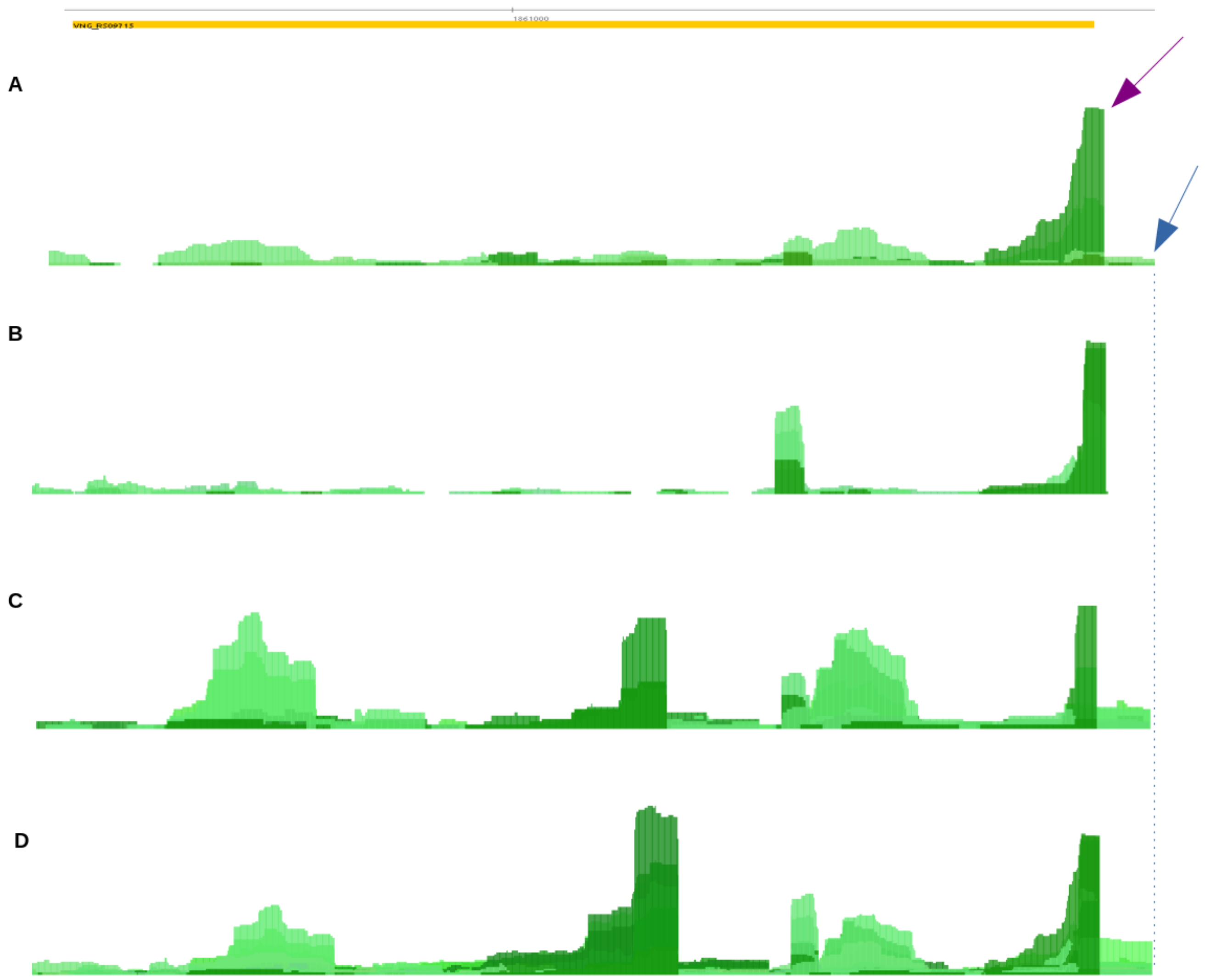


Figure (4-2-4): *TSS mapping of the Phosphate uptake regulator gene (pst1) at different time points. A: RNA transcription after 17 hours of growth. B: RNA transcription after 37 hours of growth. C: RNA transcription after 86 hours of growth D: RNA transcription of libraries was used as references (REF) for the data creation. Purple arrow points to the primary genuine gTSS (TSS1) and the blue arrow points to the secondary alternative alTSS (TSS2). Gene is shown as observed in the browser with 5'→3' direction from right to left, mirrored relative to Figure (4-2-3).*

We claim that this specific result validates our general approach to find alTSS and thus gives confidence to the overall list of genes. At the same time, we decided to use the same approach in some other organisms for which dRNA-seq is publicly available.

4.3 alTSS in Other Organisms

We are aware that the research community in archaea is much smaller than the research community that studies bacteria for a number of reason. Although our methodology was primarily developed to answer biological questions in *H. salinarum*, and not as a general bioinformatics driven effort, we recognize that the application of everything developed her is straightforward. Therefore we applied our pipeline to other organisms in order to provide these research communities with a list of putative alternative alTSS and the visualization tools to explore them.

We arbitrarily chose 2 bacterial organisms for which high-quality dRNA-seq were available and all 4 archaeal organisms for which dRNA-seq experiments were published (circa Dec/2019): *Caulobacter crescentus* NA1000; *Thermus thermophilus* HB8; *Methanocaldococcus jannaschii* DSM 2661; *Haloferax volcanii* DS2; *Thermococcus onnurineus* NA1.

Alternative TSS were studied and analyzed the switching in TSS usage by different growth conditions. The following tables show 4 types of {gTSS, alTSS} pairs, A, B, C or D, as explained in Figure (3-3-1), everything using the same criteria and pipelines developed for *H. salinarum*, our main organism of interest.

Table (4-3-1): List of 10 example genes in *Caulobacter crescentus* NA1000 where located alTSS in different types. The terms: Rep, is the replicon (RefSeq number); Start and End are the genomic coordinates for the CDS; Strand indicates the DNA *loci* direction; Protein Name shows known aliases for the putative functional annotation; Annotation shows the gTSS-to-alTSS counts ratio and positional distance; Type refers to 1 out of 4 possible gTSS-to-alTSS relative positioning; The sample condition "bulk condition" indicates that the dRNA-seq data sample was collected without specific individual conditions but rather as a collective sample. [Also can find the full table in Appendix 4].

Rep	**Start**	**End**	**Strand**	**Protein.Name**	**Annotation**	**Type**	**Condition**
NC_011916.1	20755	20766	reverse	|YP_002515395.1| hypothetical protein	0.5|11	Type A	Bulk
NC_011916.1	73680	73616	forward	|YP_002515447.2| hypothetical protein	0.54|64	Type A	Bulk
NC_011916.1	162409	162393	forward	|YP_002515528.1|adenine nucleotide exchange factor GrpE	0.61|16	Type A	Bulk
NC_011916.1	180200	180184	forward	|YP_002515543.2| polysaccharide export protein	0.68|16	Type A	Bulk
NC_011916.1	244691	244725	reverse	|YP_002515603.3|putative membrane spanning protein	0.55|34	Type A	Bulk
NC_011916.1	678985	678949	forward	|YP_002516005.1| chemotaxis receiver domain protein CheYII	0.56|36	Type C	Bulk
NC_011916.1	772889	772904	reverse	|YP_009020502.1| hypothetical protein	0.61|15	Type A	Bulk
NC_011916.1	938924	938935	reverse	|YP_002516233.3|holdfast inhibitor protein HfiA	0.55|11	Type C	Bulk
NC_011916.1	1084026	1084015	reverse	|YP_002516374.1|riboki-se	0.51|-11	Type D	Bulk
NC_011916.1	1209970	1209981	reverse	|YP_002516476.1|ADP-heptose--LPS heptosyltransferase	0.51|11	Type A	Bulk

Table (4-3-2): List of 10 example genes in *Haloferax volcanii* DS2 where 934 genes with alTSS in different types. The terms: Rep, is the replicon (RefSeq number); Start and End are the genomic coordinates for the CDS; Strand indicates the DNA *loci* direction; Protein Name shows known aliases for the putative functional annotation; Annotation shows the gTSS-to-alTSS counts ratio and positional distance; Type refers to 1 out of 4 possible gTSS-to-alTSS relative positioning; The dRNA-seq data condition was specified according to the lab where the organism has been cultured, "bulk condition" indicates that the dRNA-seq data sample was collected without specific individual conditions but rather as a collective sample, “Frankfurt” indicates the lab in Frankfurt city, “Ulm” indicates the lab in Ulm city, and “Wurzburg” indicates the lab in Wurzburg city. [Also can find the full table in Appendix 5].

Rep	**Start**	**End**	**Strand**	**Protein.Name**	**Annotation**	**Type**	**Location**
NC_013964.1	1545	1415	forward	WP_004041151.1| acetylornithine deacetylase	0.79|130	Type C	Bulk
NC_013964.1	43049	43195	reverse	WP_004041116.1|D-xylonate dehydratase	0.93|146	Type A	Bulk
NC_013964.1	74645	74526	forward	WP_004041088.1|Lrp/AsnC family transcriptional regulator	0.95|119	Type A	Ulm
NC_013964.1	90692	90749	reverse	WP_004041075.1|ABC transporter permease	0.77|57	Type A	Bulk
NC_013964.1	114882	114763	forward	WP_013034983.1|glycosyl hydrolase family 88	0.87|119	Type A	Frankfurt
NC_013964.1	123906	123919	forward	WP_004041045.1|alcohol dehydrogenase	0.94|-13	Type D	Wurzburg
NC_013964.1	127509	127480	forward	WP_004041042.1|IclR family transcriptional regulator	0.82|29	Type C	Frankfurt
NC_013964.1	149406	149475	reverse	WP_004041023.1| hypothetical protein	0.78|69	Type A	Ulm
NC_013964.1	165626	165595	forward	WP_004041008.1|oleate hydratase	0.81|31	Type A	Bulk
NC_013964.1	171253	171299	reverse	WP_013034963.1|ISH3 family transposase	0.67|46	Type C	Frankfurt

Table (4-3-3): List of 10 example genes in *Methanocaldococcus jannaschii* DSM 2661 where 250 genes with alTSS in different types. The terms: Rep, is the replicon (RefSeq number); Start and End are the genomic coordinates for the CDS; Strand indicates the DNA *loci* direction; Protein Name shows known aliases for the putative functional annotation; Annotation shows the gTSS-to-alTSS counts ratio and positional distance; Type refers to 1 out of 4 possible gTSS-to-alTSS relative positioning; The dRNA-seq data condition was specified as two replications of optimal growth culture, "rep.01" indicates the first dRNA-seq data sample replicate, "rep.02" indicates the second dRNA-seq data sample replicate. [Also can find the full table in Appendix 6].

Rep	**Start**	**End**	**Strand**	**Protein.Name**	**Annotation**	**Type**	**Sample**
NC_000909.1	4154	4178	forward	transporter	0.72|-24	Type B	rep.02
NC_000909.1	7250	7265	reverse	WP_064496362.1|formate dehydrogenase	0.59|15	Type A	rep.01
NC_000909.1	16804	16744	forward	WP_010869509.1|IS6 family transposase	0.72|60	Type A	rep.01
NC_000909.1	25222	25262	forward	WP_083774557.1| hypothetical protein	0.85|-40	Type B	rep.01
NC_000909.1	39994	39972	reverse	WP_064496373.1|DNA-directed RNA polymerase subunit F	0.6|-22	Type B	rep.01
NC_000909.1	42749	42734	forward	WP_010869534.1| TIGR00375 family protein	0.56|15	Type A	rep.02
NC_000909.1	50591	50495	reverse	WP_064496895.1|50S ribosomal protein L31e	0.52|-96	Type B	rep.01
NC_000909.1	55839	55893	reverse	WP_010869547.1|3,4-dihydroxy-2-butanone-4-phosphate synthase	0.74|54	Type A	rep.01
NC_000909.1	60257	60246	forward	WP_010869553.1|ferredoxin	0.67|11	Type C	rep.01
NC_000909.1	61820	61840	forward	WP_064496382.1| hypothetical protein	0.73|-20	Type D	rep.01

Table (4-3-4): List of 10 example genes in *Thermococcus onnurineus* NA1 where located 121 genes with alTSS in different types. The terms: Rep, is the replicon (RefSeq number); Start and End are the genomic coordinates for the CDS; Strand indicates the DNA *loci* direction; Protein Name shows known aliases for the putative functional annotation; Annotation shows the gTSS-to-alTSS counts ratio and positional distance; Type refers to 1 out of 4 possible gTSS-to-alTSS relative positioning; The dRNA-seq data condition was specified as two replications of optimal growth culture, "rep.01" indicates the first dRNA-seq data sample replicate, "rep.02" indicates the second dRNA-seq data sample replicate. [Also can find the full table in Appendix 7].

Rep	**Start**	**End**	**Strand**	**Protein.Name**	**Annotation**	**Type**	**Sample**
NC_011529.1	27498	27620	forward	WP_012570987.1|exosome complex exonuclease Rrp41	0.71|-122	Type B	rep.02
NC_011529.1	32404	32421	reverse	WP_012570993.1| homoserine dehydrogenase	0.56|17	Type A	rep.02
NC_011529.1	55604	55715	reverse	WP_012571017.1| hypothetical protein	0.9|111	Type A	rep.01
NC_011529.1	65572	65557	forward	WP_012571030.1|50S ribosomal protein L29	0.78|15	Type A	rep.01
NC_011529.1	65730	65824	forward	WP_012571031.1| translation initiation factor	0.56|-94	Type B	rep.02
NC_011529.1	67442	67421	forward	WP_012571036.1|30S ribosomal protein S4e	0.65|21	Type A	rep.02
NC_011529.1	70718	70763	forward	WP_012571043.1|50S ribosomal protein L18	0.64|-45	Type B	rep.02
NC_011529.1	72527	72564	forward	WP_012571046.1|50S ribosomal protein L15	0.53|-37	Type B	rep.02
NC_011529.1	76580	76555	forward	WP_012571052.1|50S ribosomal protein L14e	0.64|25	Type A	rep.02
NC_011529.1	85941	85846	forward	WP_012571065.1|50S ribosomal protein L13	0.65|95	Type A	rep.02

Table (4-3-5): List of 10 example genes in *Thermus thermophilus* HB8 where 238 genes with alTSS in different types. The terms: Rep, is the replicon (RefSeq number); Start and End are the genomic coordinates for the CDS; Strand indicates the DNA *loci* direction; Protein Name shows known aliases for the putative functional annotation; Annotation shows the gTSS-to-alTSS counts ratio and positional distance; Type refers to 1 out of 4 possible gTSS-to-alTSS relative positioning. The dRNA-seq data condition was specified by the length of RNA size regions used in defining proteins, "bulk condition" indicates that the dRNA-seq data sample was collected without specific individual conditions but rather as a collective sample, RNA_size refers to cDNA libraries used for sample collection in size ranges of “Short”19-30 nt, “Medium” 35-50 nt, and “Long” 50-100. [Also can find the full table in Appendix 8].

Rep	**Start**	**End**	**Strand**	**Protein.Name**	**Annotation**	**Type**	**RNA_size**
NC_006461.1	1221	1235	forward	|BAD69825.1|enolase (2-phosphoglycerate dehydratase)	0.62|-14	Type B	Long
NC_006461.1	8712	8699	reverse	|BAD69831.1|phage shock protein A	0.94|-13	Type D	Bulk
NC_006461.1	14280	14269	forward	|BAD69836.1| geranylgeranyl diphosphate synthetase	0.79|11	Type C	Long
NC_006461.1	38376	38365	forward	|BAD69860.1|hypothetical protein	0.68|11	Type A	Bulk
NC_006461.1	53043	53056	reverse	|BAD69877.1|R-methyltransferase	0.6|13	Type A	Bulk
NC_006461.1	57257	57246	forward	|BAD69882.1|molybdenum cofactor biosynthesis protein D/E	0.93|11	Type C	Medium
NC_006461.1	57875	57850	forward	|BAD69883.1|hypothetical protein	0.9|25	Type C	Bulk
NC_006461.1	66629	66535	forward	|BAD69894.1|ABC-transporter, ATP-binding subunit	0.94|94	Type C	Medium
NC_006461.1	67430	67326	forward	|BAD69895.1|conserved hypothetical protein	0.69|104	Type C	Medium
NC_006461.1	78135	78022	forward	|BAD69899.1| phosphoribosylanthranilate isomerase	0.93|113	Type A	Medium

The alTSS were detected in the aforementioned organisms as response to changing in growth conditions. The datasets presented above are a re-analysis of dRNA-seq data from published papers in which alTSS were not the focus. Meanwhile, there are works designed to study alTSS but with no datasets publicly available, therefore not considered in this PhD thesis. One of such cases is (Li et al., 2015), where the archaeal model *Methanolobus psychrophilus* was examined using two degrees 18ºC and 8ºC to understand cold-shock and adaptation.

Just to illustrate the usefulness of the resource created we selected a single gene from one of these organisms non-central to our group, the model organism *Caulobacter crescentus* NA1000 (currently also known as *Caulobacter vibrioides*), and explored an interesting bioinformatics derived hypothesis. The well-characterized gene *ftsZ* (CCNA_02623 *loci*), encoding a very important cell division protein, had a relatively long 5'UTR that latter revealed to host an uORF (CCNA_03971). Ribo-seq evidence supports that this uORF is actually translated to a small protein (https://www.ncbi.nlm.nih.gov/gene/7332973), turning the long 5'UTR into a bicistronic operon with genes very different in terms of size, as shown in Figure (4-3-1). Our alTSS analysis suggests that, like the conditional operon behavior we studied before illustrated by Figure (1-6), there is a *ftsZ*-only transcript and a CCNA_03971-*ftsZ* transcript. Figure (4-3-1) shows that the alTSS detected still yield a 5'UTR truly untranslated. Both transcript isoforms could coexist and perhaps be regulated differentially by functional reasons that would be eventually elucidated when the small protein encoded at CCNA_03971 is experimentally characterized.

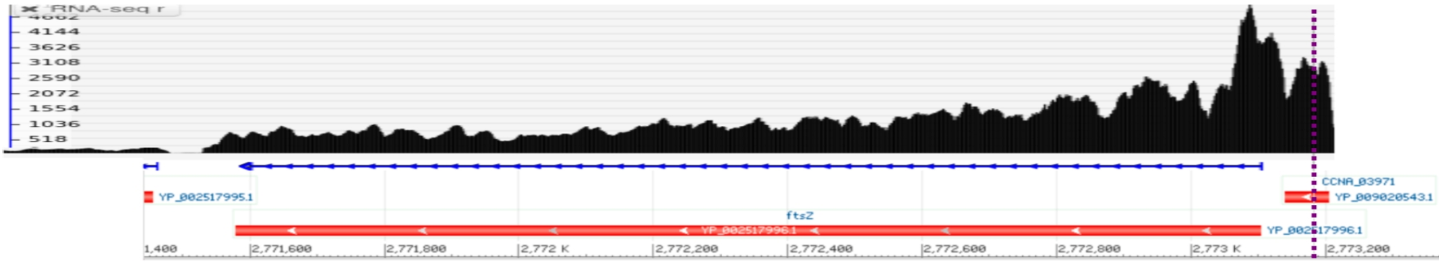
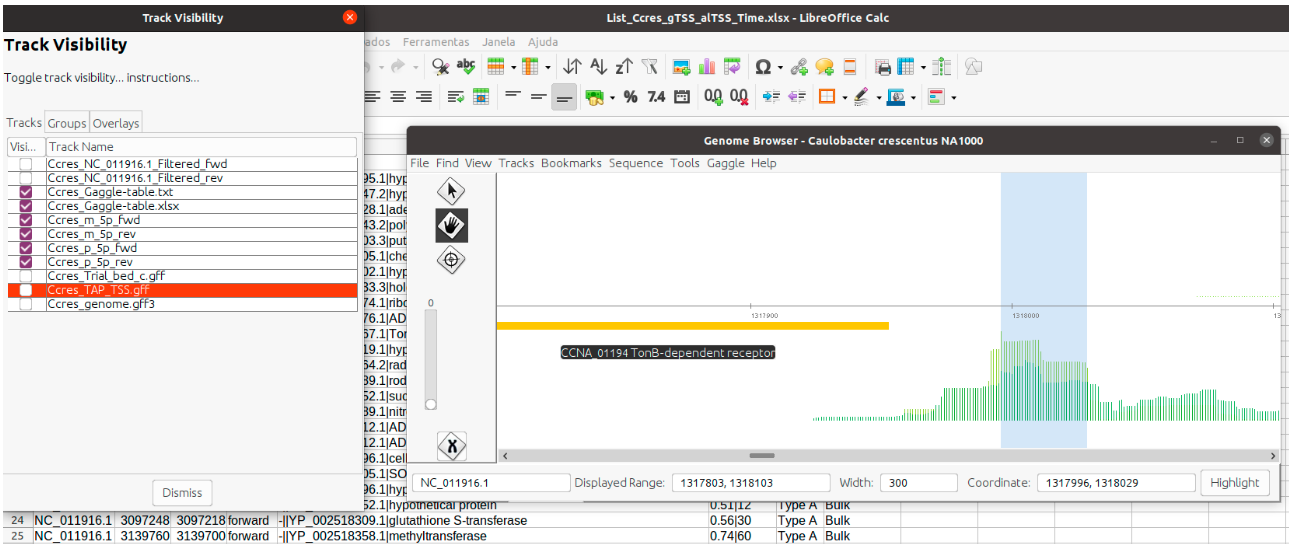


Figure (4-3-1): *alTSS mapping of Caulobacter crescentus NA1000 shows that there is a transcript isoform that do not have an uORF. The schematic is a merge of CauloBrowser data (*http://web.stanford.edu/group/golden_gate_clon/cgi-bin/genome_info_browser/html/genome_info_browser.py*) and NCBI's gene webpage (*https://www.ncbi.nlm.nih.gov/gene/7332973*) put to the same scale. CauloBrowser RNA-seq data shows the established genuine gTSS and NCBI's annotation shows the small protein that makes the 5'UTR no so UTR after all. The alTSS is marked by the dotted purple vertical line. Gene is shown with 5'→3' direction from right to left.*

It is beyond the scope of the present PhD work to explore the implications of the alTSS found for other than *H. salinarum* but we made all the data available along with GGB files to allow proper visualization, like an example in Figure (4-3-2) in Appendix 9.


Track Visibility
List_Ccres_gTSS_alTSS_Time.xlsx - LibreOffice Calc
ados Ferramentas Janela Ajuda
Genome Browser - Caulobacter crescentus NA1000
File Find View Tracks Bookmarks Sequence Tools Gaggle Help
CCNA_01194 TonB-dependent receptor
1317900
1318000
Dismiss
Highlight

Track Visibility

Toggle track visibility... instructions...

Tracks | Groups | Overlays

Visi...	Track Name
☐	Ccres_NC_011916.1_Filtered_fwd
☐	Ccres_NC_011916.1_Filtered_rev
☑	Ccres_Gaggle-table.txt
☑	Ccres_Gaggle-table.xlsx
☑	Ccres_m_5p_fwd
☑	Ccres_m_5p_rev
☑	Ccres_p_5p_fwd
☑	Ccres_p_5p_rev
☐	Ccres_Trial_bed_c.gff
☐	Ccres_TAP_TSS.gff
☐	Ccres_genome.gff3

NC_011916.1 Displayed Range: 1317803, 1318103 Width: 300 Coordinate: 1317996, 1318029

					52.1|hypothetical protein	0.51|12	Type A	Bulk
24	NC_011916.1	3097248	3097218	forward	-||YP_002518309.1|glutathione S-transferase	0.56|30	Type A	Bulk
25	NC_011916.1	3139760	3139700	forward	-||YP_002518358.1|methyltransferase	0.74|60	Type A	Bulk

Figure (4-3-2): *Example of a analysis section using* Caulobacter crescentus *NA1000 data. Screenshot shows alTSS table, the GGB controls to display and hide tracks, and the arbitrary example TonB-dependent receptor gene sucA (CCNA_01194 loci). The highlighted blue selection shows a 34bp apart alternative TSSs that may have different regulatory properties. Light green profile is the TEX+ library alignment and the dark green is the TEX- library. Gene is shown as observed in the browser with 5'→3' direction from right to left.*

4.4 Internal TSSs (iTSS) and fake TSS (fTSS)

We defined one of the main conceptual objects of this PhD work, the alTSS, as positions that appear outside (upstream) of a coding sequence, as depicted before in Figure (3-3-1) and Figure (1-10). However, its is possible to obtain a perfectly valid alternative TSS for a gene if it is located downstream to the genes start codon, inside the gene. In fact, those cases would have even more dramatic consequences since they are not changing 5'UTR sequences but rather translated protein sequences and thus, enabling protein isoforms. In this work we deliberately avoided the use of the term "alTSS" to refer to these internal alternative TSS and kept the established usual term "iTSS" where the "i" stands for internal. We could introduce the terms "aliTSS" or "ialTSS" but we try to keep the acronyms as few as possible.

Our research group studied before these internal alternative TSS (Ten-Caten et al., 2018) and illustrated in Figure (1-8) panel a, aiming to discover the intraRNAs messenger transcripts that could translate protein isoforms, as illustrated in Figure (1-4) in green. However, in that previous work some potential alternative internal TSS were arbitrarily and intentionally left out. Therefore we used the framework developed for the present PhD thesis and revisit (Ten-Caten et al., 2018) data in order to find additional alTSS (of the internal type, i.e. novel iTSS) .

The filtering criteria used by (Ten-Caten et al., 2018) was: "*(...) we applied filters excluding iTSS which: (i) do not pass stringent statistical significance cutoff of <10^{-15}, (ii) are located too close to CDS' edges (90 bp or 10% of CDS length margin), or (iii) are upstream of sub-sequences prone to form structured molecules.*" We applied restrictions similar to (i) and (iii) but not (ii). Instead of (ii) we looked for alternative TSS internal to coding regions from the CDS start codon up to 90 bp downstream from it, as this region was left out by the original authors' search for iTSS. Instead of a 10^{-15} cutoff in (i) we used 10^{-6}. We perform the same (iii) filtering with a -20 kcal/mol cutoff (see Methods section for details). We employed a p-value cutoff of 10^{-6} and conducted various analyses using differing cutoff values. Notably, while there were numerical differences in the resulting outcomes, the overall conclusions remained consistent.

Using this criteria we were able to find **96** protein coding genes showing alternative start sites of the iTSS kind. Table (4-4-1) shows the results of the 96 genes and also the table is presented as an Appendix Table 10 online.

Table (4-4-1): Arbitrary sample of genes (10 out of 96) genes that have internal alternative alTSS (iTSS) in *Halobacterium salinarum* NRC-1 considering all time points from the growth curve into a single collapsed dataset. ID is the TSS identifier; Replicon can be chromosome or plasmids; Position is the inferred coordinate for the iTSS; Strand refers to forward or reverse; Difference refers to the amount of reads that TEX+ libraries have more than TEX- libraries; p.Value is the statistical significance of the dRNA-seq enrichment; Positional Uncertainty is the error-bar associated with the iTSS position; Common.Name is the *loci* hosting the iTSS. [Full table also available in Appendix 10].

ID	Replicon	Position	Strand	Difference	*p*-Value	Positional Uncertainty	Common.Name
TSS_1093_3	chr	368837	+	19	0	0	VNG_RS01860
TSS_6466_3	chr	439546	-	22	0	0	VNG_RS02240
TSS_4226_3	chr	1649846	+	26	0	0	VNG_RS08655
TSS_1582_3	chr	573300	+	28	0	0	VNG_RS02975
TSS_9834_3	chr	1783219	-	29	0	1	VNG_RS09315
TSS_989_3	chr	318234	+	30	0	1	VNG_RS01600
TSS_747_3	chr	211759	+	31	0	4	VNG_OE1409F
TSS_2883_3	chr	1098002	+	31	0	0	VNG_RS05755
TSS_5128_3	chr	1992093	+	32	0	6	VNG_RS10415
TSS_3563_3	chr	1360343	+	37	0	0	VNG_RS07145

All 96 iTSS were visually inspected one-by-one and it was very clear that most of them were false positives or simple result from CDS misannotation. Out of 96, 17 (18%) iTSS are in fact regular usual genuine TSS that were considered internal because the CDS start codon was misplaced in the genome annotation used in this work. Some of those 17 were already corrected (TSS_6373_3 or TSS_9946_3 for example, available in the table appendix 10) by NCBI standard procedures by the time this text is being revised (January 2023). Figure (4-4-1) illustrates a radical case of misannotated CDS, VNG_RS04865, where our dRNA-seq analyses was originally searching for alternative internal alTSS but actually found the real gene's TSS, which in turn indicates a novel unknown *H. salinarum* specific protein instead of the current annotation of a broken pseudogene. Current official annotation for this *loci* states that this is a pseudogene in frame 1 reverse strand, therefore 5'→3' direction is right to left in Figure (4-4-1). However, iTSS coincides with an classical ATG start codon on frame 2 with Ribo-seq data showing ribosome occupancy along the extent of the newly annotated CDS until the stop codon ~300 bp way. Other CDS are still misannotated and small corrections to CDS sequence could be considered as (minor) positive collateral effect from the reanalyzes of dRNA-seq data. It is not the focus of this PhD work but it would inform future discoveries on this model organism.

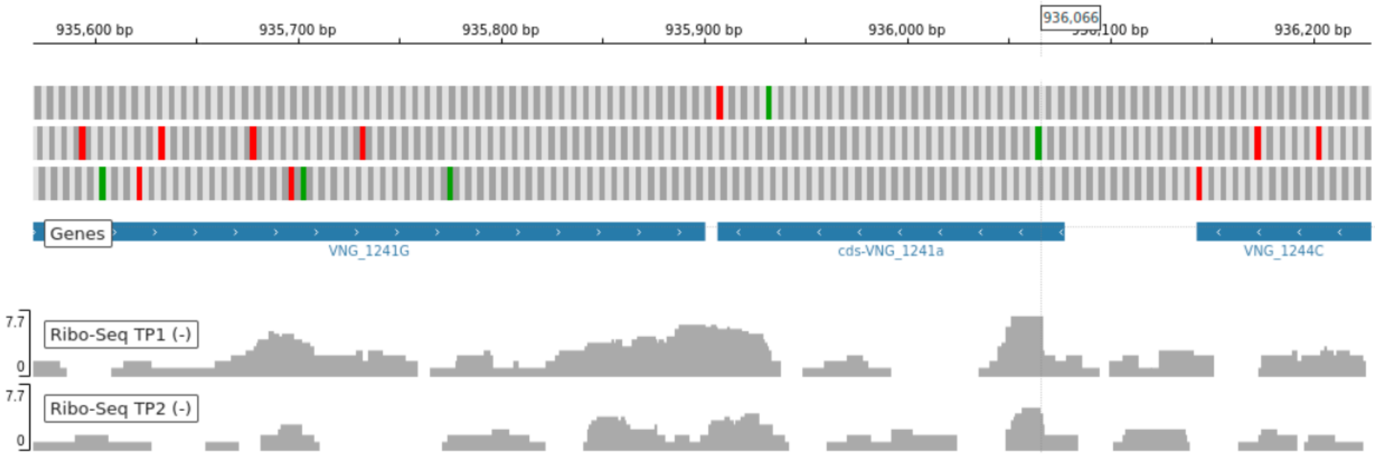
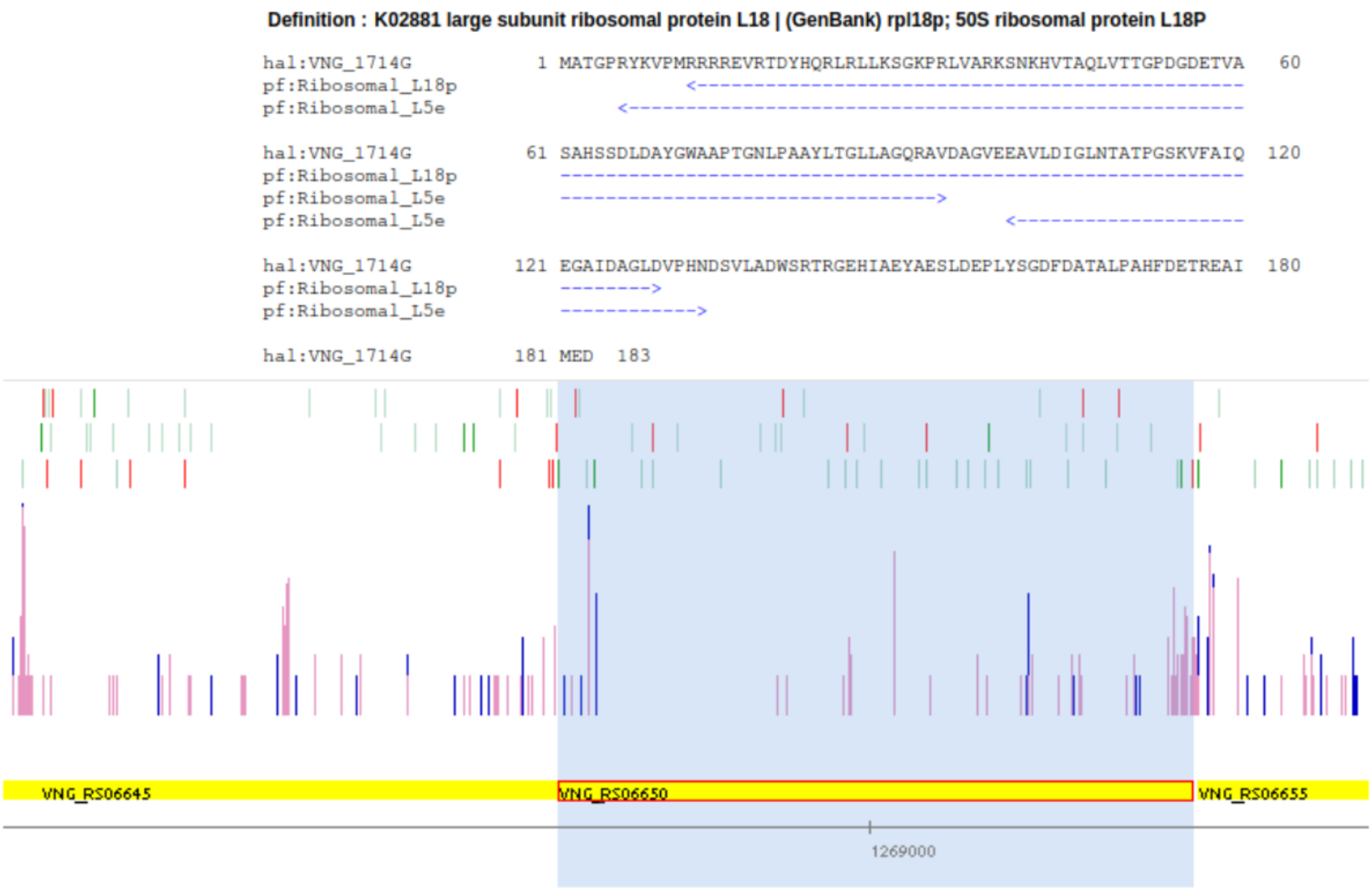


Figure (4-4-1): *Ribo-seq data showing an illustrates a radical case of misannotated CDS, VNG_RS04865, where our dRNA-seq analyses was originally searching for alternative internal alTSS. Gene is shown as observed in the browser with 5'→3' direction from left to right.*

The genes selected for display in Table (4-4-1) were a illustrative sample among the 16 iTSS (17%) considered likely to produce valid alternative transcripts and perhaps protein isoforms. They were considered likely since the dRNA-seq signal could not be explained by confounding factor such as secondary structure forming sequences adjacent to the aligned reads peaks, tested manually using RNAbows (http://rna.williams.edu/rnabows/) diverse prediction methods. Formally it is impossible, without experimental bench lab work, to be certain that the shorter RNA starting at the iTSS actually translate to a shorter protein isoform. It is possible that the RNA being transcribed from the iTSS point on is a ncRNA and not an mRNA. It is also possible that the mRNA translates not an isoform but rather a totally different protein out of the original annotated CDS frame. What the dRNA-seq data shows confidently is that an RNA molecule is transcribed in an overlapped fashion. The proximity with alternative internal start codons (not necessary ATG as other are known to exist in archaea) also plays a role on our classification of "potentially transcribed" status of those 16 genes listed in Table (4-4-1), from which an example is shown in Figure (4-4-2). This example, a ribosomal protein L18 (VNG_RS06650 or VNG1714G, synonymous identifiers), has 2 potential ATG start codons, and could generate two protein isoforms differing only at the first 10 amino acids: MATGPRYKVP. At the present we were unable to determine *in silico* if this amino-acid sequence is functional (i.e. signaling, recognition, docking, etc), however it is interesting to note that the KEGG-based motif search (https://www.kegg.jp/ssdb-bin/ssdb_motif?kid=hal:VNG_1714G) indicates that the characteristic L18p domain starts exactly at the second ATG enabled by the internal alTSS. This could mean that both isoforms are functional but with different properties. We consider this case a promising case for biological discovery and not only a simple start codon misannotation because there is plenty of transcription signal between both putative start codons in our and publicly available *H. salinarum* transcriptomes (data not show).

Experimental follow-up research is needed to test the bioinformatics hypothesis raised in this PhD work, in the same spirit carried out by (Ten-Caten et al., 2018), whose Western Blots showed the translation of alternative proteins related to iTSSs.



Definition : K02881 large subunit ribosomal protein L18 | (GenBank) rpl18p; 50S ribosomal protein L18P

```
hal:VNG_1714G          1 MATGPRYKVPMRRRREVRTDYHQRLRLLKSGKPRLVARKSNKHVTAQLVTTGPDGDETVA   60
pf:Ribosomal_L18p                   <------------------------------------------------
pf:Ribosomal_L5e              <------------------------------------------------------

hal:VNG_1714G         61 SAHSSDLDAYGWAAPTGNLPAAYLTGLLAGQRAVDAGVEEAVLDIGLNTATPGSKVFAIQ  120
pf:Ribosomal_L18p        ------------------------------------------------------------
pf:Ribosomal_L5e         ---------------------------------->
pf:Ribosomal_L5e                                                <--------------------

hal:VNG_1714G        121 EGAIDAGLDVPHNDSVLADWSRTRGEHIAEYAESLDEPLYSGDFDATALPAHFDETREAI  180
pf:Ribosomal_L18p        -------->
pf:Ribosomal_L5e         ------------>

hal:VNG_1714G        181 MED  183
```

Figure (4-4-2): *dRNA-seq signal profile for the ribosomal protein L18 gene (VNG_RS06650) which has 2 potential ATG start codons at +3 frame, and could generate two protein isoforms differing only at the first 10 amino acids. Gene is in forward strand therefore 5'→ 3' direction is from left to right. The blue over pink highest histogram bar marks the alTSS. Blue lines refers to TEX+ reads and Pink lines refers to TEX- reads.*

Most interesting is the fact that 63 (66%) iTSS were clearly identified as false positives in spite of the filter applied similarly to (Ten-Caten et al., 2018). It is know that the conceptual model supporting the whole dRNA-seq experiment interpretation assumes, as shown in Figure (1-8) panel a, that the TEX exonuclease enzyme will digest all RNA molecules without a protective triphosphate 5' end (5'PPP). However, there is a known limitation of this model which is the fact that the enzyme cannot move one degrading RNA polymers if encounters a strongly structured paired segment. Although this TEX limitation is known, most of the published works in dRNA-seq ignore this fact and consider it as an inevitable false positive. Some studies, such as (Ten-Caten et al., 2018), implement countermeasures to minimize the false positive error, as shown in Figure (3-4-1) in Methods section. In this PhD thesis we used a criteria based on Minimum Free Energy (MFE) tilling sequence segments and quantifying *in silico* their potential to form intrabase pairs and thus form secondary structures which stop the TEX enzyme and mimic TSS signals. We refer these false positive, indistinguishable from true TSS signals if only dRNA-seq data is considered, fake TSS (fTSS).

Manual inspection revealed that 66% of the 96 iTSS found using filters against fTSS, were indeed fTSS. This indicates that more bioinformatics research needs to be carried out in order to develop methods that take the RNA primary sequence into account to establish TSS (genuine gTSS, iTSS, antisense aTSS, alike) with more precision. If the primary aim of the study is to find genuine gTSS, the external to dRNA-seq information on CDS annotation can be easily integrated and the challenge is greatly simplified. However, for more subtle entities such alternative alTSS, internal iTSS, antisense aTSS and so on, really only on dRNA-seq data leads to avoidable false positives with a mixture of TSS and fTSS as results.

Although considered a nuisance that has to be taken care of to study transcription starts, we recognize that the result from our work could also be considered an opportunity to experimentally determine important potentially functional RNA secondary structures. As far as we are aware, the majority of claims and inferences regarding RNA secondary structures, such as hairpin loops for example, are made using theoretical and computational analysis. The field of secondary prediction is well advanced with established tools such as RNAfold (and all other famous tools in ViennaRNA http://rna.tbi.univie.ac.at/) or Mfold (updated to UNAfold http://www.unafold.org/) but it is important to bear in mind that they are all computational. There are sequencing protocols develop specifically to probe high-throughput RNA structures, such as SHAPE-Seq (Mortimer et al., 2012), but dRNA-seq is not one of them. Combining primary sequence analysis, trough strategies like ours, with dRNA-seq data, deployed for transcription studies, we propose in this PhD work that one is also able to extract experimental secondary structure information from this kind of data “as a bonus”.

Before exploring this idea into a new pipeline to validate experimentally predictions of secondary structures, we checked its consistency in a well known experimentally validated stem-loop hairpin from a system very important for *H. salinarum*, its gas vesicle system. It is known that *H. salinarum* buoyancy and light scattering is due to internal protein complexes that trap gas similar to a submarine’s Ballast tank. One of the most important genes in this system is the *gvpA* (VNG_RS12320 and VNG_RS10625 *locus* duplicated in both plasmids) has a well-characterized stem-loop at its end likely involved in mRNA stabilization (Pfeifer, 2015). We detected a fTSS, which would be correctly filter out when focusing on TSS since its MFE = -25.2 < -20 kcal/mol, that co-localized precisely with the aforementioned stem-loop hairpin, as show in Figure (4-4-3).

Moreover, an adjacent TEX+ > TEX- signal that barely did not classify as fTSS (thus, rather is a false positive TSS) since has MFE = -19.5 > -20 kcal/mol also co-localized with the other side of the hairpin.

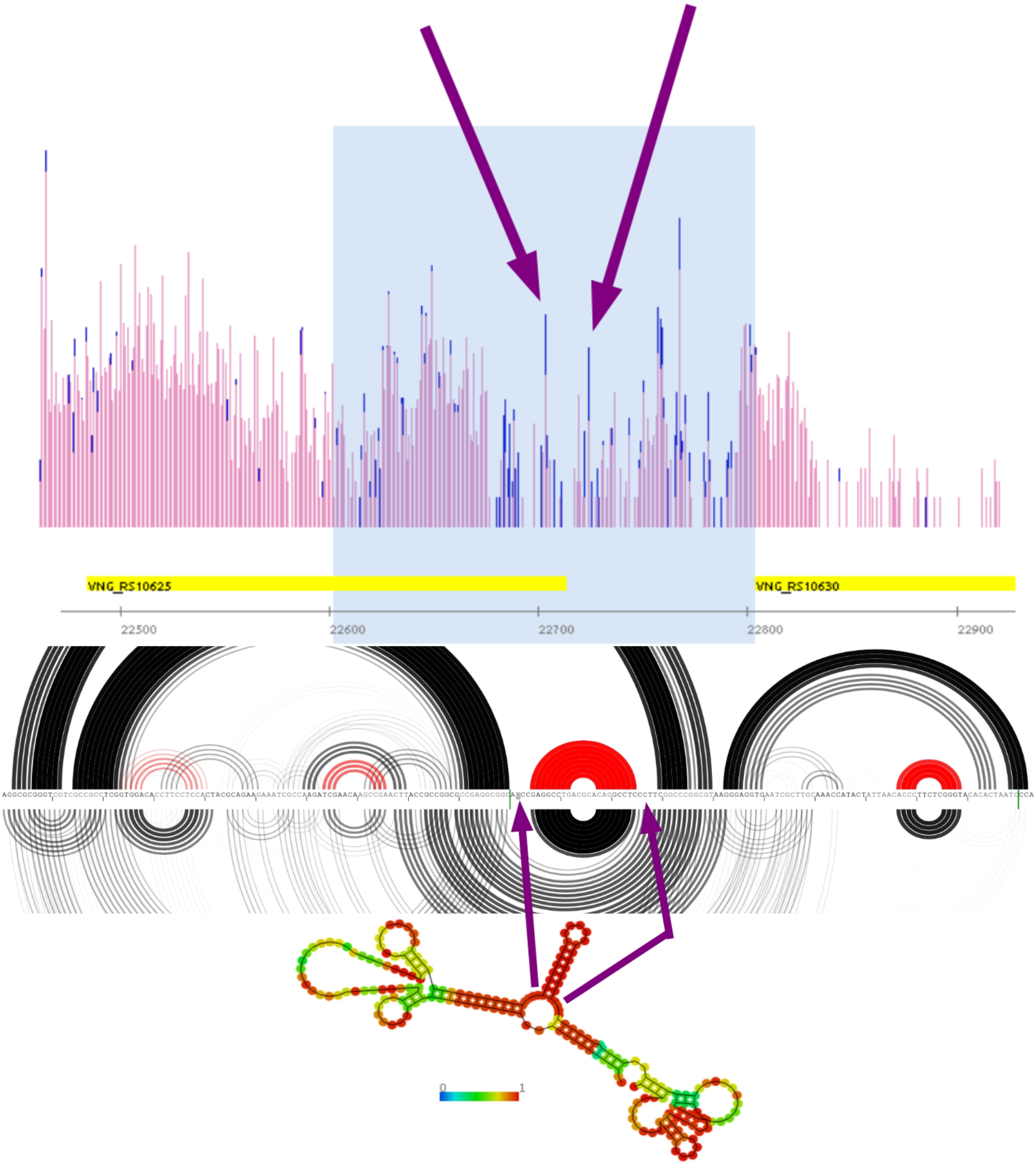


Figure (4-4-3): *Model for fTSS that co-localized precisely with the aforementioned stem-loop hairpin. A) Gaggle Genome Browser screenshot shows the fTSS position in left and in the right another fTSS also co-localized with the other side of the hairpin. Blue lines refers to TEX+ reads and Pink lines refers to TEX- reads. B) RNA-bows and 2D structure can show the possible hairpin between the both fTSS position. Gene is shown as observed in the browser with 5'⟶3' direction from left to right.*

Therefore, we gained confidence that dRNA-seq method can be modified to yield information beyond the scope of its original intent by properly interpreting fake TSSs. Previously, our research group characterized a set of transcripts transcribed near the 3' end of transposable elements belonging to the IS1341 group. These transcripts overlap the IS coding regions, are transcribed in the same orientation, presented a strong secondary structure computational signature, and show expression patterns distinct from their cognate gene (often anti-correlated) in several experimental conditions (Gomes-Filho et al., 2015). The IS1341 group belong to a very ancient archaeal transposable element family, IS200/605, and conservation of these discovered sense overlapping transcripts (sotRNAs) granted the establishment of two new RNA families in Rfam database v14.0: sot0042 and sot2652 (Rfam accessions: RF02656 and RF02657, respectively).

Originally described in *Halobacterium salinarum* NRC-1, sot0042 is now annotated in additional 52 species (144 sequences), mostly Halobacteria. Secondary structure predictions showed a putative tetraloop hairpin motif for which almost nothing is known (UUCA tetraloop) (Gomes-Filho et al., 2015). Therefore, a dRNA-seq dataset originally designed for TSS finding can be now use for secondary structure experimental validation, as shown in Figure (4-4-4).

Having turned this fake TSS issue into a useful dRNA-seq data feature, finding a well-known structural motif in an important *H. salinarum* gene and experimentally confirming a computational prediction from our own group which introduced a new archaeal RNA family, we moved on to address an original biological question in the scope of the current PhD work: transcription. The dRNA-seq protocol was designed originally to deal with transcription initiation events but now we could apply it to transcription termination events.

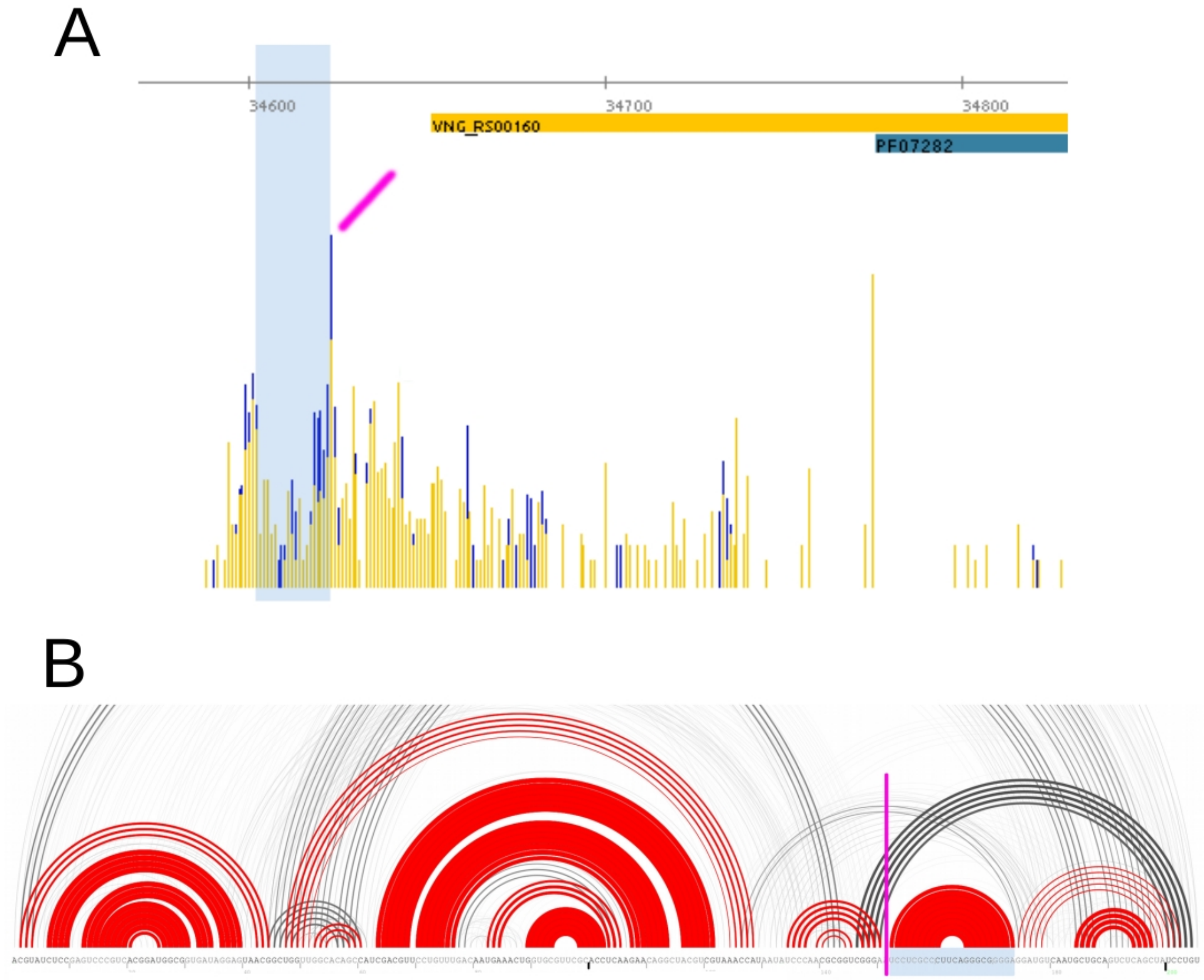


Figure (4-4-4): *Identification of tetraloop motif in sense overlapping transcript VNG_sot0042. (A) Distribution of aligned reads starting at a given genomic coordinate (horizontal ruler) for TEX+ (blue) and TEX- (yellow) datasets. Vertical bars in blue and yellow are superimposed signals (arbitrarily scaled log2 of normalized counts) at the same position. Coding sequence is in reverse strand (orange rectangle, locus ID inside, 5’→3’ direction is right to left). Domain annotation from PFAM database is shown (blue rectangle, ID inside). Tilted magenta marker points to the statistically significant TEX+ > TEX- signature (TSS_5399_3) compatible with the UUCA tetraloop motif prediction (light blue highlight). (B) Rainbow diagram representation of the predicted Minimum Free Energy (MFE) structure (red) and full partition function (black) with arc thickness proportional to the thermal average probabilities of base pairs. Diagram uses internal sotRNA 5’→3’ coordinates from left to right. Magenta and light blue markers are the same as in (A).*

4.5 Transcription Termination Sites (TTS)

One specific type of secondary structure that plays a role in transcription is the hairpin stem-loops that signal RNA polymerases to stop, abruptly or slowly, their RNA synthesis, as shown schematically in Figure (1-5) at the termination stage. This is not the only pathway to signal termination since in *Methanothermococcus thermolithotrophicus* an Oligo-dT sequence was first reported to work as an intrinsic signal for transcription termination (Thomm et al., 1993) and exhaustive characterization of intrinsic termination in *Thermococcus kodakarensis* confirmed that Poly(T) sequences were associated with intrinsic termination independent on any hairpin structures (Santangelo et al., 2009). Nevertheless, the RNA polymerase encounter with a secondary structure is indeed one of the most known signals to finalize its ongoing march synthesizing an RNA.

We engaged in finding strong secondary structure signatures using dRNA-seq near the end of annotated *H. salinarum* NRC-1 coding sequences expecting that these structures were terminators and, thus, establishing genomic coordinates defining Transcription Termination Sites (TTS). These TTS coordinates mark the boundaries where such structural features exist.

We defined a list of putative TTS for further investigation filtering the whole TSS list obtained by (Ten-Caten et al., 2018). As discussed in details in previous sections, an artifact signal produced by the inability of TEX enzyme to digest structured RNA sub-sequences yields fake TSS (fTSS), generally filter out if the focus is to establish TSS. On the other hand, what is filter out in that case is filter in if we are focusing on finding TTS. We obtained **86** putative TTS eliminating cases with MFE > -20 kcal/mol and distance between fTSS and coding sequence 3’ border greater than 50 nucleotides (|fTSS – CDS 3’ border | > 50). This filter is interpreted as keeping the lowest MFE candidates (likely structured) at most 50 nt away from the gene’s stop codon.

The following Table (4-5-1) shows the 86 TSS, the full table also can be found on-line as Appendix 11. Some secondary structures of the termination sites in *Halobacterium salinarum* NRC-1 were modeled, an example of the termination site sequences in gene VNG_RS07425 using RNAbows shown in Figure (4-5-1) for the 100 bp sequence around the predicted TTS position, using default representation (Aalberts & Jannen, 2013).

Table (4-5-1): Arbitrary sample of genes (10 out of 86) putative TTS with MFE < -20 kcal/mol and distance between fTSS and coding sequence 3' border less than 50 nucleotides. [Also can get the full table in Appendix 11].

ID	Replicon	Position	Strand	Difference	*p*-Value	Positional Uncertainty	Common.Name
TSS_1580_3	chr	573025	+	21	$3.10\ 10^{-13}$	0	VNG_RS02970
TSS_4141_3	chr	1616886	+	35	0	2	VNG_RS08500
TSS_7021_3	chr	706760	-	22	0	2	VNG_RS03615
TSS_3695_3	chr	1417232	+	44	0	0	VNG_RS07435
TSS_8033_3	chr	1077297	-	38	0	1	VNG_RS05655
TSS_10413_3	chr	1989672	-	239	0	11	VNG_RS10410
TSS_10513_3	chr	2009721	-	4	0.00066	0	VNG_RS10500
TSS_6829_3	chr	594371	-	12	0.000013	0	VNG_RS03105
TSS_2849_3	chr	1089863	+	64	0	2	VNG_RS05715
TSS_3234_3	chr	1240668	+	116	0	3	VNG_RS06445

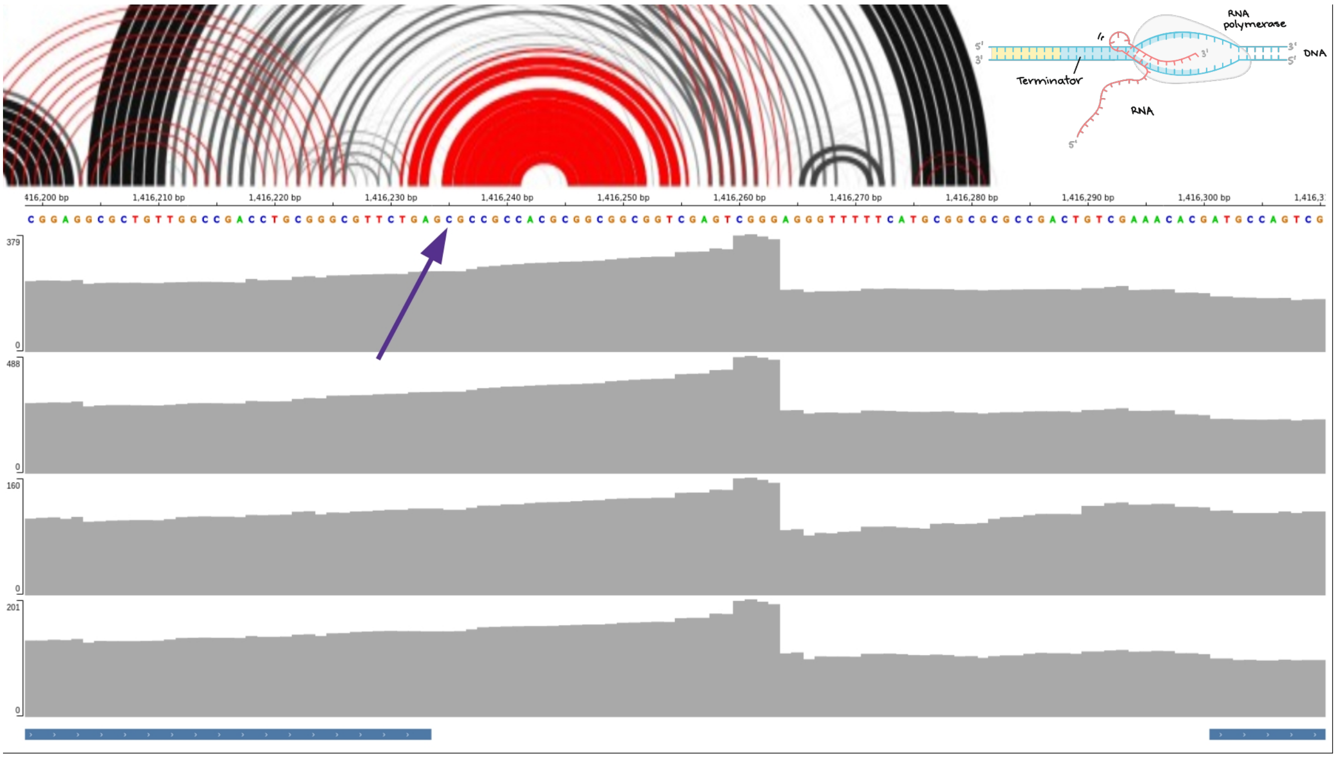


__Figure (4-5-1):__ Example of a probable termination site in __Halobacterium salinarum__ NRC-1, VNG_RS07425 loci. *Rainbow diagram representation of the predicted Minimum Free Energy (MFE) structure (red) and full partition function (black) with arc thickness proportional to the thermal average probabilities of base pairs. At same horizontal genomic scale there are 4 time-course growth-curve points of regular RNA-seq not from our group published and publicly available at* https://halodata.systemsbiology.net/viewgene/VNG_1912G*. The purple arrow points to the dRNA-seq inferred TTS (fake fTSS id: TSS_3689_3). In all third party transcriptome data there is a drop in transcripts counting of ~2-fold ~28nt downstream to the TTS indicating an mRNA end. Gene is shown with 5'→3' direction from left to right. [To get similar figures for more genes in Appendix 12]. In right corner above we repeat Figure (1-5) part 3 termination structure.*

Among the 86 putative TTS, we highlight TSS_9254_3 since it is in between two genes that are organized as a bicistronic operon. Koide and collaborators showed that *H. salinarum* have operons that are conditionally regulated (Koide et al., 2009), i.e. depending on the environmental stimuli a known operon can express independently some genes of the set, as shown in Figure (1-6). Most cases in which this phenomena was observed have the downstream genes from an array extra-expressed due to proper TSS located in short intergenic spaces or even inside the adjacent upstream gene where the promoter region lies. This independence allow stoichiometric decoupling among genes on the operon. However we were able to find an example in which the decoupling between the two genes in the operon seems to be mediated by a transcription termination of the upstream gene, as shown in Figure (4-5-2).

The gene *atpD*, coding for a subunit of a V-type ATP synthase at *loci* VNG_RS08275, is in operon with a small CPxCG-related zinc finger protein coding gene at *loci* VNG_RS08270. Our TTS data shows that there is a termination site and more than 10-year old published tiling-array data (Koide et al., 2009) shows that the expression levels of both genes are different breaking apart at the TTS region. A BLAST similarity search at DNA level retrieved only 12 archaeas for which these two genes were adjacent as operons in spite of VNG_RS08275 being vastly conserved (data not shown). Therefore it is reasonable to conclude that these two genes do not need to be co-expressed and the "premature" termination of a biscistronic mRNA is a likely mechanism to decouple them.

Improvements could be made in the TTS detection procedure to better predict secondary structures immediately downstream to fTSS positions without the usage of the crude MFE approximation. This open up new possibilities of research and a similar cataloging of alternative TTS (alTTS) in a similar philosophy that applied to the search of alternative start sites alTSS.

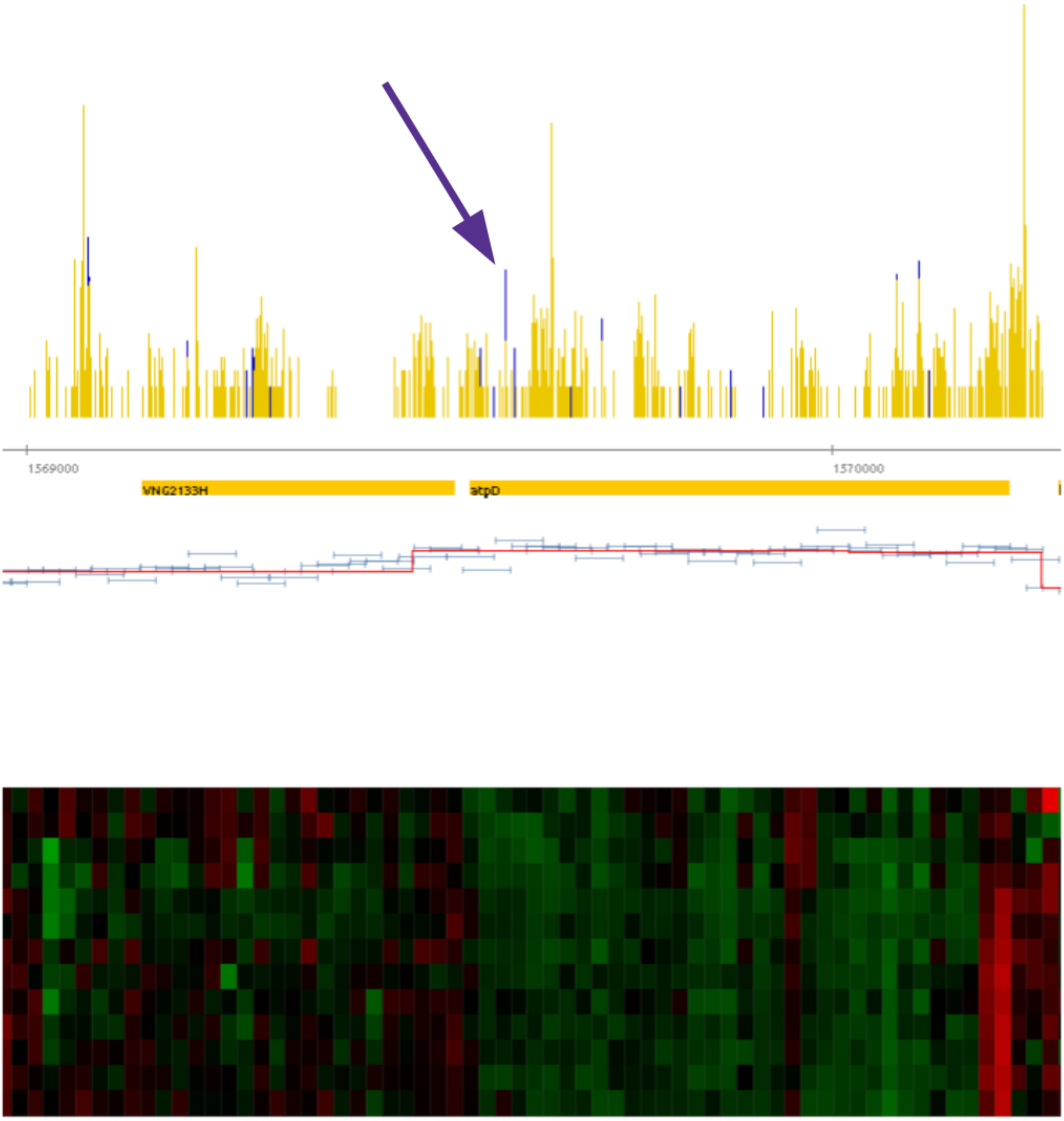


Figure (4-5-2): *Putative transcript termination site (TTS) at loci VNG_RS08275, coding for V-type ATP synthase subunit D. The decoupling between the two genes in the operon seems to be mediated by a transcription termination of the upstream gene. A) Gaggle Genome browser shows the termination sit middle of operon by the end of coding sequence from gene VNG_RS08275 and the next CDS of gene VNG_RS08270, for TEX+ (blue) and TEX- (yellow) datasets. B) Tilling array taken from (Koide et al., 2009) in the dark green area that have separation of next area in the same operon. This operon in in the reverse strand so 5'→3' direction is from right to left.*

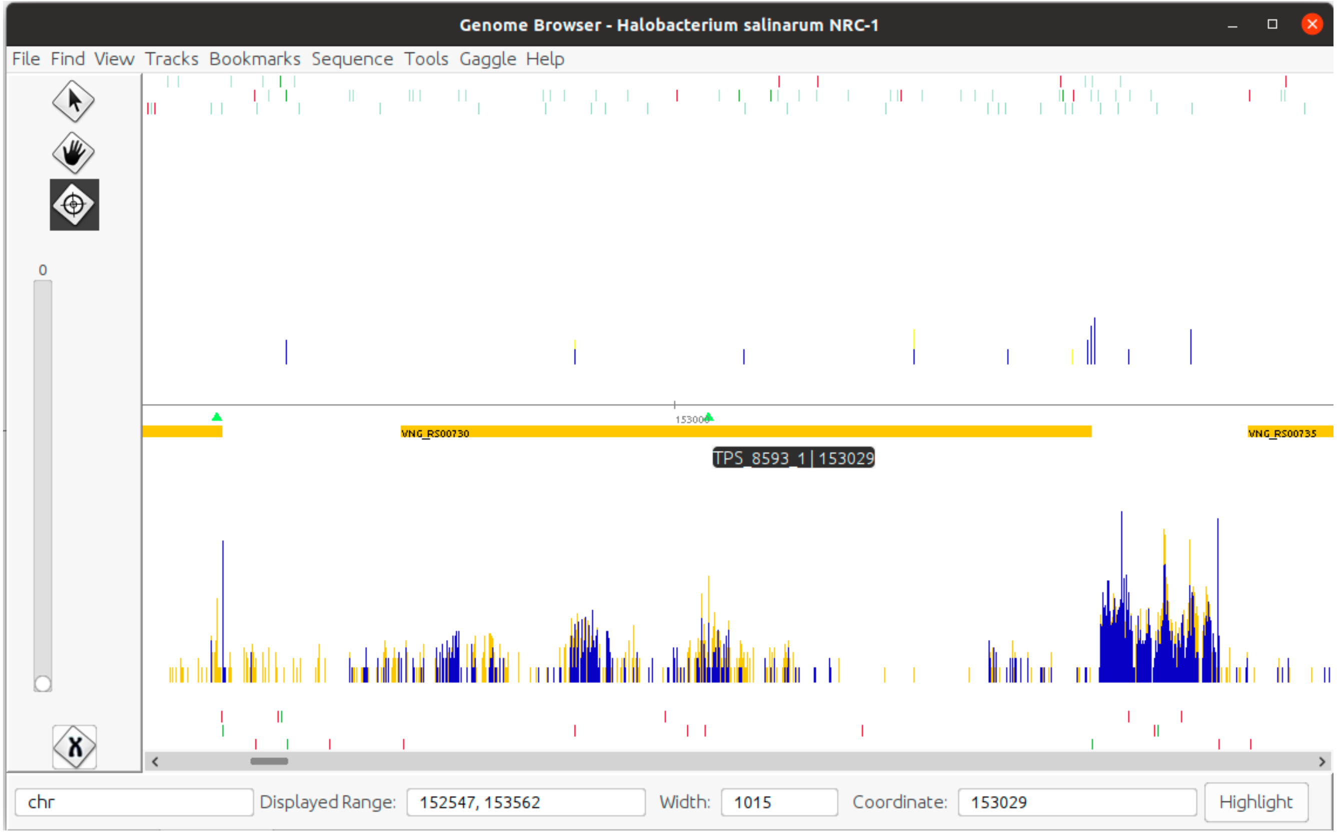
4.6 Transcription Processing Sites (TPS)

The study of the details involved in a dRNA-seq experiment, from both biochemical and bioinformatics points of view, allowed us to extend its usage to beyond transcript start site finding. We adapted this protocol to also map Transcription Processing Sites (TPS). Given that this topic concealed a higher probability of impact we decided to fsubmit a manuscript about it before the alTSS or TTS topics. In this section, the published article (Ibrahim et al., 2021) is reproduced.

We were successful in publishing this result and, to avoid repetition, in this section we attach the manuscript, commenting some aspects that are outside the scope of the final product. After its publication, which is the main original contribution from this PhD thesis, the “TPS paper” was extensively used in other works.

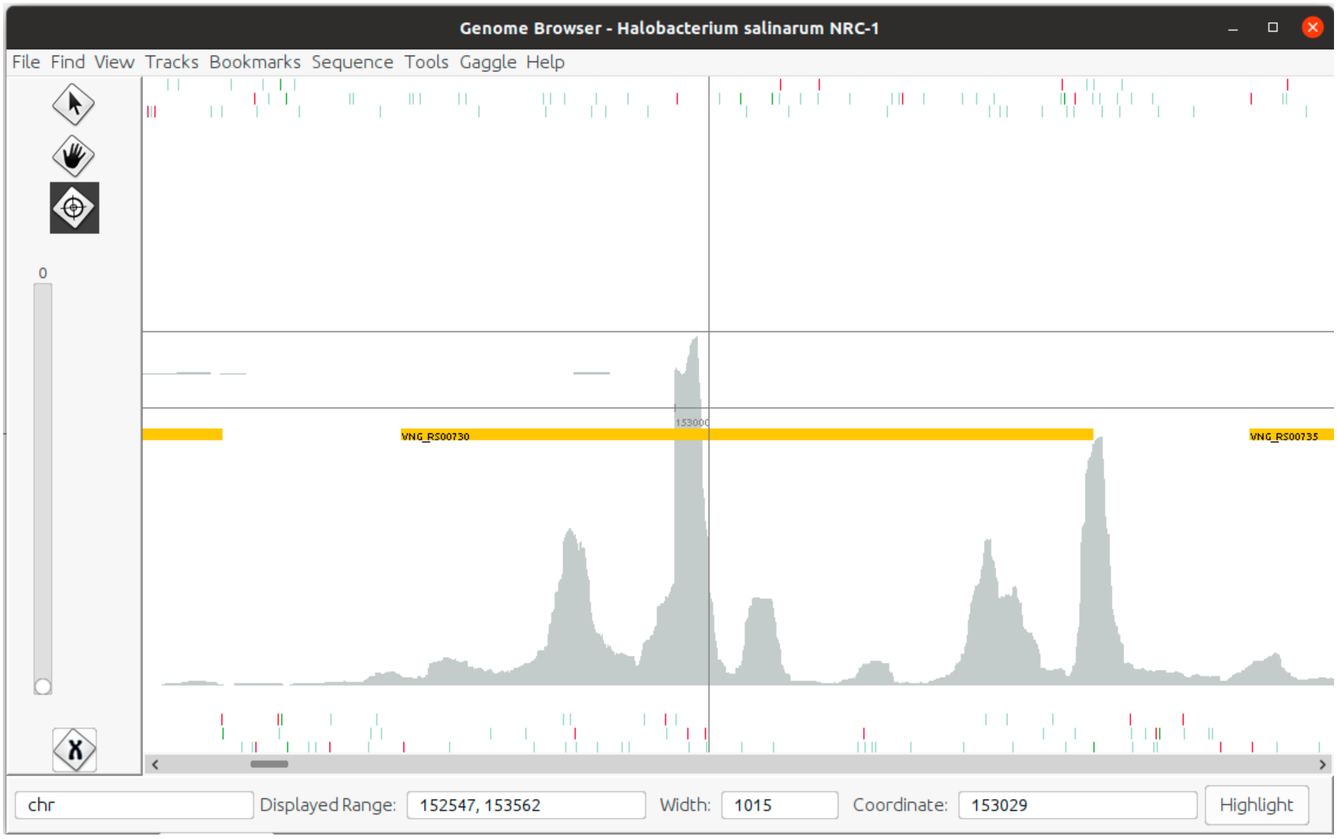
The work from Onga and other coauthors that does not include the author of this thesis (Onga et al., 2022), draw one of their main conclusions supported by the TPS map provided by (Ibrahim et al., 2021). In their Figure 3b they show a TPS inside the *rpl15e* gene, which codes for the 50S ribosomal protein L15e, as one of the main piece of evidence to support a post-transcriptional regulation claim, as shown in our Figure (4-6-1). Therefore, repercussions of the current PhD project was already taking place.

(a)


Genome Browser - Halobacterium salinarum NRC-1
File Find View Tracks Bookmarks Sequence Tools Gaggle Help
0
153000
VNG_RS00730
VNG_RS00735
TPS_8593_1 | 153029
Highlight

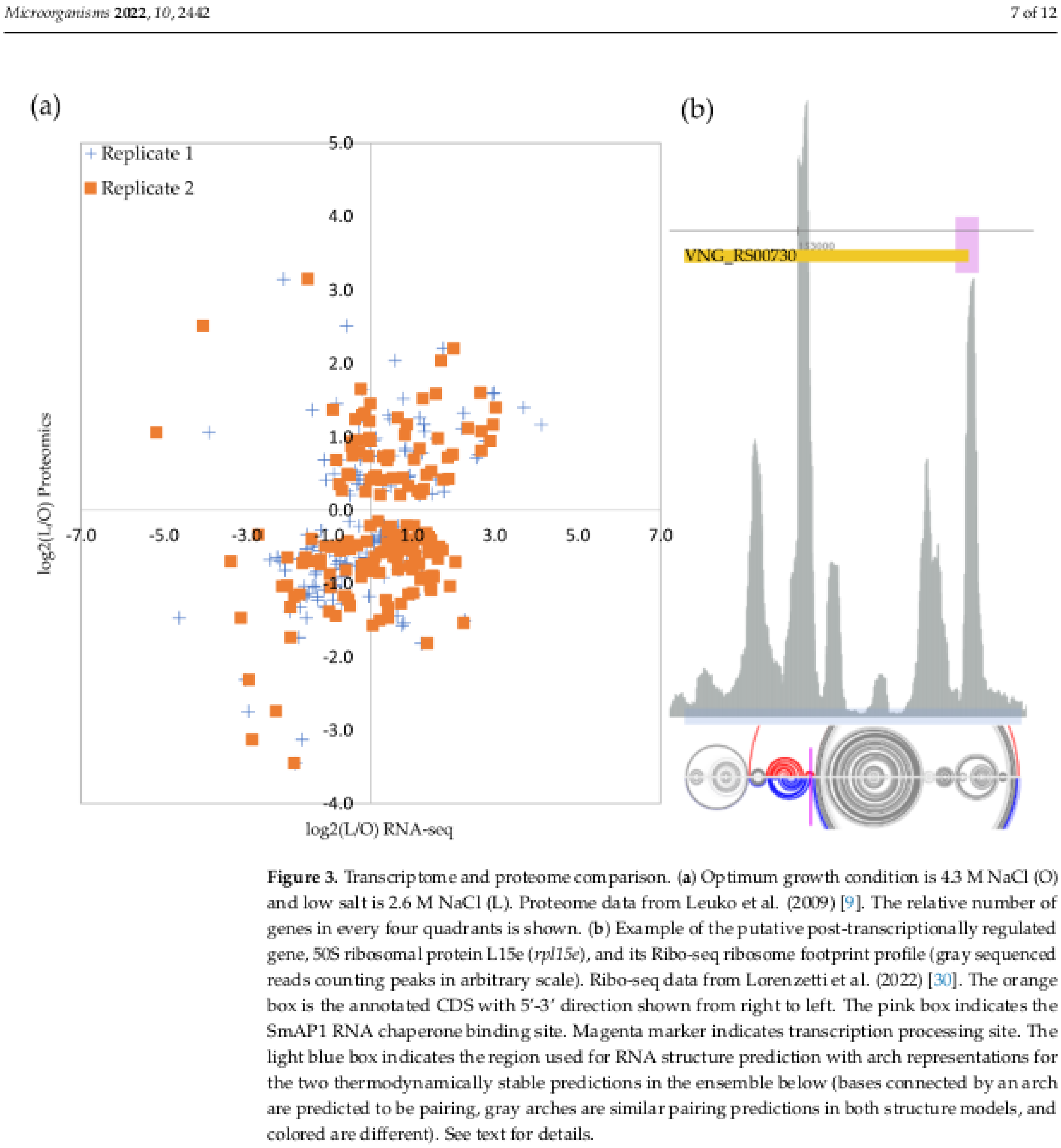
chr Displayed Range: 152547, 153562 Width: 1015 Coordinate: 153029

(b)


Genome Browser - Halobacterium salinarum NRC-1
File Find View Tracks Bookmarks Sequence Tools Gaggle Help
0
153000
VNG_RS00730
VNG_RS00735
Highlight

chr Displayed Range: 152547, 153562 Width: 1015 Coordinate: 153029

(c)



Microorganisms **2022**, *10*, 2442 7 of 12

Figure 3. Transcriptome and proteome comparison. (**a**) Optimum growth condition is 4.3 M NaCl (O) and low salt is 2.6 M NaCl (L). Proteome data from Leuko et al. (2009) [9]. The relative number of genes in every four quadrants is shown. (**b**) Example of the putative post-transcriptionally regulated gene, 50S ribosomal protein L15e (*rpl15e*), and its Ribo-seq ribosome footprint profile (gray sequenced reads counting peaks in arbitrary scale). Ribo-seq data from Lorenzetti et al. (2022) [30]. The orange box is the annotated CDS with 5′-3′ direction shown from right to left. The pink box indicates the SmAP1 RNA chaperone binding site. Magenta marker indicates transcription processing site. The light blue box indicates the region used for RNA structure prediction with arch representations for the two thermodynamically stable predictions in the ensemble below (bases connected by an arch are predicted to be pairing, gray arches are similar pairing predictions in both structure models, and colored are different). See text for details.

Figure (4-6-1): *Usage of the main contribution of this PhD thesis (Ibrahim et al., 2021) on a third party work (Onga et al. 2022). The gene on focus is rpl15e (see text for details). (a) screenshot of the Gaggle Genome Browser (GGB) utilization section made available for exploring TPS results and presented in (Ibrahim et al., 2021). for TEX+ (blue) and TEX- (yellow) datasets. Vertical bars in blue and yellow are superimposed signals at the same position. Coding sequence is in reverse strand (orange rectangle, locus ID inside, 5'→3' direction is right to left). (b) screenshot of a GGB use to visualize public Ribo-seq data and TPS also presented as a resource in (Ibrahim et al., 2021). (c) screenshot of Figure 3 at page 7 from (Onga et al. 2022) where the direct usage of visualization (b) was made (see text for details and refer to their paper for full biological interpretation).*

The work from Lorenzetti and coauthors that does not include the author of this thesis (Lorenzetti et al., 2023), extensively used the TPS map provided by (Ibrahim et al., 2021). They integrated TPS data with many other datasets available from our group and from the Institute for Systems Biology (https://isbscience.org/, Seattle-USA) to create a comprehensive *Halobacterium salinarum* NRC-1 Atlas, a web portal where one can explore a multi-omics growth curve experiment for this archaeal organism: https://halodata.systemsbiology.net/. Analysis from this PhD thesis is shown in a web-based genome browser, shown in Figure (4-6-2), that is a resource which in turn is also used in this PhD thesis to help analyze TTS or alTSS data, as shown for example in previous figures: Figure (4-2-1), Figure (4-4-1) and Figure (4-5-1).

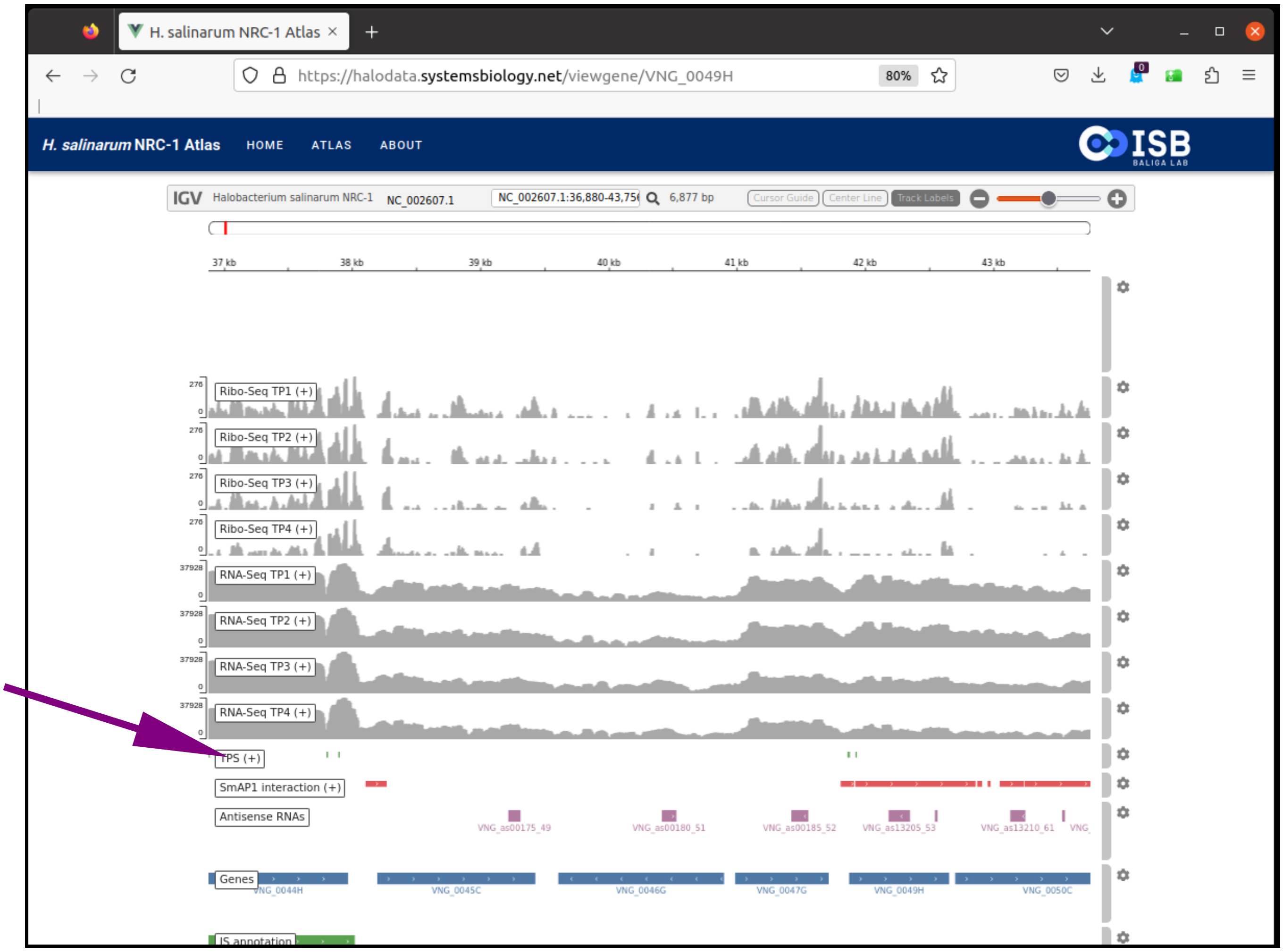
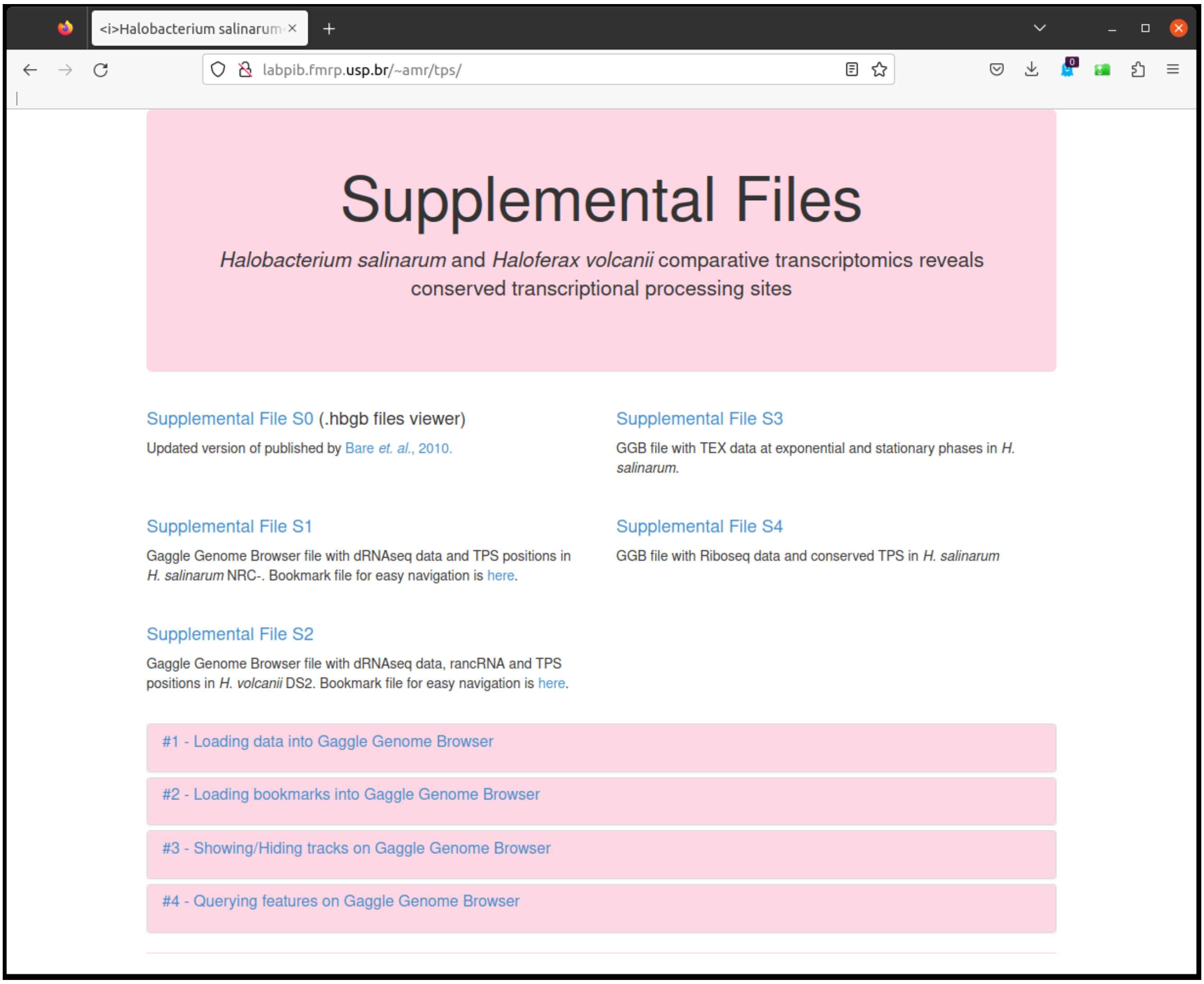


Figure (4-6-2): *Usage of the main contribution of this PhD thesis (Ibrahim et al., 2021) on a third party work (Lorenzetti et al., 2023). Screenshot from a regular web-browser. The genomic window is a arbitrary just to show some Atla's features and tracks. The purple arrow highlights the TPS map published in (Ibrahim et al., 2021). Gene is shown as observed in the browser with 5'→3' direction from left to right.*

Finally, we reproduce the Ibrahim and coauthors manuscript, which has 19 pages 10 supplemental figures and a companion web-site, from which a screenshot is shown as Figure (4-6-3). The companion web-site is where Gaggle Genome Browser files for many data mentioned on the manuscript can be interactively explored, as shown in several figures in this PhD dissertation.



Supplemental Files

Halobacterium salinarum and *Haloferax volcanii* comparative transcriptomics reveals conserved transcriptional processing sites

Supplemental File S0 (.hbgb files viewer)

Updated version of published by Bare *et. al.*, 2010.

Supplemental File S1

Gaggle Genome Browser file with dRNAseq data and TPS positions in *H. salinarum* NRC-. Bookmark file for easy navigation is here.

Supplemental File S2

Gaggle Genome Browser file with dRNAseq data, rancRNA and TPS positions in *H. volcanii* DS2. Bookmark file for easy navigation is here.

Supplemental File S3

GGB file with TEX data at exponential and stationary phases in *H. salinarum.*

Supplemental File S4

GGB file with Riboseq data and conserved TPS in *H. salinarum*

#1 - Loading data into Gaggle Genome Browser

#2 - Loading bookmarks into Gaggle Genome Browser

#3 - Showing/Hiding tracks on Gaggle Genome Browser

#4 - Querying features on Gaggle Genome Browser

Figure (4-6-3): *Screenshot of the companion web-site for the main contribution of this PhD thesis (Ibrahim et al., 2021) where all processed data can be downloaded and explored. Available at* http://labpib.fmrp.usp.br/~amr/tps/

genes

MDPI

Article

Halobacterium salinarum and *Haloferax volcanii* Comparative Transcriptomics Reveals Conserved Transcriptional Processing Sites

Amr Galal Abd El-Raheem Ibrahim [1,†], Ricardo Z. N. Vêncio [1,†], Alan P. R. Lorenzetti [2] and Tie Koide [2,*]

[1] Department of Computation and Mathematics, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto 14040-900, Brazil; amrgalal@usp.br (A.G.A.E.-R.I.); rvencio@usp.br (R.Z.N.V.)

[2] Department of Biochemistry and Immunology, Ribeirão Preto Medical School, Universidade de São Paulo, Ribeirão Preto 14040-900, Brazil; alorenzetti@usp.br

[*] Correspondence: tkoide@fmrp.usp.br; Tel.: +55-16-3315-3107

[†] These authors contributed equally to this work.

Abstract: Post-transcriptional processing of messenger RNA is an important regulatory strategy that allows relatively fast responses to changes in environmental conditions. In halophile systems biology, the protein perspective of this problem (i.e., ribonucleases which implement the cleavages) is generally more studied than the RNA perspective (i.e., processing sites). In the present in silico work, we mapped genome-wide transcriptional processing sites (TPS) in two halophilic model organisms, *Halobacterium salinarum* NRC-1 and *Haloferax volcanii* DS2. TPS were established by reanalysis of publicly available differential RNA-seq (dRNA-seq) data, searching for non-primary (monophosphorylated RNAs) enrichment. We found 2093 TPS in 43% of *H. salinarum* genes and 3515 TPS in 49% of *H. volcanii* chromosomal genes. Of the 244 conserved TPS sites found, the majority were located around start and stop codons of orthologous genes. Specific genes are highlighted when discussing antisense, ribosome and insertion sequence associated TPS. Examples include the cell division gene *ftsZ2*, whose differential processing signal along growth was detected and correlated with post-transcriptional regulation, and biogenesis of sense overlapping transcripts associated with IS*200*/IS*605*. We hereby present the comparative, transcriptomics-based processing site maps with a companion browsing interface.

Keywords: transcript processing sites; TPS; RNA processing sites; dRNA-seq; *Halobacterium salinarum* NRC-1; *Haloferax volcanii* DS2; comparative transcriptomics; post-transcriptional regulation; gene expression

check for updates

Citation: Ibrahim, A.G.A.E.-R.; Vêncio, R.Z.N.; Lorenzetti, A.P.R.; Koide, T. *Halobacterium salinarum* and *Haloferax volcanii* Comparative Transcriptomics Reveals Conserved Transcriptional Processing Sites. *Genes* **2021**, *12*, 1018. https://doi.org/10.3390/genes12071018

Academic Editors: María José Bonete, Mónica Camacho and Julia Esclapez

Received: 30 March 2021
Accepted: 27 May 2021
Published: 30 June 2021

1. Introduction

Post-transcriptional processing of mRNAs is an important regulatory mechanism of gene expression [1,2]. The phenomenon can be seen from two complementary viewpoints: the protein and RNA perspectives. For halophilic archaea, RNA-binding proteins, such as chaperones, and exo- and endoribonucleases, were characterized and the genetic perturbation was assessed at the systems-level [3,4], giving a general view from the protein perspective. On the other hand, although decay and turnover rates have been measured at the systems-level [5], there are no comprehensive genome-wide maps of processing sites comparable to that available for the methanogenic psychrophilic archaeon *Methanolobus psychrophilus* [6] or the hyperthermophile *Pyrococcus furiosus* [7], which would address the problem from the RNA perspective. Halophile research would benefit from the availability of such processing site maps for its two main model organisms, *Halobacterium salinarum* and *Haloferax volcanii*.

Halobacterium salinarum is a photosynthesizing archaeon that does not rely on either chlorophyll or bacteriochlorophyll [8]. It shows no turgor pressure and uses the "salt-in" strategy to achieve osmotic balance [9]. *H. salinarum* is an obligatory halophile that grows optimally at 4.3 mol/L NaCl concentrations and gets lysed at low salt concentrations. Its intracellular concentrations of K^+ and Na^+ are measured as ≈4 mol/L and ≈1.4 mol/L, respectively, with Cl^- being just 10% higher than the growth medium [10]. Classical and modern biological and biotechnological achievements were undertaken using *H. salinarum* as a model, ranging from the structural elucidation of bacteriorhodopsin [11] to vaccine improvement [12]. These achievements leveraged the proposal of an environmental/gene regulatory influence network that can predict transcriptional changes to new environmental or genetic perturbations [13,14] and consolidated *H. salinarum* as a model organism for gene regulation among archaea. *Haloferax volcanii*, on the other hand, grows optimally at 1.7 to 2.5 mol/L NaCl, a relatively moderate salt requirement in the halophile context [15]. *H. volcanii* is incapable of phototrophic growth or floating using a gas vesicle system, but forms biofilms and is more genetically stable due to less numerous insertion sequences [16]. It is able to produce most of its amino acids and degrade a set of sugars that *H. salinarum* cannot metabolize [15,16]. Overall, *H. volcanii* is a valuable model organism for its practical aspects [15], even participating in the early moments of the ongoing CRISPR revolution [17]. The amount of transcriptome data available for *H. salinarum* and *H. volcanii* is unparalleled in the third domain of life (≈53% of experiments and ≈24% of RNA entries for archaea at NCBI's Gene Expression Omnibus and Sequence Read Archive databases, respectively, in January 2021).

Classically, a transcriptomics experiment is designed to understand the dynamic behavior of messenger transcripts as a proxy for protein-level changes. However, high-resolution hybridization (tiling microarrays) and sequencing-based (RNA-seq) transcriptome measuring technologies were able to reveal new genomic features that are difficult to be identified only by computational analysis of DNA sequences [18,19], expanding the scope of traditional transcriptomics to beyond differential expression. With an exponential decrease in costs, sequencing platforms were used in combination with several kinds of sample preparation strategies to explore different aspects of the transcriptome, creating a family of specialized RNA-seq protocols [20,21]. One such protocol is the differential RNA sequencing (dRNA-seq) method. The dRNA-seq method works by comparing differences between TEX (Terminator 5′ phosphate dependent exonuclease) treated samples against control samples in search of differential enrichment of 5′ triphosphorylated ends, thus unveiling primary transcripts. However, often neglected information that could be provided by the same protocol is the set of enriched 5′ monophosphorylated transcripts, which thus, unveils processed transcripts. Specific protocols for the efficient detection of RNA processing sites were developed (RiboMeth-seq [7], 5′P-seq [6] or pRNA-seq [22]), but legacy dRNA-seq datasets, originally deployed to map transcription start sites (TSS), can be readily reanalyzed. Instead of comparing an exonuclease-treated (TEX+) enrichment signal against a non-treated (TEX−) control (thus, TEX+ > TEX−), a search for TEX−mediated depletion signals (thus, TEX− > TEX+) were carried out, identifying transcriptional processing sites (TPS) [23].

In this work, we mapped TPS along the genomes of two halophilic model organisms, *Halobacterium salinarum* NRC-1 and *Haloferax volcanii* DS2, using previously acquired dRNA-seq data [24,25] and performed comparative transcriptomics to assess conservation. Finally, the processing site map just built was used to highlight specific cases in which post-transcriptional regulation seems to be relevant.

2. Materials and Methods

2.1. dRNA-seq Raw Data Alignment Reanalysis

Raw dRNA-seq data from *H. salinarum* NRC-1 [24] and *H. volcanii* DS2 [25] were reanalyzed with our in-house pipeline ("Caloi-seq", https://github.com/alanlorenzetti/frtc/ (accessed on 27 May 2021)) as a byproduct of our group's previous work on internal TSS (iTSS) and intraRNAs [24]. Although not fully explored in that work, which focused on iTSS, the pipeline created alignment files that are useful for the current work and that are considered the input data for our current reanalysis. Briefly, the read quality was inspected using FastQC v0.11.7; paired-end libraries were processed using Trimmomatic v0.36; reads that passed the filters as R1-R2 pairs were aligned to *H. salinarum* NRC-1 and *H. volcanii* DS2 reference genomes (NCBI's Assembly accessions ASM680v1 and ASM2568v1) using HISAT2 v2.1.0 in paired-end mode. Orphan reads R1 and R2 that passed the trimming process but not as pairs were aligned using single-end mode. The resulting alignment files were used as input for MMR to treat multi-mapped reads and the output was submitted to an additional filtering step to keep only the aligned R1 reads. Furthermore, for visualization purposes, we processed MMR output files using deepTools v2.5.3 [26] to compute the genome-wide read coverage. The raw input files, under the accession numbers SRP137801 and SRP076059, were obtained from NCBI's SRA database.

2.2. *Transcript Processing Site (TPS) Mapping*

We used TSSAR v1.0.1 [27] to identify TPS positions, deliberately changing the tool's TEX+ input for the TEX− R1 alignment file and vice versa in order to highlight statistically significant 5′ monophosphorylated depletion signals (TEX− > TEX+). Therefore, TSSAR's "TSS" outputs are actually TPS. TSSAR parameters were *p*-value (*p*) of $p < 10^{-9}$ with a minimum of 10 reads and a distance of "TSS" grouping of at least 5 nt.

2.3. *TPS Conservation*

TPS were defined as "conserved" if the normalized position D in orthologous gene pairs (from the 1554 available at OrthologeDB [28]) was less than two units apart (as introduced in [29]): $|D_{\text{Hsa}} - D_{\text{Hvo}}| < 3$. The normalized position is the genomic coordinate x, where the TPS was found, divided by the CDS length: $D = 100\,|x_{\text{CDSstart}} - x_{\text{TPS}}| / |x_{\text{CDSstart}} - x_{\text{CDSend}}|$; therefore, it is standardized within the interval zero and 100 if TPS is exactly at the start codon or at the last CDS nucleotide, respectively. Moreover, $D < 0$ and $D > 100$ were allowed to represent out-of-bound TPS at the 5′ and 3′ of a CDS, respectively.

2.4. *Secondary Structure Prediction*

Secondary structure predictions were carried out, using RNAfold v2.4.8 at the RNAfold web server [30] using default parameters, except for energy parameters, where "Turner model, 1999" was selected. For a few arbitrarily selected sequences, additional confirmatory predictions were carried out using the RNAbow, AllPairsMFE method with default parameters [31]. In order to highlight only high-probability base pairings in specific cases, we used the RNAstructure, ProbablePair method, with a > 0.99 probability cutoff [32]. RNA multiple alignment was carried out using LocARNA-P [33] with the "Turner model, 1999" energy parameters.

2.5. *Gene Set Enrichment Analysis*

Gene functional classification and Gene Ontology enrichment analysis were performed using the PANTHER system [34] at http://pantherdb.org/ (accessed on 28 March 2019) The default cutoff of the False Discovery Rate (FDR) of <5% was used in the Binomial test in the PANTHER Overrepresentation Test suite (Released 20181113). In order to avoid biases toward the orthologous genes' functional classes, only categories/terms enriched that are not also enriched in an orthologous vs. whole genome comparison were considered.

2.6. *Ribo-seq Data Analysis*

H. salinarum NRC-1 ribosomal profiling (Ribo-seq) data were reanalyzed using the same pipeline used for dRNA-seq without special treatment, which is usually deployed when dealing with translation initiation site identification. In order to generate the genome browser visualization, all time point samples and their replicates were combined by summing their read coverage. The data are publicly available under NCBI's SRA accession SRP119792 [35].

2.7. Differential Expression and Qualitative Analysis

The expression profiles along the genome coordinates were considered mostly qualitatively and the signal profiles were evaluated visually using the Gaggle Genome Browser [36] tool, with no special attention to numeric *y*-axis scales or normalization, except when comparisons between libraries or experiments were made. All expression profiles were examined in $\log_2$ or absolute scales with appropriate identification. Traditional *H. salinarum* RNA-seq data, a complementary dataset published along with Ribo-seq data (SRP119792), were used in specifically marked cases. *Natrinema* sp. J7-2 differential expression analysis was carried out after signal normalization to ensure comparability between three experiments. The sum of coverage in all genomic positions was used as the normalization factor, so the signal integral was set to 1, re-scaled to "per million" for convenience and shown without $\log_2$ scale. The *H. salinarum* differential expression analysis used $\log_2$ and a simple normalization factor that rescaled the least sequenced libraries to match the most sequenced library.

3. Results

3.1. Transcript Processing Site (TPS) Mapping

We identified transcript processing sites (TPS) in *H. salinarum* NRC-1 and *H. volcanii* DS2 by reanalyzing public dRNA-seq. We searched for TEX+ induced depletion signals using the TSSAR v1.0.1 stand-alone tool [27]. High number of sequencing reads starting at a specific given position in TEX− libraries that were depleted in TEX+ libraries are signatures of 5′ monophosphorylated molecules and, therefore, the outcome of a cleavage event of a longer precursor. In order to properly explore the generated map, interactive stand-alone Gaggle Genome Browser [36] interfaces were created for both organisms. The interfaces were also used as a source for some of the figures and supplemental figures (Supplemental Files S1 to S4; available at http://labpib.fmrp.usp.br/~amr/tps/ (accessed on 27 May 2021)).

This approach identified 2098 processing sites in 1183 coding sequences (CDS) in *H. salinarum* (Table S1). This number represents 43% of all 2782 CDS considered (annotation from [24]). Most genes showed a single TPS (59%) and almost all genes presented, at most, three TPS (91%). Only two genes had more than 10 TPS: a halolysin (VNG_RS10060 *locus*) and an integrase (VNG_RS00870). The region with the highest TPS density (8 events in 200 nt, 0.04 nt-1) is associated with an RNase H domain containing exoribonuclease (VNG_RS04745, Figure S1a).

In order to learn about the spatial distribution of TPS along genes, each TPS coordinate was normalized considering its cognate CDS length, creating a relative position coordinate system. Values 0 and 100 indicate TPS at the start and stop codons, respectively, and values outside this range are allowed to account for putative 5′ and 3′ UTR processing. Figure 1a shows the overall distribution of TPS positions along *H. salinarum*'s CDS in which most of the TPS are located near the start codons.

The reanalysis of *H. volcanii* data yielded 3515 TPS distributed in 1408 CDS (Table S2) out of 2828 chromosomal CDS considered (49%; RefSeq annotation for ASM2568v1). Genes with a single TPS were the most common (45%) and cases showing at most three TPS were the majority (79%). Figure 2b shows the distribution of TPS normalized positions along *H. volcanii*'s CDS, showing a preferential location near the start codon as in *H. salinarum*, and a higher concentration near the stop codons. The region with the highest TPS density (11 in 200 nt, 0.055 nt^{-1}) is surrounding the chaperone *cspA4* gene's stop codon (HVO_RS14265 *locus*, Figure S1b).

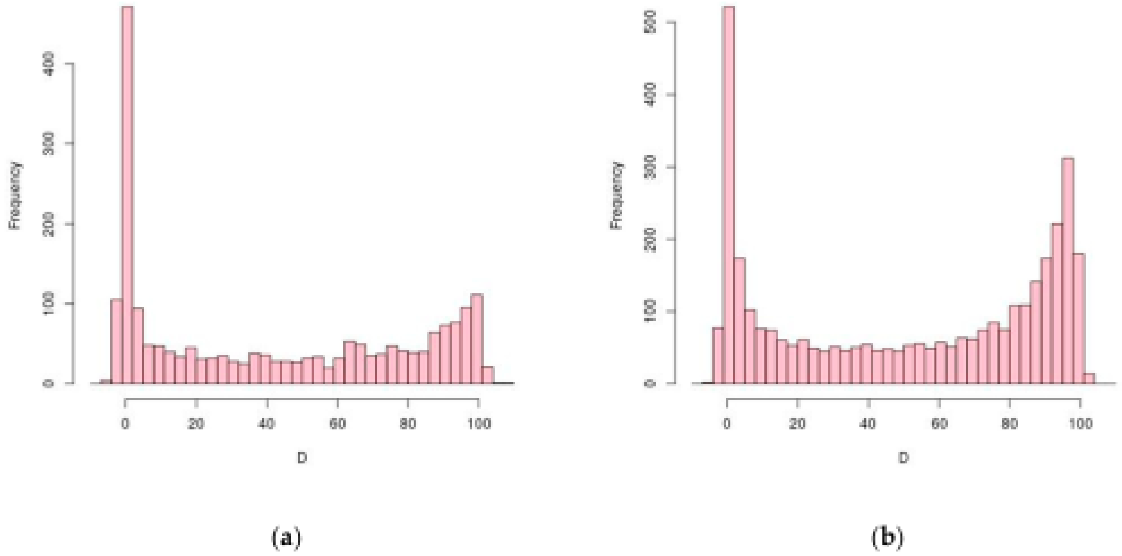


Figure 1. Distribution of CDS length-normalized positions (*D*) of TPS (transcriptional processing sites). Processing sites at start and stop codons are *D* = 0 and *D* = 100, respectively. Histograms have size 3 bins. (**a**) *H. salinarum* NRC-1; (**b**) *H. volcanii* DS2.

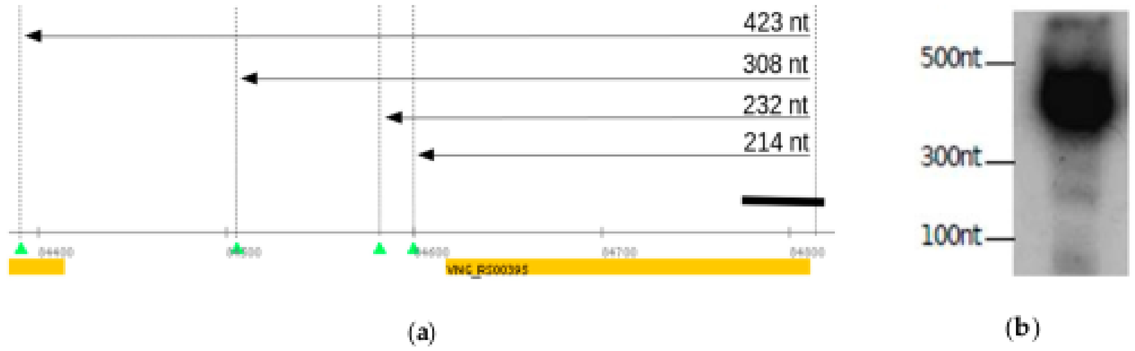


Figure 2. Experimental validation of predicted TPS using gene *cspA1* (VNG_RS00395 *locus*). (**a**) The consecutive TPS predict mRNAs of various sizes (arrows) from TSS to TPS_8361_1, TPS_8363_1, TPS_8365_1 and TPS_8366_1. Northern blot probe location is marked by the black rectangle. (**b**) Northern blot adapted from Figure 3 in [37].

Here, we focused only on annotated CDS and adopted a relatively stringent statistical significance cutoff, $p < 10^{-9}$, an inclusion criterion that, if relaxed, could result in more TPS candidates (Tables S3 and S4). Moreover, tRNAs or rRNAs, which are not the focus of this work, are well-known cases of intense processing and, as expected, harbor plenty of TPS signals (see Supplemental Files S1 and S2 for exploration in the genomic context).

Aiming the validation of our approach, we reused the previously published Northern blot results from our group in which some of the detected bands can now be explained by the TPS analysis. Back then, the *cspA1* gene, which encodes for a highly conserved chaperone protein (VNG_RS00395 *locus*), was probed to demonstrate the ubiquitous presence of TSS associated RNAs (TSSaRNA) in all three domains of life [37]. However, it was left unexplained in [37]: (i) the fainter bands longer than the TSSaRNA and (ii) the strongest signal between 300 and 500 nt, much longer than the 195 bp CDS. Transcripts starting at the genuine TSS (5 nt 5′UTR) and ending at TPS would result in truncated molecules of 423, 308, 232 and 214 nt in length, respectively (Figure 2a and Figure S2a). This prediction is consistent with the experimental observation of bands between the 500, 300 and 100 nt molecular markers (Figure 2b, adapted from Figure 3 in [37]).

Similarly, processing sites provide reasonable explanations for the previously published, seemingly odd Northern blot results: the conserved arginine deiminase pathway *arcRACB* gene cluster was interrogated by different probes revealing diverse isoforms, two of which unusually started inside *arcA*, forming *arcA'CB* and *arcA''CB* transcripts (Figure 4C in [38]). These 25-year-old results are remarkably consistent with sites TPS_20932_1 and TPS_193446_1 (Figure S2b).

Cases in *H. volcanii* can also be mined from the literature. The published Northern blots showing fragmented expression patterns of the polycistronic *tsg* operon, which encodes an ABC-type sugar transport system, validated our TPS finding approach: the dRNA-seq reanalysis showed processing break-points that would result in 7.1, 5.2, 3.8, 2.9, 1.5, and 1.9 knt transcripts that are consistent with experimental bands and the original authors' hypotheses of the partial termination and processing of a large polycistronic primary transcript, resulting in several transcripts observed by Northern blot (from operon's TSS to TPS_086838, TPS_086886, TPS_086946, TPS_086991 and TPS_087101 at the 5′-end part; from TPS_086892 to TPS_086838 at the 3′-end part; matching with Figure 6b,d in [39], among other combinations).

Similar observations can be made even between different organisms. The published Northern blot results for gas vesicle system genes in the moderately halophilic microorganism *Haloferax mediterranei* showed different transcripts isoforms arising from the *gvpDEFGHIJKLM* operon [40], consistent with our *H. salinarum* TPS data (*H. volcanii* has no gas vesicle system). The cleavage breakpoint in *H. mediterranei*'s transcript that generated the *gvpDE'* (2 kb) processed transcript is equivalent to *H. salinarum*'s TPS_19964_1 (VNG_RS10615 *locus*), which in turn predicts a 2087 nt transcript in *H. salinarum* (Figure S2c). Analogously, the two *H. mediterranei* 0.45 and 1.3 kb *gvpD'* transcripts are consistent with TPS_18331_1 and TPS_19982_1, respectively. Additional validation is provided by TPS found within the main gas vesicle system gene *gvpA1* (VNG_RS10625 *locus*): it is located at the basis of a stem–loop structure, experimentally implicated in the mRNA stability [40]. This stem–loop is 64 nt upstream to the stop codon in *H. mediterranei* and the TPS is 61 nt upstream to the stop codon in *H. salinarum* (Figure S2d).

Taking together, these results validate our dRNA-seq re-purposing to find transcript-processing sites (TPS).

3.2. TPS Conservation

H. salinarum and *H. volcanii* are both halophiles but they have distinct biological properties, including requirement/tolerance in environmental salt concentrations. They belong to the same phylum and class (Halobacteria) but to different orders (Halobacteriales and Haloferacales). Therefore, we assume that the TPS present in both organisms are probably relevant.

We defined TPS as "conserved" if the relative normalized position D in which they are found in orthologous gene pairs are approximately the same (i.e., $|D_{\text{Hsa}} - D_{\text{Hvo}}| < 3$, in a 0 to 100 scale). This criterion yielded 244 TPS (Table S5) in 178 orthologous gene pairs out of 1554 available at OrtholugeDB [28].

Enrichment analysis [34] using all genes as background showed that Gene Ontology (GO) terms are overrepresented only in the 178 orthologous genes bearing conserved TPS but not in the whole orthologous gene set (Table S6). Among these, we highlighted the most enriched "biological processes", "molecular functions" and "cellular components" GO branches, respectively; all with approximately 10-fold enrichment with a false discovery rate (FDR), controlled at less than 5%: iron–sulfur cluster assembly (GO:0016226); ATPase activity, coupled to transmembrane movement of ions, rotational mechanism (GO:0044769); and proton-transporting two-sector ATPase complex (GO:0016469).

Conserved TPS are concentrated upstream of start codons and around stop codons (Figure 3). The histogram of D for conserved TPS has similar characteristics of general TPS (Figure 1).

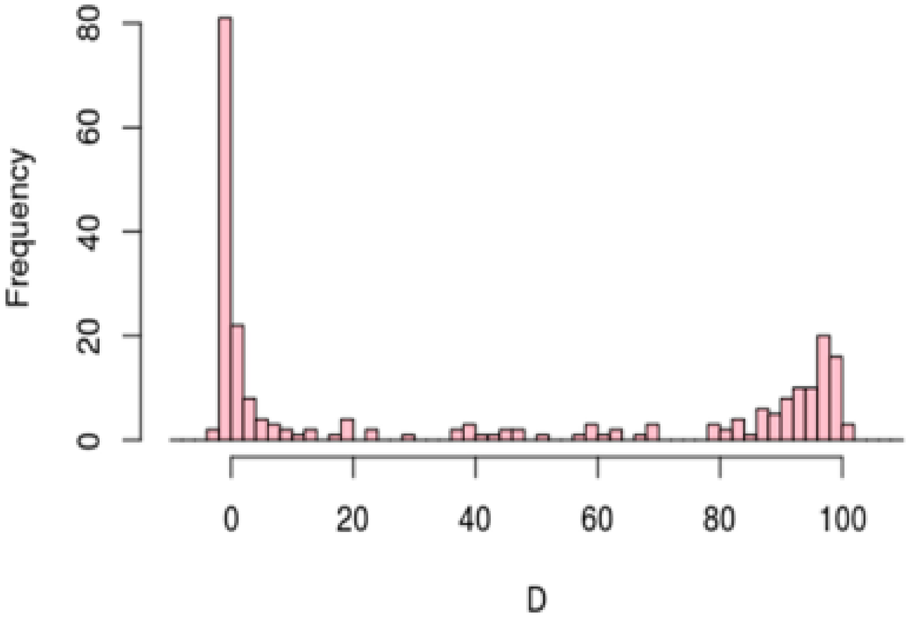


Figure 3. Distribution of CDS length-normalized positions (*D*) of TPS conserved between *H. salinarum* NRC-1 and *H. volcanii* DS2. Processing sites at start and stop codons are $D = 0$ and $D = 100$, respectively.

3.3. *H. salinarum Specific Internal TPS*

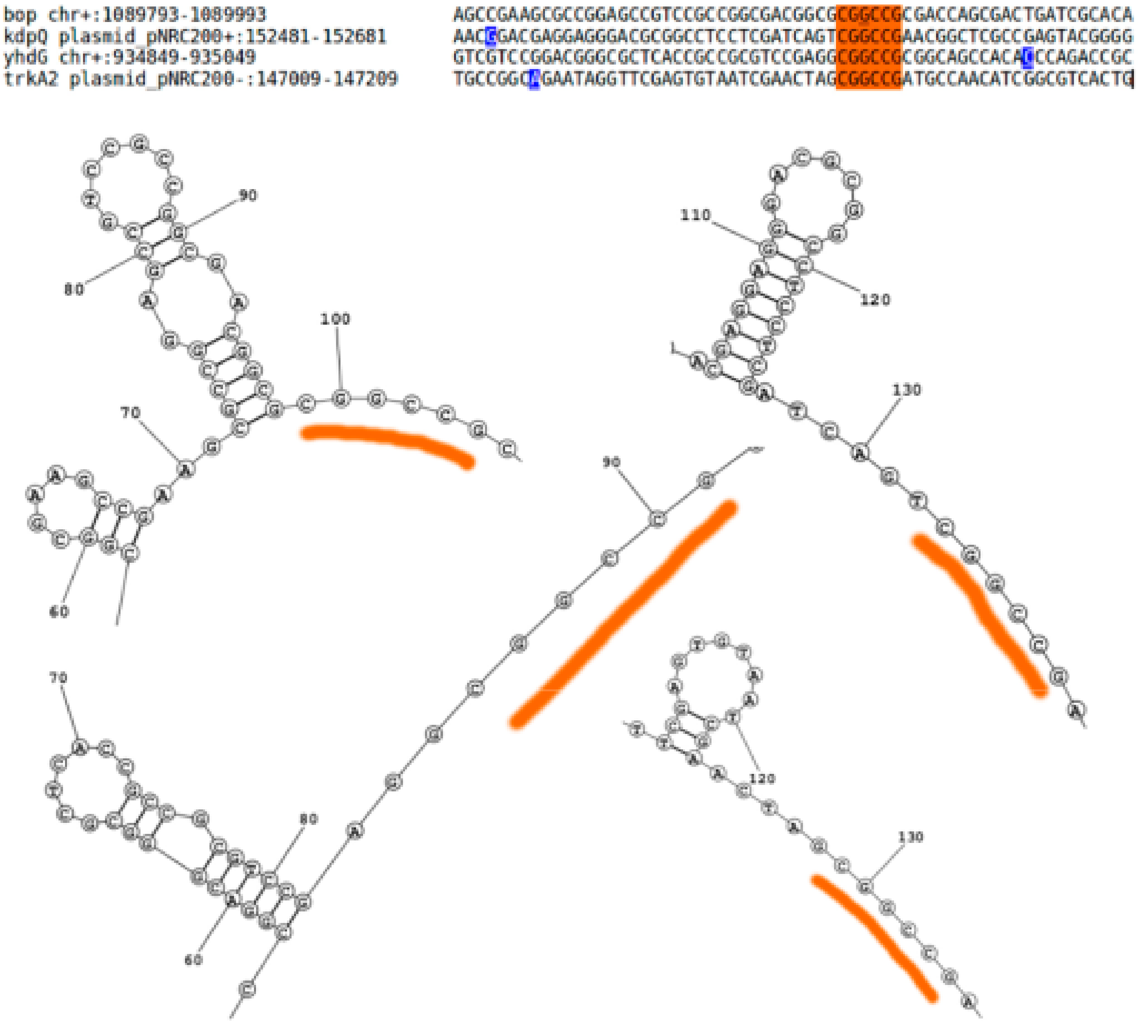
Since *H. salinarum* and *H. volcanii* are organisms that show important lifestyle differences, we then evaluated the orthologous genes with internal TPS present in one but not in the other, with special attention to salt homeostasis related genes, which may indicate differential post-transcriptional regulatory strategies.

Out of the 1554 orthologous gene pairs, 74 *H. salinarum* genes had internal TPS ($25 < D < 75$) with no counterpart in the *H. volcanii* equivalent region (Table S7). Reciprocally, 165 *H. volcanii* genes had internal TPS with no *H. salinarum* counterparts (Table S8). Gene set enrichment analysis showed no significant output among the *H. salinarum* specific internal TPS and just one molecular function in the *H. volcanii* case: ATPase activity, coupled to transmembrane movement of ions, rotational mechanism (GO:0044769, 12-fold enrichment, FDR < 2%).

From the *H. salinarum* specific internal TPS set, we highlight the *kef1* gene, which encodes a sodium transporter protein (TPS_13995_1, VNG_RS07995 *locus*, $D = 73$). We did not find an equivalent TPS in the moderate halophile *H. volcanii*, but there is evidence for an internal transcript in two other non-moderate halophilic archaea: *Natrinema* sp. J7-2 and *Haloquadratum walsbyi*. *H. walsbyi*, a hyperhalophilic organism isolated from a brine pool in Egypt [41], was probed under different illumination conditions by regular RNA-seq [42]. This experiment revealed a coverage signal peak around an equivalent position inside *kef1* (HQ_RS10745 *locus*, $D = 71$) (Figure S3b). *Natrinema* sp. J7-2 (formerly known as *H. salinarum* J7) is an extremophile isolated from a Chinese salt mine that requires, at least, a 10% and optimally 25% salt concentration to grow. Its transcriptome was studied by regular RNA-seq in three salinity conditions: low (15% NaCl, 2.6M), optimal (25% NaCl, 4.3M) and high (30% NaCl, 5.1M). Reanalysis of these data showed not only signals of an equivalent TPS inside *Natrinema* sp. J7-2's *kef1* gene (NJ7G_RS00730 *locus*, $D = 74$) but also that the cleavage output varies according to the environmental salt concentration (Figure S3c). We observed the same trend in our replication of this experiment in *H. salinarum* NRC-1: TPS_13995_1 increases > 4-fold in low salt relative to optimal growth condition (Onga et al., in prep.).

Given the importance of RNA processing to the salinity adaptation of *H. salinarum* [4], we investigated the association between internal TPS and the endoribonuclease activity of a regulatory RNase, namely VNG2099C (VNG_RS08138 *locus*). It was shown that the archaeal

RNase VNG2099C contributes to growth, the regulation of ion transport and is involved in environment-dependent physiologic transitions [4]. Four main genes highlighted in that study were relevantly dysregulated by VNG2099C deletion: *bop*, which encodes the light-driven proton-pump bacteriorhodopsin (VNG_RS05715 *locus*); *kdpQ*, encoding a positive auto-regulator for the Kdp potassium transport channel (VNG_RS11195); *trkA2*, encoding one regulatory subunit of a H$^+$-K$^+$ symporter (VNG_RS11170); and *yhdG*, encoding an ornithine-arginine antiporter (VNG_RS04855). We found TPS in *bop*, *kdpQ* and *trkA2*. Only *yhdG* did not have a TPS automatically identified by our dRNA-seq reanalysis since it is borderline statistically significant (TPS_3805_1, $p = 2.8\ 10^{-6} > 10^{-9}$), although clearly present (Figure S4d). A sequence-based common feature among these genes was not reported before and our own extensive multiple alignment or Markov chain-based motif search could not detect putative cleavage sites either. However, conjecturing that the TPS found inside those genes are related to VNG2099C's endonuclease activity, and performing structure-aware RNA multiple alignment [33] around these TPS sites, we were able to identify a signature: CGGCCG downstream of a strong stem–loop secondary structure (Figure 4).



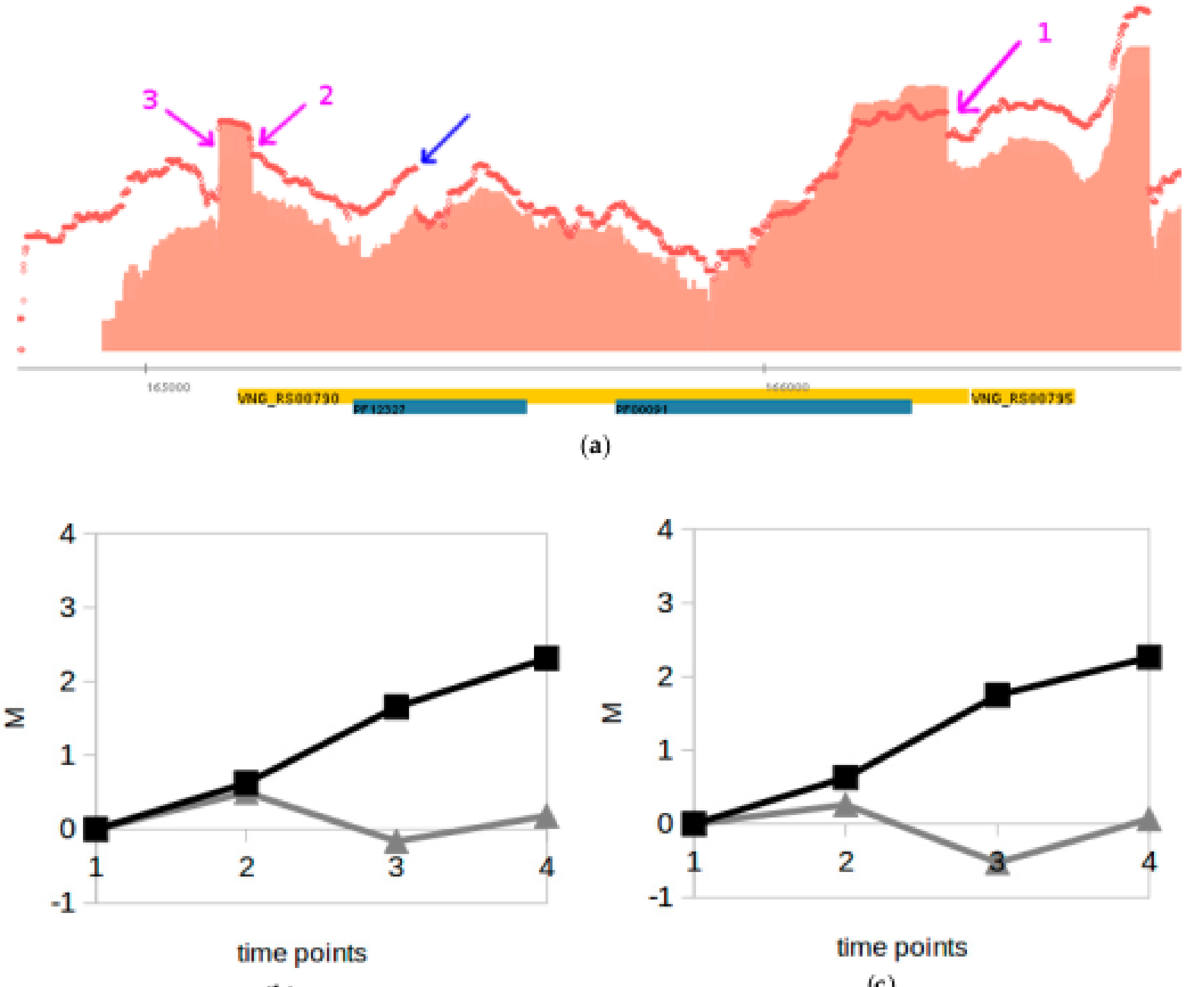
```
bop chr+:1089793-1089993                  AGCCGAAGCGCCGGAGCCGTCCGCCGGCGACGGCGCGGCCGCGACCAGCGACTGATCGCACA
kdpQ plasmid_pNRC200+:152481-152681       AACGGACGAGGAGGGACGCGGCCTCCTCGATCAGTCGGCCGAACGGCTCGCCGAGTACGGGG
yhdG chr+:934849-935049                   GTCGTCCGGACGGGCGCTCACCGCCGCGTCCGAGGCGGCCGCGGCAGCCACAGCCAGACCGC
trkA2 plasmid_pNRC200-:147009-147209      TGCCGGCAGAATAGGTTCGAGTGTAATCGAACTAGCGGCCGATGCCAACATCGGCGTCACTG
```

Figure 4. Putative signature found in dysregulated genes after VNG2099C RNase deletion using structure-aware RNA multiple alignment guided by TPS sites. Sequences from −100 to +100 around TPS (blue underlined bases) inside genes *bop*, *kdpQ*, *yhdG* and *trkA2* were used as alignment input. All genes harbor the CGGCCG motif downstream to a strong stem-loop (red highlights). The secondary structure predictions were filtered to report only base pairings with >0.99 probability (zoomed out version in Figure S4). Secondary structure predictions are, from top left to bottom right, for genes: *bop*, *kdpQ*, *yhdG* and *trkA2*.

Although it is not clear how the VNG2099C RNase mechanism generates these cuts, the sequence-structure putative signature paves the way for future research.

3.4. *TPS Relationship with Gene Expression during Growth*

H. salinarum dRNA-seq data were acquired at different time points across a typical growth curve experiment (Figure S5a), allowing a time-dependent comparison of processing products abundance. Individual inspection of all 244 conserved TPS at the exponential phase (17 h, OD600 $\approx$ 0.3) versus stationary phase (37 h, OD600 $\approx$ 0.5) showed that the vast majority of the transcripts starting at these positions follow the same expression patterns of their cognate full-length genes (File S3). Four notable exceptions (Figure 5 and Figure S5) are genes: *ftsZ2*, which encodes a cell division protein (TPS_8657_1, VNG_RS00790 *locus*); *eEF1A*, encoding an elongation factor (TPS_16108_1, VNG_RS10385); a putative arsenic resistance operon repressor encoded at the VNG_RS03675 *locus* (TPS_2832_1); and *pcn*, encoding a DNA polymerase III subunit (TPS_14733_1, VNG_RS08800).



(a)

(b)

(c)

Figure 5. Example of differential processing during *H. salinarum* growth. (**a**) The conserved TPS (TPS_8657_1, arrow 1) is near the start codon inside the *ftsZ2* gene (VNG_RS00790 *locus*) which encodes a cell division protein. Pfam domain annotation (blue rectangles) and coding sequences are in reverse strand (orange rectangles), thus, $5' \rightarrow 3'$ direction is right to left. Aligned reads coverage along genomic coordinates for TEX− libraries at exponential and stationary phases are shown in light red (solid) and red (dots), respectively ($\log_2$ counts normalized and arbitrarily jointly scaled). Magenta arrows 2 and 3 point to non-conserved TPS and blue arrow points to intraRNA. (**b**) *ftsZ2* gene $\log_2$ fold-change (M) between multi-modality measurements in different time points relative to the early exponential phase from [35] (time point 1: early exponential, 2: mid exponential, 3: late exponential, 4: stationary, squares: RNA-seq data, triangles: Ribo-seq data). (**c**) Same as (**b**) but for *cdrS* (VNG_RS00795 *locus*), encoding a *ftsZ2* regulator.

In halophilic archaea, the operon *cdrS-ftsZ2* is highly conserved and *cdrS* regulates *ftsZ2*, a tubulin homolog with a pivotal role in cell division [43]. A comparison among quantitative transcriptome (RNA-seq) and ribosome occupancy (Ribo-seq) data taken from the same *H. salinarum* growth time course [35] showed that this operon is probably subject to post-transcriptional regulation (Figure 5b,c). Although mRNA levels increase over time, as seen in the dRNA-seq dataset (Figure 5a), ribosome protected fragment levels indicated that they are not translated in the same proportion (Figure 5b). The TPS location inside *ftsZ2* near the 5′ end region, coinciding with a decrease in the RNA-seq signal in the stationary phase (Figure 5a), indicates that mRNA processing is involved in post-transcriptional regulation. The possibility of asRNA-mediated processing was investigated since in *E. coli*, the asRNA interaction with *ftsZ2* 5′-end regulates cell division [44]. The dRNA-seq data we reanalyzed here show clear signs of an asRNA presence in the vicinity of *ftsZ2* gene 5′-end in *H. volcanii* (HVO_RS07495) but not in *H. salinarum* (Files S1 and S2). Interestingly, there is an intraRNA [24] that increases its expression over time (Figure 5a, blue arrow) but the region flanked by two TPS, TPS_8628_1 and TPS_8625_1, remained constant (Figure 5a).

Although not the main focus here, there are differential processing cases in genes conserved between *H. salinarum* and bacteria, but not present in *H. volcanii*. An interesting example is *arcA* (VNG_RS11635), a gene involved in the arginine deiminase pathway, which allows fermentative arginine utilization by many bacteria and has been acquired by *H. salinarum* and not by *H. volcanii*. This gene was mentioned in a previous section to support the validation effort (Figure S2b). A previous study showed conditional independence of transcription inside the *arcRACB* cluster with clear transcription factor binding sites just upstream to all the genes [45]. Here, we noticed that along such "synthesis" signs, there are also "breaking" signs with TPS spread along the cluster, specially at 5′-ends inside or outside the coding regions. We speculate that the equilibrium between binding and processing would calibrate the appropriate *arcRACB* stoichiometry according to environmental conditions. In spite of the great increase in *arcA* transcript abundance along the growth curve, the signal pattern around the TPS is different (Figure S6a).

3.5. Antisense Transcript Processing Site (aTPS) Mapping

We identified antisense transcription processing sites (aTPS) in *H. salinarum* NRC-1 and *H. volcanii* DS2, expecting that a subset of them would be signatures of sense/antisense hybridization followed by cleavage events. A total of 265 and 629 aTPS were found in *H. salinarum* and *H. volcanii*, respectively (Table S9 and Table S10).

From a sample of 8 chromosomal asRNAs validated by Northern blot and conjectured to be processed products in *H. volcanii* [46], 6 are consistent with aTPS found here: HVO_2293, HVO_2990, HVO_1027, HVO_0263, HVO_0599, HVO_2928. This validation result, together with what we observed for regular TPS, is an indication that our approach is also reasonable for aTPS finding.

Comparative transcriptomics considering Escherichia coli str. K-12 substr. MG1655, for which sense/antisense interaction was established [47] indicates that aTPS can point to double-stranded RNA (dsRNA). An example of such a phenomenon is found within the stress-related CspA (CSD domain containing) family genes. All 4 of the H. volcanii chromosomal cspA genes presented at least one aTPS along with H. salinarum cspA1 (VNG_RS00395 locus). In E. coli, sense/antisense hybrids were experimentally observed for cspD, cspE, cspF, cspH, members of CspA family that are not directly involved in the cold shock response [47]. Similarly, in H. mediterranei, Northern blot data for arcA gene show a remarkable co-localization between experimentally detected long-processed product and an aTPS counterpart in H. salinarum (Figure S2b, TPS_19346_1) (this gene is absent in H. volcanii). Evidently, the presence of an aTPS is not a definitive hallmark of asRNA interaction, but rather is additional, circumstantial evidence.

A search for positional coincidence within a 5 nt window between TPS and aTPS as a potential signature of the sense/antisense interaction returned very few candidates: 18 and 48 pairs in H. salinarum and H. volcanii, respectively (Tables S11 and S12). Antisense transcripts are notoriously non-conserved and that is also the case between H. salinarum and H. volcanii [29]. Imposing that aTPS must be present in orthologous gene pairs, we found only 5 genes: dppA2, iscU, nosF1, atpI and VNG_RS08220 (loci pairs VNG_RS09900-HVO_RS07725, VNG_RS09660-HVO_RS05210, VNG_RS09315-HVO_RS19230, VNG_RS08320-HVO_RS06205, VNG_RS08220-HVO_RS06290, respectively). The case of dppA2 (alternatively, dppD), part of a dipeptide ABC transporter dppFABC operon, is particularly interesting since its constituents were shown to be conditionally regulated in H. salinarum (Figure 4D in [45]). Eventual decoupling between dppF and downstream dppABC was attributed to environment-dependent transcription factor binding combinations. However, checking regulatory and transcriptional data (raw intensity data from [45] and module hc2556 from [14]) we see that dppA2 can also be decoupled from dppBC. Our TPS analysis suggests that there is a dppABC cleavage and the aTPS presence allows us to speculate that this could be an asRNA-mediated process. Levels of dppA2, dppB2 and dppC2 rise along the growth curve in dRNA-seq and RNA-seq data (Figure 6) but available observable transcripts diminish around both TPS and aTPS regions (Figure 6a). Blunt sequenced antisense read edges match well the sense read 3′-ends (Figure 6c), consistent with a double-strand hybridization model.

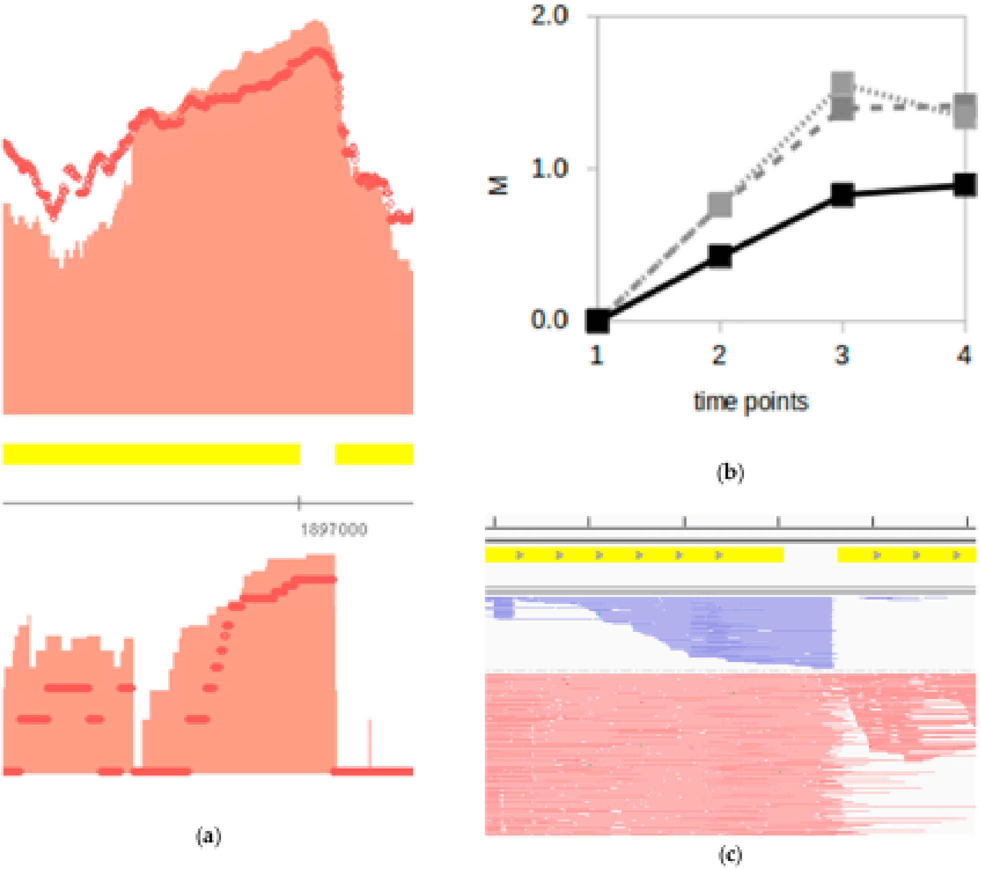


Figure 6. Example of sense/antisense processing during *H. salinarum* growth. (**a**) TPS (TPS_7176_1, top panel) and aTPS (TPS_15790_1, bottom panel) are located between *dppA2* and *dppB2* (VNG_RS09900 and VNG_RS09905 *loci*, 600 bp zoomed in out of a ~1.7 kbp gene), which encode members of a dipeptide ABC transporter system. Coding sequences are in forward strand (yellow rectangles) thus 5′→3′ direction is left to right, upper panel shows dRNA-seq data for forward strand and lower panel for reverse strand. (**b**) *dppA2*, *dppB2* and *dppC2* gene $\log_2$ fold-change (M) between transcriptome data in different time points relative to the early exponential phase from [35] (time point 1: early exponential, 2: mid-exponential, 3: late exponential, 4: stationary, filled black: *dppA2*, dashed grey: *dppB2* data, dotted light grey: *dppC2*). (**c**) Sample of aligned reads in the same window of (**a**) but not log scaled. Reads aligned to the reverse (antisense to *dppA2*) and forward strands are shown in blue and red, respectively (reads in red cropped vertically and horizontally for clarity since they are ~10-fold more abundant in this window). Tick marks are 100 bp apart.

3.6. TPS Associated with Ribosome Dynamics

From all 244 conserved TPS, at least 157 (64%) are inside or closely adjacent to *H. salinarum* Ribo-seq peaks (Table S13). Examples of such TPS/Ribo-seq coincidences are shown in Figure S7 for a few arbitrarily selected genes and the complete dataset can be explored using the File S4.

Interestingly, *pan1* (VNG_RS01995), which encodes the PAN-A proteasome-activating nucleotidase, has two alternative transcripts translated in two protein isoforms in *H. salinarum* [48] that correspond to the TPS found. The difference between both transcripts (Figure S7a) roughly matches the Ribo-seq peak and is consistent with conserved TPS locations in both *H. salinarum* (TPS_9619_1) and *H. volcanii* (TPS_058741, HVO_RS08770) considering the nominal positional uncertainty. Previous searches for ribosome associated ncRNAs (rancRNAs) in *H. volcanii* [49] led to the discovery of 856 *loci* internal to annotated CDS from which 56 (7%) coincide with the TPS hereby presented (Table S14). One of those rancRNAs is located between *pan1* transcript isoforms and we speculate that it may be involved in the choice for protein isoforms in both organisms.

Since Ribo-seq profiles along genomic coordinates can be reflexes of prolonged pauses or stalled ribosomes during the translation process [50], we hypothesize that these TPS found in the vicinity of Ribo-seq peaks may be related to "No-Go" decay [51].

3.7. TPS Identify ncRNAs Derived from Insertion Sequence Transcripts

Originally described in the extremophile model archaeon *H. salinarum* NRC-1, sot0042 RNA family (Rfam accession RF02656) is now annotated in 54 other species (144 sequences), mostly Halobacteria. These RNAs are transcribed from within insertion sequences (IS) in the same strand as their protein-coding genes. The previous effort to establish the sot0042 RNA family was not able to elucidate its biogenesis [52]. Given that Northern blot experiments detected VNG_sot0042 but not the cognate protein coding *tnpB* (VNG_RS00160 *locus*) full-length transcripts, the processing scenario was further investigated here.

A clear TEX+ < TEX− depletion signal (TPS_8065_1) is detected coinciding with the VNG_sot0042 start site, supporting the processing biogenesis hypothesis (Figure 7a). This observation is not restricted to the sot0042 RNA family but rather applies to almost all other sense overlapping RNAs (sotRNAs) previously found. Of all 10 sotRNAs previously identified near IS 3′ ends in *H. salinarum* [52], seven show clear TEX+ depletion as evidence of being processing products of their cognate IS transcripts: VNG_sot0026 (TPS_7956_1), VNG_sot0042 (TPS_8065_1), VNG_sot0044 (TPS_263_1), VNG_sot0286 (TPS_1226_1), VNG_sot6181 (TPS_20588_1), VNG_sot6221 (TPS_20696_1) and VNG_sot6361 (TPS_19531_1). From the remaining three out of 10 cases, VNG_sot0013 and VNG_sot2652 showed TEX+ depletion evidence (Figure S8) but were not automatically identified by our statistical pipeline. Generally, the *loci* coordinates inferred originally using sRNA-seq data [37] agreed with the current dRNA-seq based TPS analysis within a <5 nt margin. Analysis of public *H. volcanii* dRNA-seq data showed the same results (Figure 7b) for the sot0042 family member harbored within the HVO_2075 *locus* (TPS_077749).

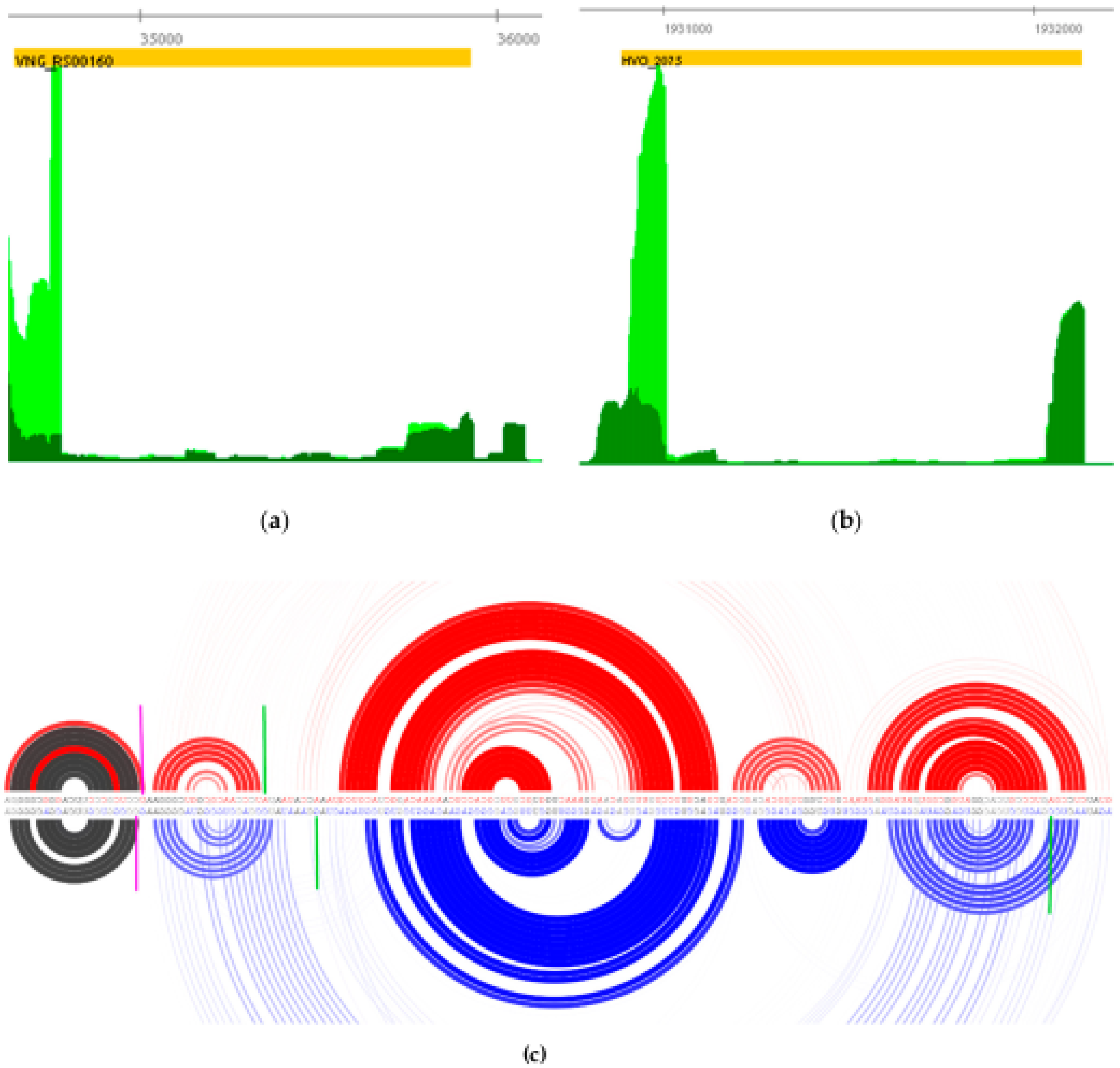


Figure 7. Gene *tnpB* transcript processing site signal (TEX− > TEX+) in (**a**) *H. salinarum* and (**b**) *H. volcanii* dRNA-seq datasets. Coding sequences are in reverse strand (orange rectangle), thus 5′→3′ direction is right to left. Aligned reads coverage along genomic coordinates for TEX+ and TEX− are shown in dark green and light green, respectively (normalized counts arbitrarily scaled). Statistical analysis showed the same results for the sot0042 family (RF02656) member within the VNG_RS00160 *locus* (TPS_8065_1) and HVO_2075 *locus* (TPS_077749). (**c**) Predicted rainbow diagram of sot0042 family representatives in *H. salinarum* (top, red) and *H. volcanii* (bottom, blue), 5′→3′ direction is right to left. Arc thickness is proportional to the thermal average probabilities of base pairs. Magenta markers point to hairpin induced false positive TSS. Green markers indicate transcript processing sites (TPS) inferred by statistically significant signal suppression after TEX treatment (right to left): TPS_077749, TPS_077728, TPS_8058_1.

Our observations are not limited to Haloarchaea but rather can also be seen in Thermococci and even in Bacteria. Our reanalysis of *Thermococcus kodakaraensis* KOD1 dRNA-seq data (PRJNA242777) shows strong evidence for sotRNA processing biogenesis for two out of all six IS*605*-related annotated transposases [52] (Figure S9). Similarly, reanalysis of several bacterial dRNA-seq data shows TPS inside the IS *locus* near the characteristic transposase DNA-binding protein domain OrfB_Zn_ribbon (PFAM database accession: PF07282). Considering over 950 bacterial species currently annotated as presenting PF07282 as their last C-terminal protein domain, at least five also had their primary transcriptome investigated by dRNA-seq: *Escherichia coli* K-12 (PRJNA238884), *Helicobacter pylori* 26,695 (PRJNA343039), *Streptomyces coelicolor* M145 (PRJNA285265), *Synechocystis* sp. PCC 6803 substr. GT-I (PRJNA224696) and *Mycobacterium tuberculosis* H37Rv (PRJEB1807). All of these five bacteria showed a clear TPS signal approximately at the same region near the PF07282 3′-end boundary (Figure S10), a feature also shown in our own data for the extremophile archaeon *H. salinarum*.

4. Discussion

The present bioinformatics study is embedded in the spirit of maximum data reuse. The identification of transcript processing sites (TPS) in *H. salinarum* NRC-1 and *H. volcanii* DS2 was achieved by reanalyzing public dRNA-seq data. Usually, dRNA-seq is performed to point to TEX+ > TEX− signals. However, using the same underlying statistical model, it is possible to additionally search for TEX− > TEX+ signals, yielding processing sites. Therefore, we used a well-established tool, TSSAR, forcefully inverting the labels of treated and untreated libraries to map putative TPS sites using the same published datasets. We validated this approach using published Northern blot experiments (Figure 2 and Figure S2) originally performed to support other scientific questions. The observed hybridization bands match well the putative transcripts with breakpoints at the TPS now identified.

TPS maps were made available through interactive files for the Gaggle Genome Browser (GGB), which renders navigation-enabled versions of Figures 2a, 5a, 6a and 7 and so on. The GGB platform is easy to use, and further instructions are available at the companion website http://labpib.fmrp.usp.br/~amr/tps/ (accessed on 27 May 2021) along with dRNA-seq, Ribo-seq and the differential expression publicly available data used in this work in GGB format (Files S1 to S4). This is the main result of our work: a resource that can pave the way for diverse downstream research initiatives.

The majority of TPS are concentrated around 5′ ends. Both halophiles have a relatively high rate of leaderless transcripts (≈70%) [24,25], which is consistent with the observed low frequency of $D < 0$ cases. A second population is located around 3′ ends, especially in *H. volcanii*. Small- and large-scale studies on archaeal 3′ UTR indicated the regulatory relevance of such regions [53,54]; however, the majority of TPS found at 3′ ends are contained within CDS and not at UTR ($D < 100$). In bacteria, 3′ ends can also work as a reservoir of regulatory sRNAs [55]. Many of the $D > 100$ cases are located between consecutive genes in operons. This is consistent with what was found in *E. coli* using Term-seq data [56], where mRNA processing of genes within an operon creates differential transcript abundances for equally transcribed operon members. Differential transcription initiation of operon members was reported before in *H. salinarum* [45] but differential decay has not been observed [5]. However, an example of an intra-operonic differential processing case was shown in (dppFABC operon, Figure 6). RNA half-life measurements using massive sequencing platforms in multiple environmental conditions may resolve this issue. Alternatively, and as a possible technical limitation, it is possible that such 5′-end enrichment of TPS would be simply a reflection of sequencing selection bias.

Our approach detects transcripts or segments of transcripts that were cleaved from longer precursors, but it cannot elucidate the mechanism or biogenesis of such a molecular break. It is impossible to know whether detected transcripts are a result of stalled 5′→3′ exoribonuclease digestion or endoribonuclease cleavage since the final observable is the same. Exoribonucleases acting in the 3′→5′ direction are not expected to generate an enrichment of reads all starting at the same genomic position. The internal TPS, on the other hand, are expected to be results of endonuclease activity.

We carried out extensive unfruitful searches for sequence motifs that could unify all internal TPS. No specific endoribonucleolytic cleavage signal that initiates mRNA decay is known in spite of endoribonuclease-high conservation in archaea [1]. Therefore, it is reasonable to assume that more than one RNase and more than one mechanism are operational at the same time. Our approach is not able to discriminate each case. We hypothesize that much of the difficulty in identifying sequence targets for endonucleolytic RNases may be explained by the possibility that no specific sequence motif is recognized, but rather structural motifs. However, TPS could aid the localization of putative cleavage motifs if integrated with additional data: we managed to find a putative RNase cleavage signature related to physiologically relevant regulatory action, using published RNase deletion information (Figure 4) [4]. The principled sub-selection of sequences surrounding the TPS mapped in this work may be the starting point to find missing signatures.

We hypothesize that TPS located approximately at the same relative positions in orthologous genes between *H. salinarum* and *H. volcanii* could be functionally relevant or a product of a conserved post-transcriptional regulation mechanism. The distribution of conserved TPS locations (Figure 4) is similar to overall sets from both organisms (Figure 2).

Iron–sulfur cluster assembly, a virtually omnipresent process due to its importance in providing protein co-factors, was found to present TPS more often than other biological processes. In bacteria, the iron–sulfur assembly protein concentrations are regulated post-transcriptionally [57]. It is reasonable to hypothesize that the TPS found overrepresented in iron–sulfur cluster assembly-related genes are evidence of relevant post-transcriptional regulation. Relevant to their lifestyle, active transmembrane ion transport was also found to resort to post-transcriptional regulation more than other molecular activities, even when taking into account the bias over orthologous genes between *H. salinarum* and *H. volcanii*. The same molecular activity was found to be overrepresented when considering *H. volcanii*-specific TPS, which may indicate that active ion transport by a rotational mechanism is tightly controlled in this organism relative to *H. salinarum*.

One of the main differences between *H. salinarum* NRC-1 and *H. volcanii* DS2 is the gas vesicle system, present in the former but not in the latter. This system, composed by at least two structural proteins (GvpA and GvpC) and several regulatory proteins, allows some halophilic organisms to float and search for optimal conditions along the water column [40,58]. *H. volcanii* is often used as a model to study gas vesicle system genes through transformation and heterologous expression [15]. We found some TPS inside the main gas vesicle system genes and conjecture that the post-transcriptional regulation layer may not be completely captured, using the complemented "natural null mutant" *H. volcanii* model.

Our temporal analyses of *H. salinarum* dRNA-seq data revealed cases of differential processing during growth. Although the vast majority of TPS breakpoint signals follow the same pattern of the overall gene, i.e., increased signal when the overall gene expression signal increases and vice versa, a few cases display opposite patterns, i.e., decrease in transcript abundance at TPS breakpoint while the cognate genes increase expression and vice versa (Figure S5 and Figure 6). We highlighted *ftsZ2*, a gene involved in cell division, which increased its expression from exponential to stationary phases and at the same time decreased transcript cleavage at the conserved TPS breakpoint (Figure 5). The elongation factor *eEF1A*, on the other hand, showed a negligible decrease in expression levels as *H. salinarum* grew, but a marked breakpoint at the conserved TPS position, indicating an additional post-transcriptional effort to stop translation. Other non-conserved TPS exist, showing differential regulation (even within these highlighted examples) but we kept our focus on those shared between *H. salinarum* and *H. volcanii*. The source of this analysis is available, and growth-dependent post-transcriptional regulation candidate events can be explored in the interactive browser made available as supplemental data.

H. salinarum ribosome footprinting [35] allowed us to notice a trend between concentration of ribosome-bound RNA fragments and TPS breakpoints (Figure S6 and Figure 7). Often, we find colocalization between TPS and Ribo-seq local peaks (64% of conserved TPS). It is known from the Ribo-seq analysis field that the signal profile reflects the dynamics of an array of ribosomes translating an mRNA [59,60]. Internal signal peaks are often interpreted as ribosome pauses. We are not able to resolve whether the processing event pointed by the TPS is the following: (i) some sort of response to paused/stalled ribosomes, such as "No-Go" decay (NGD) [51,61]; or (ii) products of mRNA precursors interacting/interfering with ribosomes [62]. In both cases, local enrichment peaks are expected in Ribo-seq signals.

Finally, we kept advancing the understanding of ncRNAs associated with insertion sequences (IS). Our research group characterized a set of transcripts near the 3′ end of transposable elements belonging to the IS*1341* group [52]. These transcripts overlap the IS coding regions in the same orientation, harbor a strong predicted secondary structure signature and show expression patterns distinct from their cognate gene (often anti-correlated) in several experimental conditions. The IS*1341* group belongs to a very ancient archaeal transposable element family, IS*200*/IS*605*, and conservation of these discovered sense overlapping transcripts (sotRNAs) granted the establishment of two new RNA families in the Rfam database v14.4 [63], sot0042 and sot2652 (Rfam accessions: RF02656 and RF02657, respectively). The TPS mapping carried out here showed that transcription start sites (TSS) of sotRNAs are, in fact, processing sites (TPS) and, therefore, these ncRNAs are derived from larger precursors. Our observations are not limited to Haloarchaea but rather can also be detected in Thermococci and even in bacteria. Since the level of nucleotide sequence conservation among all these bacterial and archaeal organisms upstream or downstream of the TPS is very low, the interpretation of such positional coincidence with the transposase domain is still elusive to us. Taken together, these results allow us to recognize that sotRNAs, with sot0042 and sot2652 RNA families, being only particular cases, are processing products from harboring IS transcripts.

Future steps of this work include in silico and experimental efforts. It would be important to apply the straightforward dRNA-seq "inversion" to other organisms, including non-halophiles. Additional organisms are continuously being considered and results are made available at http://labpib.fmrp.usp.br/~rvencio/tpsdraft/ (accessed on 27 May 2021). Experimental follow-ups on post transcriptional regulation predictions, such as for the physiologically relevant gas vesicle system (Figure S2) or FtsZ2 cell division system (Figure 5), must be carried out at protein level to confirm decoupling between the RNA message and protein action in halophiles. High-throughput proteomics is the natural choice to search and validate post-transcriptional regulatory mechanisms pointed here. Structural (not sequence-based only)-motif finding guided by TPS (Figure 4) and RNase deletions may reveal post-transcriptional regulatory circuits.

5. Conclusions

Using an in silico comparative transcriptomics approach, we maximized the data value by reanalyzing published RNA-seq variants to gain understanding on transcript processing. We mapped transcript processing sites (TPS) in *H. salinarum* NRC-1 and *H. volcanii* DS2, and drew correlations with *H. salinarum* differential expression, RNase deletion and ribosome footprint data, especially for TPS conserved between both halophilic species. The interactive map provided as supplemental files can pave the way for experimental validation of some biological hypotheses raised in this work. Our intent is to provide a useful platform for the halophilic microorganism research community.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12071018/s1, Figure S1: Genes with highest TPS density in *H. salinarum* and *H. volcanii*. Figure S2: Experimental validation of predicted TPS. Figure S3: Example of internal TPS in salt regulation gene. Figure S4: Putative signature found in dysregulated genes after VNG2099C RNase deletion. Figure S5: Example of differential processing at TPS during *H. salinarum* growth. Figure S6: Example of translation affecting probably due to TPS during *H. salinarum* growth. Figure S7: Example of conserved TPS, Ribo-seq and rancRNA signal coincidence in *H. salinarum* and *H. volcanii*. Figure S8: Identification of processing site in sense overlapping transcripts VNG_sot0013 and VNG_sot2652. Figure S9: Identification of processing sites in sense overlapping transcripts in *T. kodakaraensis* IS*605* insertion sequence family. Figure S10: Identification of processing site in IS associated sense overlapping transcripts in Bacteria. Table S1: Transcript processing sites (TPS) identified in *Halobacterium salinarum* NRC-1. Table S2: Transcript processing sites (TPS) identified in *Haloferax volcanii* DS2. Table S3: Non-filtered TPS search in *H. salinarum* NRC-1. Table S4: Non-filtered TPS search in *H. volcanii* DS2. Table S5: Transcript processing sites (TPS) conserved between *H. salinarum* NRC-1 and *H. volcanii* DS2. Table S6: Gene Ontology functional enrichment analysis of genes with conserved TPS.

Table S7: Internal TPS present in *H. salinarum* with no correspondent in *H. volcanii*. Table S8: Internal TPS present in *H. volcanii* with no correspondent in *H. salinarum*. Table S9: Antisense transcript processing sites (aTPS) identified in *H. salinarum*. Table S10: Antisense transcript processing sites (aTPS) identified in *H. volcanii*. Table S11: TPS and aTPS pairs in *H. salinarum*. Table S12: TPS and aTPS pairs in *H. volcanii*. Table S13: Conserved TPS co-localized with local Ribo-seq peaks in *H. salinarum*. Table S14: Putative conserved rancRNAs between *H. salinarum* and *H. volcanii*. File S1: Gaggle Genome Browser file with dRNA-seq data and TPS positions in *H. salinarum* NRC-1. File S2: Gaggle Genome Browser file with dRNA-seq data, rancRNA and TPS positions in *H. volcanii* DS2. File S3: GGB file with TEX− data at exponential and stationary phases in *H. salinarum*. File S4: GGB file with Ribo-seq data and conserved TPS in *H. salinarum*.

Author Contributions: Conceptualization, T.K. and R.Z.N.V.; methodology, A.G.A.E.-R.I., R.Z.N.V. and A.P.R.L.; software, A.G.A.E.-R.I., R.Z.N.V. and A.P.R.L.; writing-original draft preparation, R.Z.N.V. and A.G.A.E.-R.I.; writing-review and editing, T.K., A.G.A.E.-R.I., A.P.R.L. and R.Z.N.V.; supervision, T.K.; project administration, T.K.; funding acquisition, T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants #2015/21038-1 and #2017/03052-2, São Paulo Research Foundation (FAPESP). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) - Finance Code 001. This study was financed in part by the Fundação de Apoio ao Ensino, Pesquisa e Assistência do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FAEPA).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank LabPIB and LaBiSisMi members for the many discussions and insights on transcript processing sites.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Clouet-D'Orval, B.; Batista, M.; Bouvier, M.; Quentin, Y.; Fichant, G.; Marchfelder, A.; Maier, L.K. Insights into RNA-processing pathways and associated RNA-degrading enzymes in Archaea. *FEMS Microbiol. Rev.* **2018**, *42*, 579–613. [CrossRef] [PubMed]
2. Auboeuf, D. Alternative mRNA processing sites decrease genetic variability while increasing functional diversity. *Transcription* **2018**, *9*, 75–87. [CrossRef]
3. Fischer, S.; Benz, J.; Späth, B.; Maier, L.K.; Straub, J.; Granzow, M.; Raabe, M.; Urlaub, H.; Hoffmann, J.; Brutschy, B.; et al. The archaeal lsm protein binds to small RNAs. *J. Biol. Chem.* **2010**, *285*, 34429–34438. [CrossRef]
4. Wurtmann, E.J.; Ratushny, A.V.; Pan, M.; Beer, K.D.; Aitchison, J.D.; Baliga, N.S. An evolutionarily conserved RNase-based mechanism for repression of transcriptional positive autoregulation. *Mol. Microbiol.* **2014**, *92*, 369–382. [CrossRef]
5. Hundt, S.; Zaigler, A.; Lange, C.; Soppa, J.; Klug, G. Global analysis of mRNA decay in Halobacterium salinarum NRC-1 at single-gene resolution using DNA microarrays. *J. Bacteriol.* **2007**, *189*, 6936–6944. [CrossRef]
6. Qi, L.; Yue, L.; Feng, D.; Qi, F.; Li, J.; Dong, X. Genome-wide mRNA processing in methanogenic archaea reveals post-transcriptional regulation of ribosomal protein synthesis. *Nucleic Acids Res.* **2017**, *45*, 7285–7298. [CrossRef] [PubMed]
7. Birkedal, U.; Beckert, B.; Wilson, D.N.; Nielsen, H. The 23S Ribosomal RNA From *Pyrococcus furiosus* Is Circularly Permuted. *Front. Microbiol.* **2020**, *11*. [CrossRef]
8. DasSarma, S.; Berquist, B.R.; Coker, J.A.; DasSarma, P.; Müller, J.A. Post-genomics of the model haloarchaeon *Halobacterium* sp. NRC-1. *Saline Syst.* **2006**, *2*, 3. [CrossRef]
9. Gunde-Cimerman, N.; Plemenitaš, A.; Oren, A. Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS Microbiol. Rev.* **2018**, *42*, 353–375. [CrossRef] [PubMed]
10. Engel, M.B.; Catchpole, H.R. A microprobe analysis of inorganic elements in Halobacterium salinarum. *Cell Biol. Int.* **2005**, *29*, 616–622. [CrossRef]
11. Henderson, R.; Baldwin, J.M.; Ceska, T.A.; Zemlin, F.; Beckmann, E.; Downing, K.H. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **1990**, *213*, 899–929. [CrossRef]
12. DasSarma, P.; Negi, V.D.; Balakrishnan, A.; Karan, R.; Barnes, S.; Ekulona, F.; Chakravortty, D.; DasSarma, S. *Haloarchaeal gas* vesicle nanoparticles displaying *Salmonella* SopB antigen reduce bacterial burden when administered with live attenuated bacteria. *Vaccine* **2014**, *32*, 4543–4549. [CrossRef] [PubMed]

13. Bonneau, R.; Facciotti, M.T.; Reiss, D.J.; Schmid, A.K.; Pan, M.; Kaur, A.; Thorsson, V.; Shannon, P.; Johnson, M.H.; Bare, J.C.; et al. A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell. *Cell* **2007**, *131*, 1354–1365. [CrossRef] [PubMed]
14. Brooks, A.N.; Reiss, D.J.; Allard, A.; Wu, W.; Salvanha, D.M.; Plaisier, C.L.; Chandrasekaran, S.; Pan, M.; Kaur, A.; Baliga, N.S. A system-level model for the microbial regulatory genome. *Mol. Syst. Biol.* **2014**, *10*, 740. [CrossRef] [PubMed]
15. Soppa, J. Functional genomic and advanced genetic studies reveal novel insights into the metabolism, regulation, and biology of *Haloferax volcanii*. *Archaea* **2011**, *2011*. [CrossRef]
16. Hartman, A.L.; Norais, C.; Badger, J.H.; Delmas, S.; Haldenby, S.; Madupu, R.; Robinson, J.; Khouri, H.; Ren, Q.; Lowe, T.M.; et al. The complete genome sequence of Haloferax volcanii DS2, a model archaeon. *PLoS ONE* **2010**, *5*, e9605. [CrossRef] [PubMed]
17. Lander, E.S. The Heroes of CRISPR. *Cell* **2016**, *164*, 18–28. [CrossRef] [PubMed]
18. Gelsinger, D.R.; Diruggiero, J. The non-coding regulatory RNA revolution in archaea. *Genes* **2018**, *9*, 141. [CrossRef] [PubMed]
19. Hör, J.; Gorski, S.A.; Vogel, J. Bacterial RNA Biology on a Genome Scale. *Mol. Cell* **2018**, *70*, 785–799. [CrossRef]
20. Saliba, A.E.; Santos, S.C.; Vogel, J. New RNA-seq approaches for the study of bacterial pathogens. *Curr. Opin. Microbiol.* **2017**, *35*, 78–87. [CrossRef] [PubMed]
21. Sharma, C.M.; Vogel, J. Differential RNA-seq: The approach behind and the biological insight gained. *Curr. Opin. Microbiol.* **2014**, *19*, 97–105. [CrossRef]
22. Gill, E.E.; Chan, L.S.; Winsor, G.L.; Dobson, N.; Lo, R.; Ho Sui, S.J.; Dhillon, B.K.; Taylor, P.K.; Shrestha, R.; Spencer, C.; et al. High-throughput detection of RNA processing in bacteria. *BMC Genom.* **2018**, *19*. [CrossRef] [PubMed]
23. Yu, S.H.; Vogel, J.; Förstner, K.U. ANNOgesic: A Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *Gigascience* **2018**, *7*, 1–11. [CrossRef] [PubMed]
24. Ten-Caten, F.; Vêncio, R.Z.N.; Lorenzetti, A.P.R.; Zaramela, L.S.; Santana, A.C.; Koide, T. Internal RNAs overlapping coding sequences can drive the production of alternative proteins in archaea. *RNA Biol.* **2018**, *15*, 1119–1132. [CrossRef]
25. Babski, J.; Haas, K.A.; Näther-Schindler, D.; Pfeiffer, F.; Förstner, K.U.; Hammelmann, M.; Hilker, R.; Becker, A.; Sharma, C.M.; Marchfelder, A.; et al. Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genom.* **2016**, *17*. [CrossRef]
26. Ramírez, F.; Ryan, D.P.; Grüning, B.; Bhardwaj, V.; Kilpert, F.; Richter, A.S.; Heyne, S.; Dündar, F.; Manke, T. deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **2016**, *44*, W160–W165. [CrossRef]
27. Amman, F.; Wolfinger, M.T.; Lorenz, R.; Hofacker, I.L.; Stadler, P.F.; Findeiß, S. TSSAR: TSS annotation regime for dRNA-seq data. *BMC Bioinform.* **2014**, *15*, 89. [CrossRef]
28. Whiteside, M.D.; Winsor, G.L.; Laird, M.R.; Brinkman, F.S.L. OrtholugeDB: A bacterial and archaeal orthology resource for improved comparative genomic analysis. *Nucleic Acids Res.* **2013**, *41*, 366–376. [CrossRef]
29. de Almeida, J.P.P.; Vêncio, R.Z.N.; Lorenzetti, A.P.R.; Caten, F.; Gomes-Filho, J.V.; Koide, T. The Primary Antisense Transcriptome of Halobacterium salinarum NRC-1. *Genes* **2019**, *10*, 280. [CrossRef]
30. Gruber, A.R.; Lorenz, R.; Bernhart, S.H.; Neuböck, R.; Hofacker, I.L. The Vienna RNA websuite. *Nucleic Acids Res.* **2008**, *36*, 70–74. [CrossRef] [PubMed]
31. Aalberts, D.P.; Jannen, W.K. Visualizing RNA base-pairing probabilities with RNAbow diagrams. *RNA* **2013**, *19*, 475–478. [CrossRef] [PubMed]
32. Bellaousov, S.; Reuter, J.S.; Seetin, M.G.; Mathews, D.H. RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.* **2013**, *41*, W471. [CrossRef] [PubMed]
33. Will, S.; Joshi, T.; Hofacker, I.L.; Stadler, P.F.; Backofen, R. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA* **2012**, *18*, 900–914. [CrossRef] [PubMed]
34. Mi, H.; Muruganujan, A.; Ebert, D.; Huang, X.; Thomas, P.D. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **2019**, *47*, D419–D426. [CrossRef]
35. López García de Lomana, A.; Kusebauch, U.; Raman, A.V.; Pan, M.; Turkarslan, S.; Lorenzetti, A.P.R.; Moritz, R.L.; Baliga, N.S. Selective Translation of Low Abundance and Upregulated Transcripts in *Halobacterium salinarum*. *mSystems* **2020**, *5*. [CrossRef]
36. Bare, J.C.; Koide, T.; Reiss, D.J.; Tenenbaum, D.; Baliga, N.S. Integration and visualization of systems biology data in context of the genome. *BMC Bioinform.* **2010**, *11*, 382. [CrossRef]
37. Zaramela, L.S.; Vêncio, R.Z.N.; Ten-Caten, F.; Baliga, N.S.; Koide, T. Transcription start site associated RNAs (TSSaRNAs) are ubiquitous in all domains of life. *PLoS ONE* **2014**, *9*, e107680. [CrossRef]
38. Ruepp, A.; Soppa, J. Fermentative arginine degradation in Halobacterium salinarium (formerly *Halobacterium halobium*): Genes, gene products, and transcripts of the *arcRACB* gene cluster. *J. Bacteriol.* **1996**, *178*, 4942–4947. [CrossRef]
39. Laass, S.; Monzon, V.A.; Kliemt, J.; Hammelmann, M.; Pfeiffer, F.; Förstner, K.U.; Soppa, J. Characterization of the transcriptome of *Haloferax volcanii*, grown under four different conditions, with mixed RNA-Seq. *PLoS ONE* **2019**, *14*, e0215986. [CrossRef]
40. Jäger, A.; Samorski, R.; Pfeifer, F.; Klug, G. Individual gvp transcript segments in Haloferax mediterranei exhibit varying half-lives, which are differentially affected by salt concentration and growth phase. *Nucleic Acids Res.* **2002**, *30*, 5436–5443. [CrossRef]
41. Bolhuis, H.; Palm, P.; Wende, A.; Falb, M.; Rampp, M.; Rodriguez-Valera, F.; Pfeiffer, F.; Oesterhelt, D. The genome of the square archaeon Haloquadratum walsbyi: Life at the limits of water activity. *BMC Genom.* **2006**, *7*, 169. [CrossRef]
42. Bolhuis, H.; Martín-Cuadrado, A.B.; Rosselli, R.; Pašić, L.; Rodriguez-Valera, F. Transcriptome analysis of *Haloquadratum walsbyi*: Vanity is but the surface. *BMC Genom.* **2017**, *18*, 510. [CrossRef]

43. Darnell, C.L.; Zheng, J.; Wilson, S.; Bertoli, R.M.; Bisson-Filho, A.W.; Garner, E.C.; Schmid, A.K. The ribbon-helix-helix domain protein cdrs regulates the tubulin homolog ftsz2 to control cell division in archaea. *MBio* **2020**, *11*, 1–22. [CrossRef]
44. Dewar, S.J.; Donachie, W.D. Antisense transcription of the ftsZ-ftsA gene junction inhibits cell division in *Escherichia coli*. *J. Bacteriol.* **1993**, *175*, 7097–7101. [CrossRef]
45. Koide, T.; Reiss, D.J.; Bare, J.C.; Pang, W.L.; Facciotti, M.T.; Schmid, A.K.; Pan, M.; Marzolf, B.; Van, P.T.; Lo, F.Y.; et al. Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol. Syst. Biol.* **2009**, *5*, 285. [CrossRef] [PubMed]
46. Straub, J.; Brenneis, M.; Jellen-Ritter, A.; Heyer, R.; Soppa, J.; Marchfelder, A. Small RNAs in haloarchaea: Identification, differential expression and biological function. *RNA Biol.* **2009**, *6*, 281–292. [CrossRef] [PubMed]
47. Lybecker, M.; Zimmermann, B.; Bilusic, I.; Tukhtubaeva, N.; Schroeder, R. The double-stranded transcriptome of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3134–3139. [CrossRef]
48. Chamieh, H.; Guetta, D.; Franzetti, B. The two PAN ATPases from Halobacterium display N-terminal heterogeneity and form labile complexes with the 20S proteasome. *Biochem. J.* **2008**, *411*, 387–397. [CrossRef]
49. Wyss, L.; Waser, M.; Gebetsberger, J.; Zywicki, M.; Polacek, N. mRNA-specific translation regulation by a ribosome-associated ncRNA in *Haloferax volcanii*. *Sci. Rep.* **2018**, *8*, 1–13. [CrossRef]
50. Gelsinger, D.R.; Dallon, E.; Reddy, R.; Mohammad, F.; Buskirk, A.R.; DiRuggiero, J. Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribosome pausing at single codon resolution. *Nucleic Acids Res.* **2020**, *48*, 5201–5216. [CrossRef]
51. Atkinson, G.C.; Baldauf, S.L.; Hauryliuk, V. Evolution of nonstop, no-go and nonsense-mediated mRNA decay and their termination factor-derived components. *BMC Evol. Biol.* **2008**, *8*, 290. [CrossRef]
52. Gomes-Filho, J.V.; Zaramela, L.S.; Italiani, V.C.d.S.; Baliga, N.S.; Vêncio, R.Z.N.; Koide, T. Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea. *RNA Biol.* **2015**, *12*, 490–500. [CrossRef]
53. Dar, D.; Prasse, D.; Schmitz, R.A.; Sorek, R. Widespread formation of alternative 3′ UTR isoforms via transcription termination in archaea. *Nat. Microbiol.* **2016**, *1*, 16143. [CrossRef]
54. Brenneis, M.; Soppa, J. Regulation of Translation in Haloarchaea: 5′- and 3′-UTRs Are Essential and Have to Functionally Interact In Vivo. *PLoS ONE* **2009**, *4*, e4484. [CrossRef] [PubMed]
55. Miyakoshi, M.; Chao, Y.; Vogel, J. Regulatory small RNAs from the 3′ regions of bacterial mRNAs. *Curr. Opin. Microbiol.* **2015**, *24*, 132–139. [CrossRef]
56. Dar, D.; Sorek, R. Extensive reshaping of bacterial operons by programmed mRNA decay. *PLoS Genet.* **2018**, *14*, e1007354. [CrossRef]
57. Mettert, E.L.; Kiley, P.J. How Is Fe-S Cluster Formation Regulated? *Annu. Rev. Microbiol.* **2015**, *69*, 505–526. [CrossRef]
58. Pfeifer, F. Haloarchaea and the formation of gas vesicles. *Life* **2015**, *5*, 385–402. [CrossRef] [PubMed]
59. Richter, J.D.; Coller, J. Pausing on Polyribosomes: Make Way for Elongation in Translational Control. *Cell* **2015**, *163*, 292–300. [CrossRef] [PubMed]
60. Calviello, L.; Ohler, U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet.* **2017**, *33*, 728–744. [CrossRef]
61. De Koning, B.; Blombach, F.; Brouns, S.J.J.; Van Der Oost, J. Fidelity in archaeal information processing. *Archaea* **2010**, *2010*, 1–15. [CrossRef] [PubMed]
62. Pircher, A.; Gebetsberger, J.; Polacek, N. Ribosome-associated ncRNAs: An emerging class of translation regulators. *RNA Biol.* **2014**, *11*, 1335–1339. [CrossRef] [PubMed]
63. Kalvari, I.; Nawrocki, E.P.; Ontiveros-Palacios, N.; Argasinska, J.; Lamkiewicz, K.; Marz, M.; Griffiths-Jones, S.; Toffano-Nioche, C.; Gautheret, D.; Weinberg, Z.; et al. Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **2021**, *49*, D192–D200. [CrossRef] [PubMed]

The journal version of the manuscript requires the usage of supplemental figures due to page constraints. Since this PhD text do not have such constraints we include in the following such figures since they strongly help to illustrate all points and claims made in the manuscript. The figures are indexed according to the original published manuscript instead of following this PhD dissertation numbering to allow proper visualization at the right context during the manuscript reading flow.

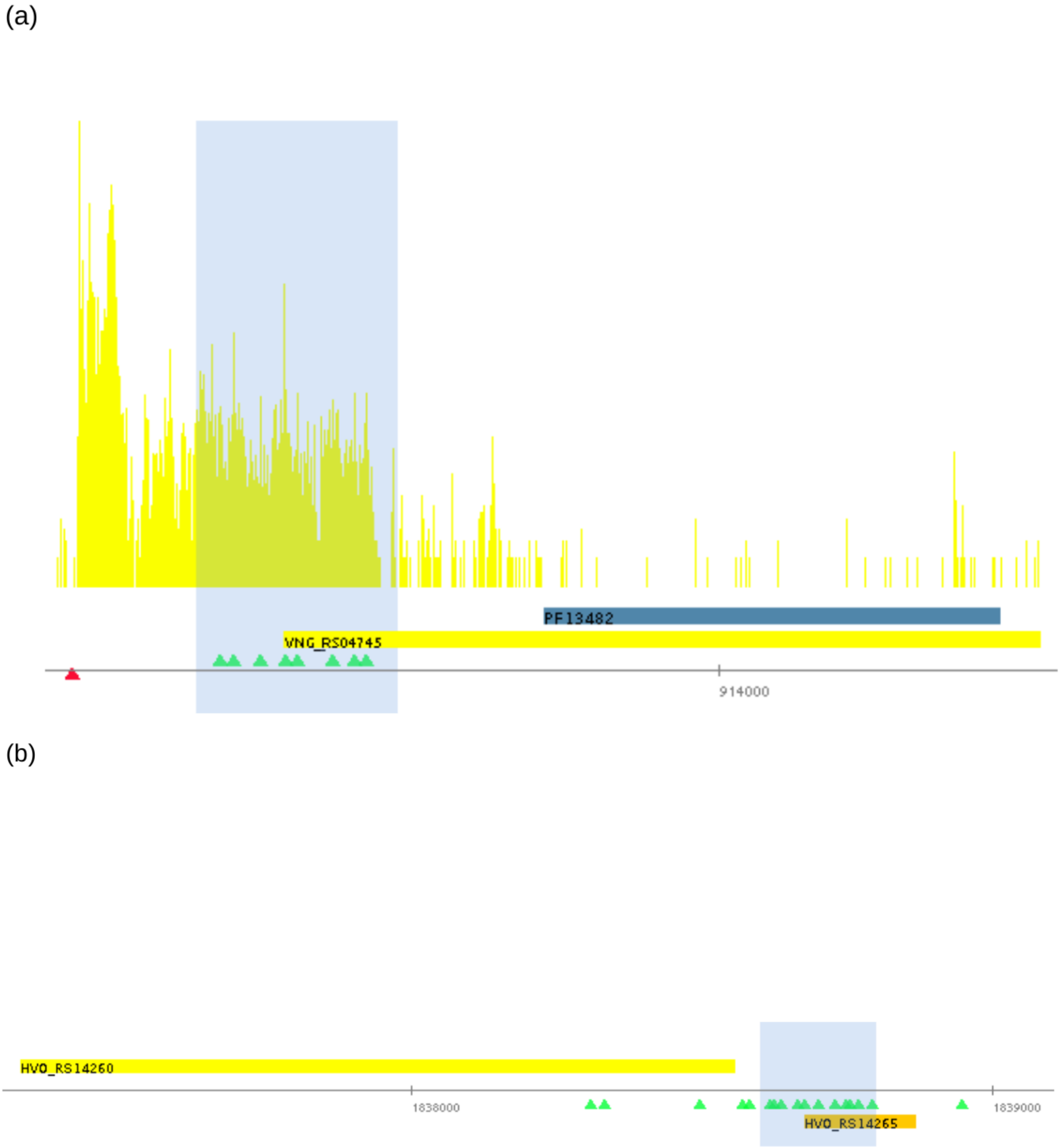
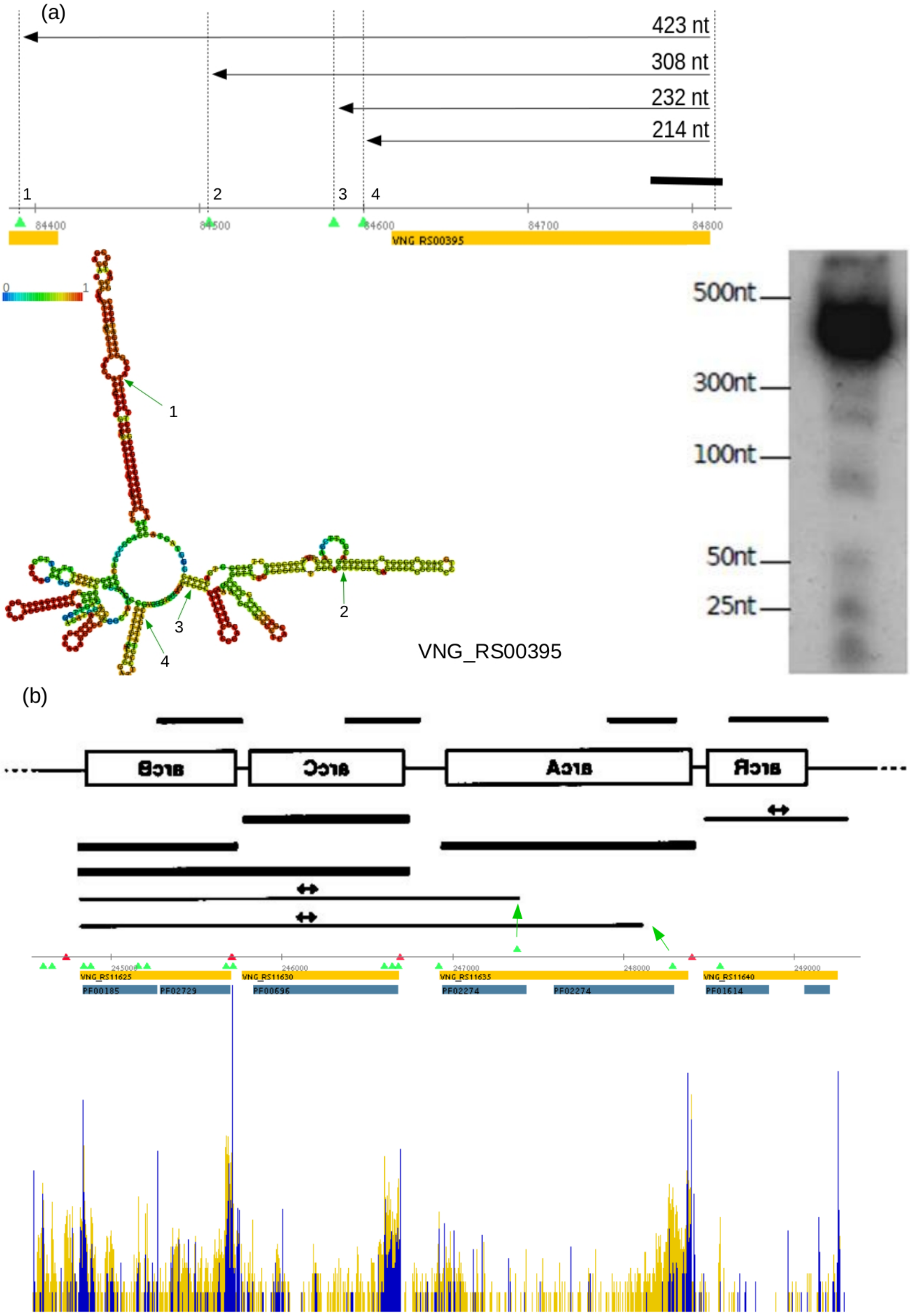
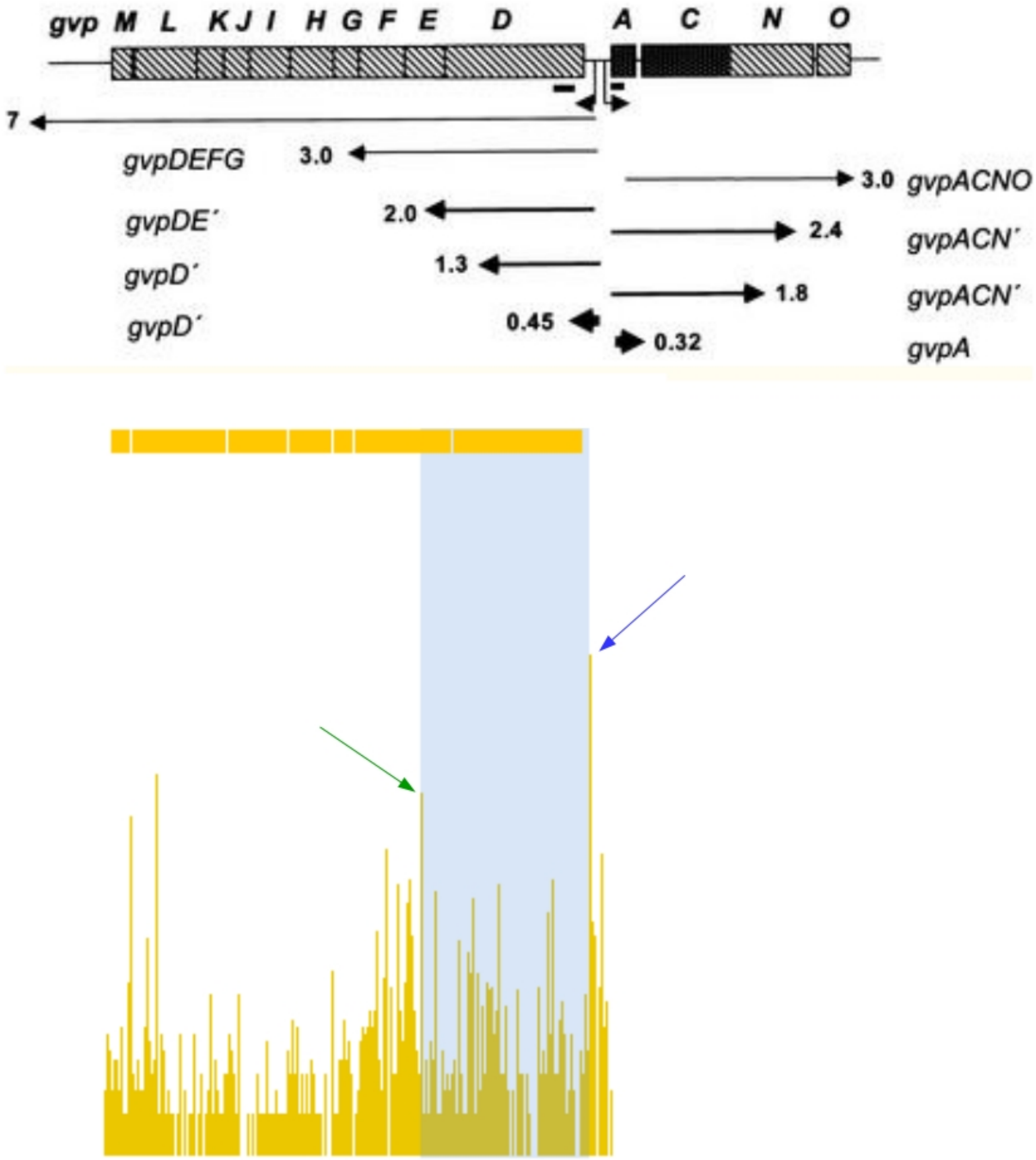


Figure S1 - Genes with highest TPS density in *H. salinarum* and *H. volcanii*.
(a) RNase H domain containing exoribonuclease (VNG_RS04745 *locus*) presents the highest TPS density in *H. salinarum*, 8 TPS (green triangles) in a 200 nt window (light blue highlight). Distribution of relative numbers of aligned reads starting at a given genomic coordinate in TEX- libraries is shown in yellow (arbitrarily log-scaled normalized counts). Pfam domain annotation (blue rectangle) and coding sequences are in forward strand (yellow rectangle) thus 5′→3′ direction is left to right. Red triangle marks ChIP-seq based binding site of TfbD transcription factor co-localized with genes' TSS (from Wilbanks *et al*., 2012). 5′ UTR is highly processed. **(b)** "cold-shock" *cspA4* gene (HVO_RS14265 *locus*) present the highest TPS density in *H. volcanii*, 11 TPS (green triangles) in a 200 nt window (light blue highlight). Coding sequence is in reverse strand (orange rectangle) thus 5′→3′ direction is right to left.


(a)
423 nt
308 nt
232 nt
214 nt
1
2
3
4
84400
84500
84600
84700
84800
VNG_RS00395
0
1
1
2
3
4
500nt
300nt
100nt
50nt
25nt
VNG_RS00395
(b)
arcB
arcC
arcA
arcR
245000
246000
247000
248000
249000
VNG_RS11625
PF00185
PF02729
VNG_RS11630
PF00696
VNG_RS11635
PF02274
PF02274
VNG_RS11640
PF01614

(c)


gvp M L K J I H G F E D A C N O
7
gvpDEFG 3.0
3.0 gvpACNO
gvpDE´ 2.0
2.4 gvpACN´
gvpD´ 1.3
1.8 gvpACN´
gvpD´ 0.45
0.32 gvpA

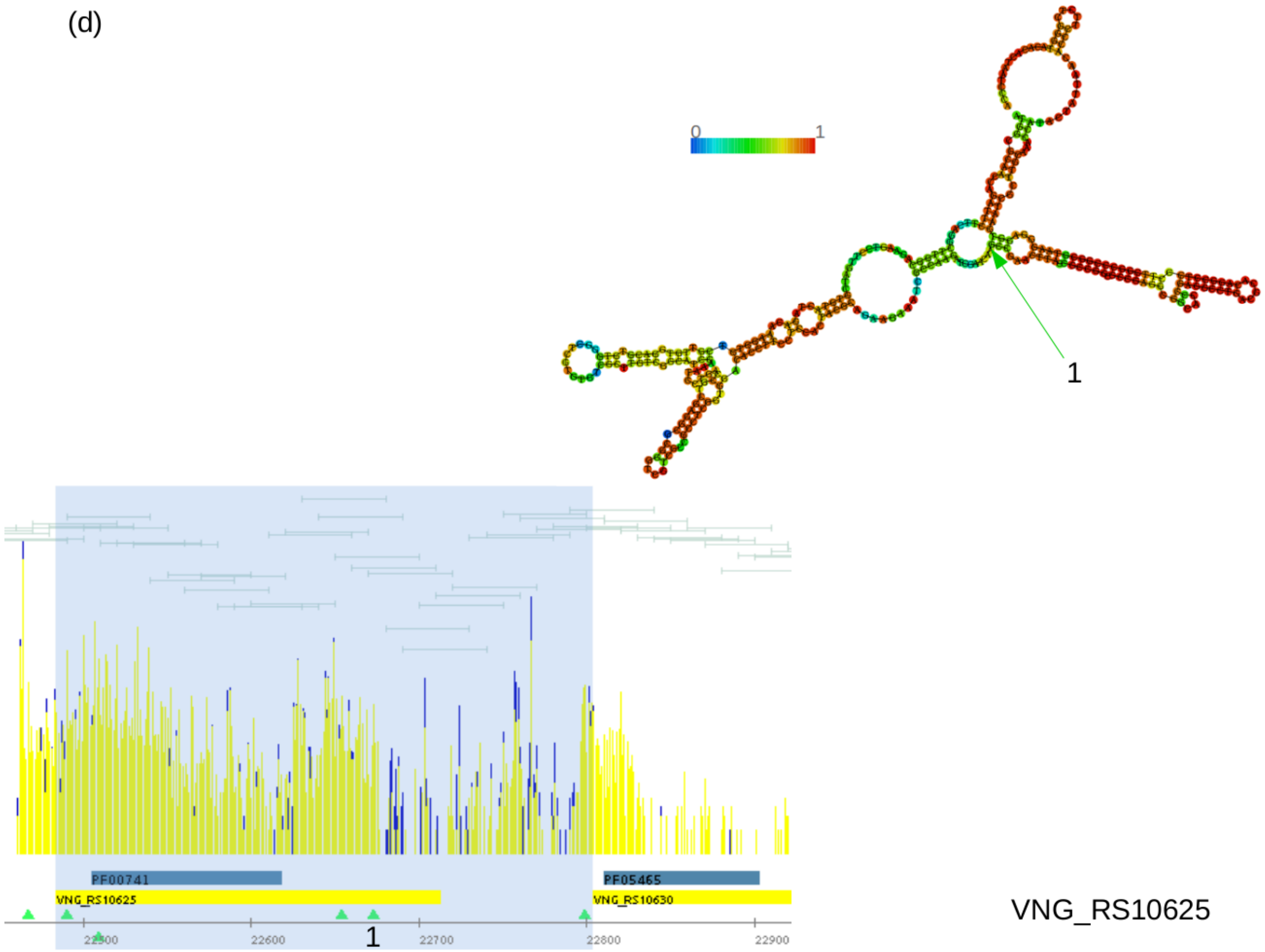
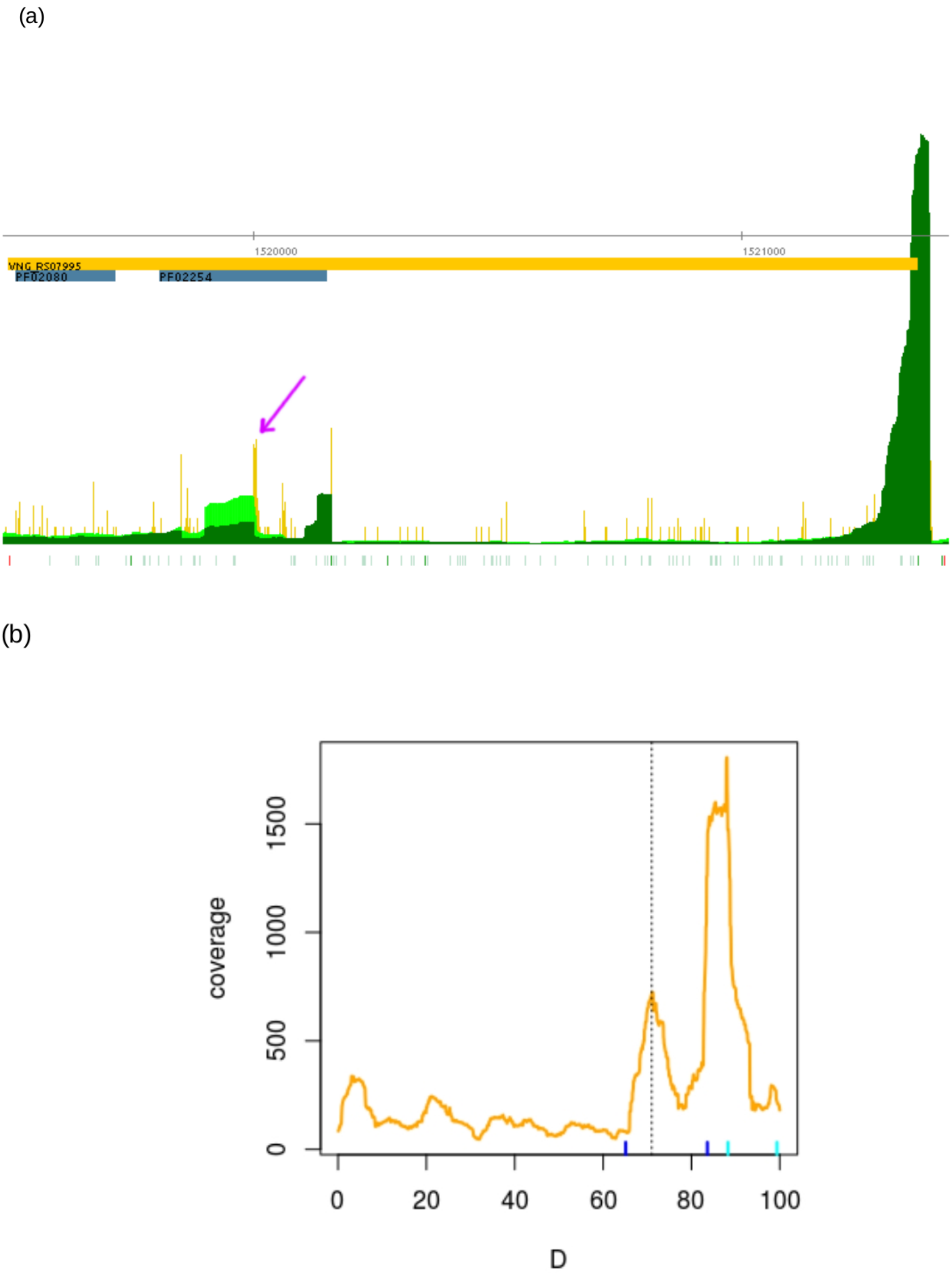


Figure S2 - Experimental validation of predicted TPS.

(a) Experimental validation of predicted TPS using gene *cspA1* (VNG_RS00395 *locus*). Consecutive TPS (green triangles) predict mRNAs of different sizes (arrows) from TSS to TPS_8361_1 (#4), TPS_8363_1 (#3), TPS_8365_1 (#2) and TPS_8366_1 (#1). Northern blot probe location is marked by the black rectangle. Right panel shows a northern blot published by our group (Zaramela *et al*., 2014) with bands consistent with transcripts cleaved at TPS sites. Lower panel shows secondary structure prediction for the larger transcript and each processing site location (numbered arrows). **(b)** Upper panel adapted from Figure 4C in Ruepp & Soppa (1996) where two-headed arrows mean inexact transcript boudaries in *H. mediterranei*. Lower panel shows an histogram using *H. salinarum* data. Distribution of relative numbers of aligned reads starting at a given genomic coordinate in TEX- and TEX+ libraries, shown in yellow and blue, respectively (arbitrarily log-scaled normalized counts). Pfam domain annotation (blue rectangle) and coding sequences are in reverse strand (orange rectangle) thus 5′→3′ direction is right to left. Green arrows point to the TPS-transcript correspondence: TPS_20943_1 and TPS_19346_1. Green triangles are TPS defined in the present work (see Supplemental File 1 to navigate them). Red triangle marks ChIP-seq based binding site of TfbD transcription factor co-localized with genes' TSS (from Wilbanks *et al*., 2012). **(c)** Conservation of TPS between *H. salinarum* and *H. mediterranei* explains the operon fragment observed in *H. mediterranei*. Upper panel taken directly from Jäger *et al*. (2002) with arrows representing observed transcripts in *H. mediterranei*. Lower panel shows *H. salinarum* dRNA-seq data. Operon representations are shown schematically approximately at scale aligned by CDS (rectangles) and TSS (blue arrow). Distribution of relative numbers of aligned reads starting at a given genomic coordinate in TEX- libraries is shown in yellow (arbitrarily log-scaled normalized counts). The *gvpDEFGHIJKLM* operon is in reverse strand in both organisms thus 5′→3′ direction is right to left. The TPS which would create an equivalent of *H. mediterranei*'s *gvpDE*′ (2.0 kb) in *H. salinarum* is TPS_19964_1 (green arrow). **(d)** TPS in *gvpA1*, encoding the most important structural protein in the gas vesicle system (VNG_RS10625). TPS (TPS_18335_1, green arrow) is located at the basis of a strong stem-loop structure involved in transcript stability. CDS are represented by yellow rectangles in forward strand (5′→3′ is left to right), Pfam domain annotations are denoted by blue rectangles. Distribution of relative numbers of aligned reads starting at a given genomic coordinate in TEX- and TEX+ libraries are shown as yellow and blue vertical lines, respectively (arbitrarily log-scaled normalized counts). Light blue highlights along genome coordinates denote the actual sub-sequence selected for detailed secondary structure prediction. Predicted structures are colored according to base pair probabilities depicted in nearby graphical scales. All structures were predicted using RNAfold web server (Gruber et al., 2008) using default parameters except energy parameters which were set to "Turner model, 1999".


(a)
1520000
1521000
VNG_RS07995
PF02080
PF02254
(b)
coverage
0
500
1000
1500
0
20
40
60
80
100
D

(c)

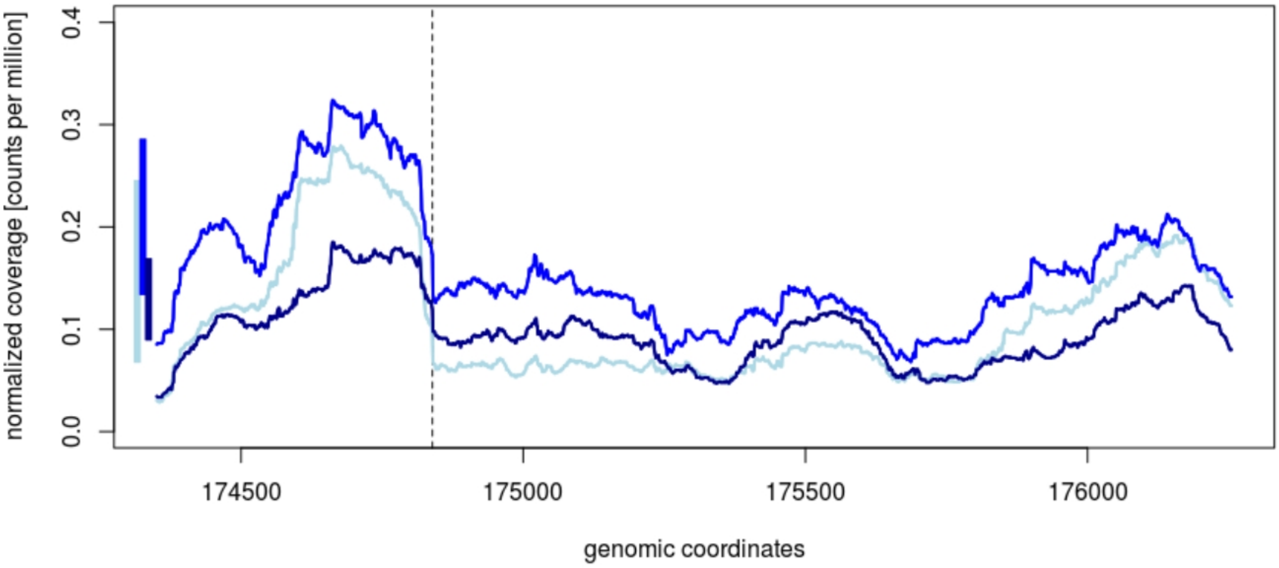
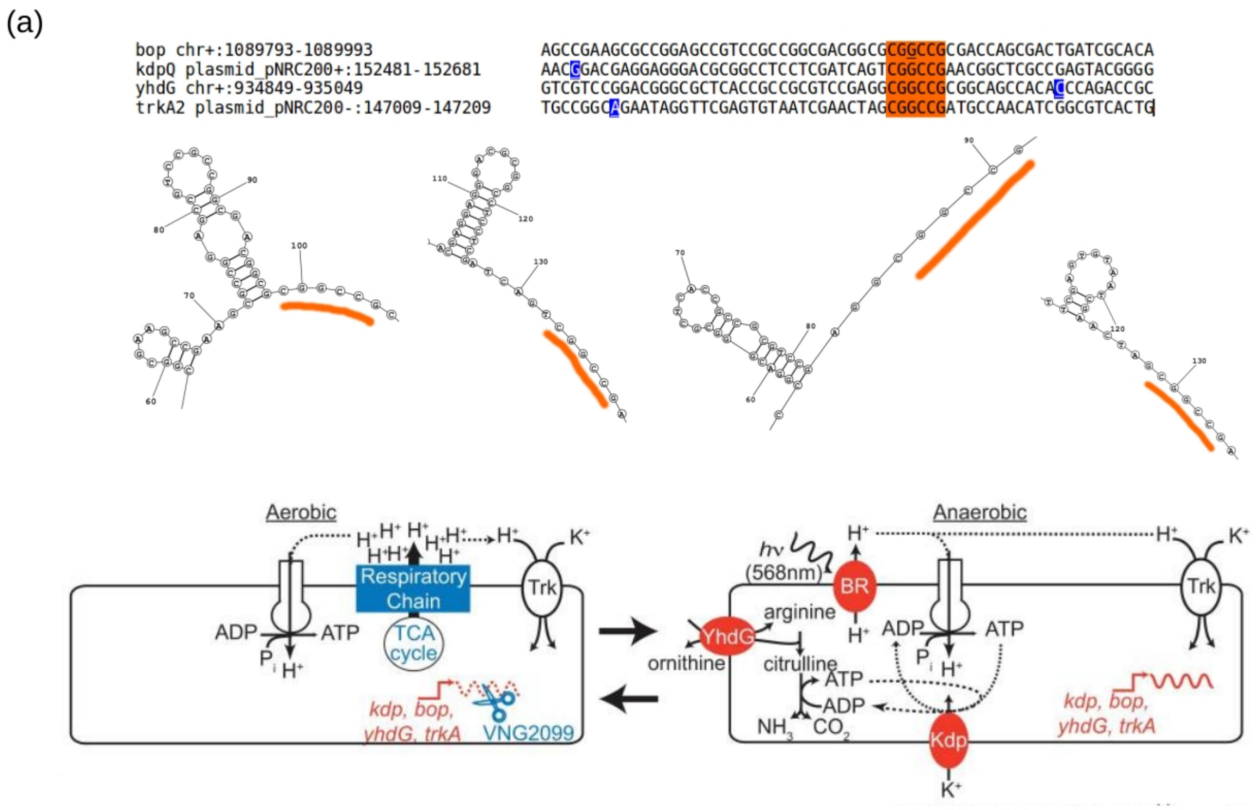
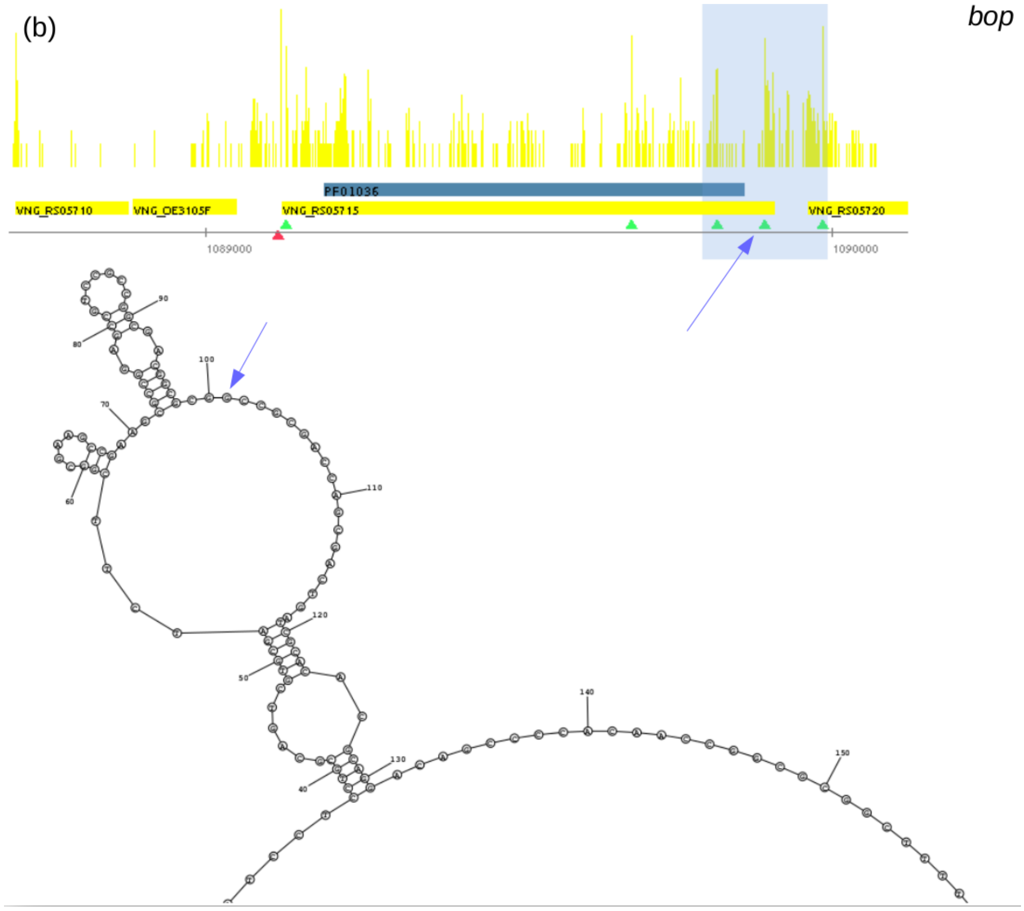


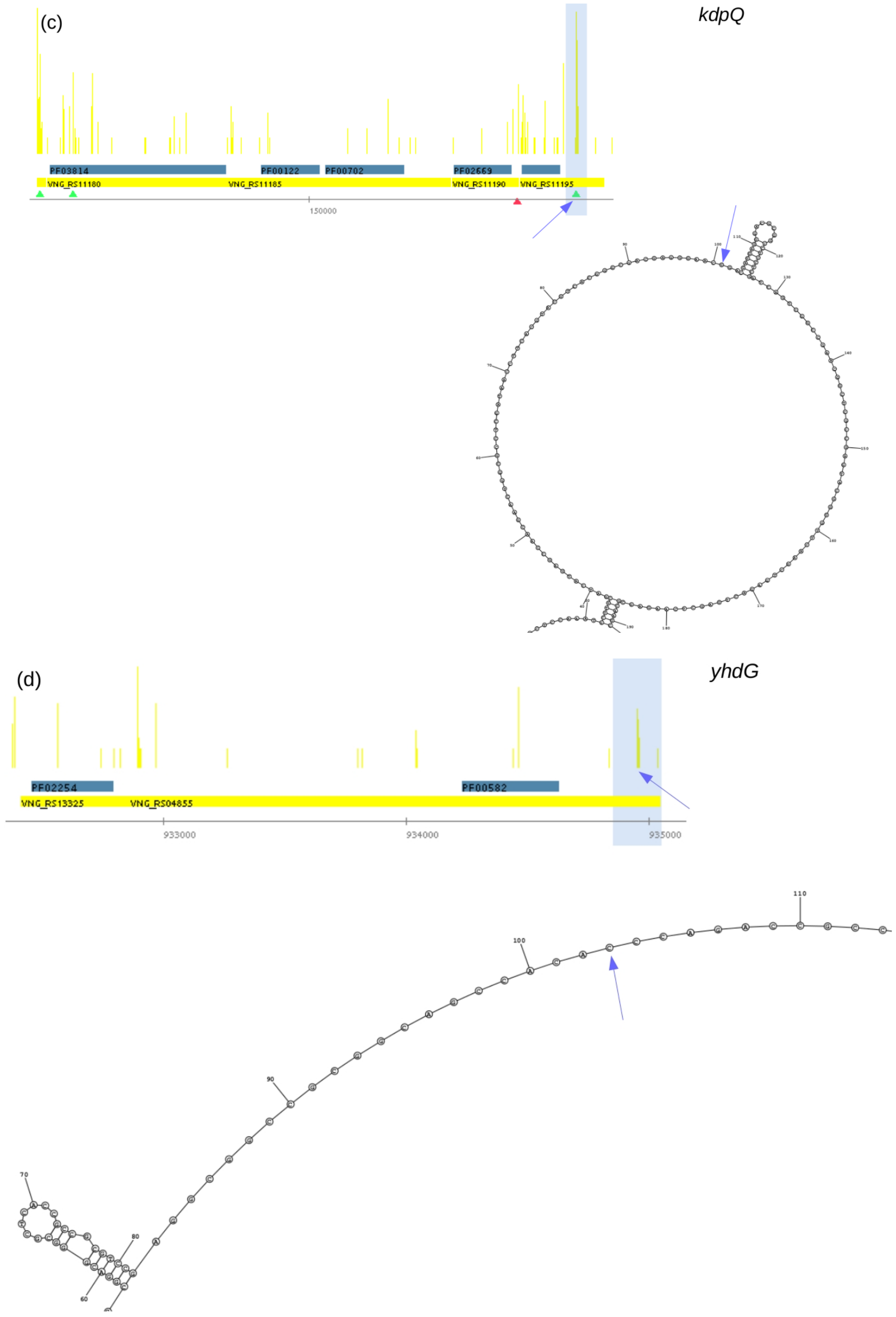
Figure S3 – Example of internal TPS in salt regulation gene.

(a) Example of transcript processing site (TPS) signal in *H. salinarum* dRNA-seq data. This is the sodium transporter *kef1* gene (VNG_RS07995 *locus*). Pfam domain annotation (blue rectangle) and coding sequences are in reverse strand (orange rectangle), thus 5'→3' direction is right to left. Distribution of relative numbers of aligned reads starting at a given genomic coordinate in TEX- libraries is shown in orange (arbitrarily log-scaled normalized counts). Aligned reads coverage along genomic coordinates for TEX+ and TEX- are shown in dark green and light green, respectively (absolute counts normalized and arbitrarily scaled). Magenta arrow points to the identified site (TPS_13995_1) along with typical TEX+ depletion signature. All possible archaeal start/stop codons in frame with *kef1* are shown in the bottom as vertical tick marks (ATG highlighted and stop codons in red). **(b)** Regular RNA-seq coverage profile of kef1 gene in *H. walsbyi*. Coverage profile of RNA-seq alignments (orange) are shown using length-normalized coordinates (D = 0 at start codon, D = 100 at stop codon inside HQ_RS10745 *locus*). Transcript abundance peak highlighted at D = 71 is positionally equivalent to TPS found in *H. salinarum*. Light and dark blue marks delimit the same Pfam domains as in (a). **(c)** Putative *Natrinema* sp. J7-2 TPS inside the sodium transporter *kef1* CDS. (NJ7G_RS00730 *locus*, reverse strand, 5'→3' is right to left). Normalized read alignment coverage is shown for low (15% NaCl, light blue), optimal (25% NaCl, blue) and high (30% NaCl, dark blue) salt concentrations along genomic coordinates only within CDS boundaries. RNA-seq data indicates that *kef1* is differentially processed at this site depending on salt concentration. The putative processing site is marked by a vertical dotted line at relative position D = 74. The expression level difference between the TPS-associated plateau and the overall gene is greater for low salt concentration (light blue vertical rectangle) and smaller for high salt concentration (dark blue vertical rectangle) although expression levels are higher in the optimal concentration (blue solid line profile).



```
bop chr+:1089793-1089993                AGCCGAAGCGCCGGAGCCGTCCGCCGGCGACGGCGCGGCCGCGACCAGCGACTGATCGCACA
kdpQ plasmid_pNRC200+:152481-152681     AACGGACGAGGAGGGACGCGGCCTCCTCGATCAGTCGGCCGAACGGCTCGCCGAGTACGGGG
yhdG chr+:934849-935049                 GTCGTCCGGACGGGCGCTCACCGCCGCGTCCGAGGCGGCCGCGGCAGCCACACCCAGACCGC
trkA2 plasmid_pNRC200-:147009-147209    TGCCGGCAGAATAGGTTCGAGTGTAATCGAACTAGCGGCCGATGCCAACATCGGCGTCACTG
```




(c)
kdpQ
PF03814
PF00122
PF00702
PF02669
VNG_RS11180
VNG_RS11185
VNG_RS11190
VNG_RS11195
150000
(d)
yhdG
PF02254
PF00582
VNG_RS13325
VNG_RS04855
933000
934000
935000
70
80
60
90
100
110

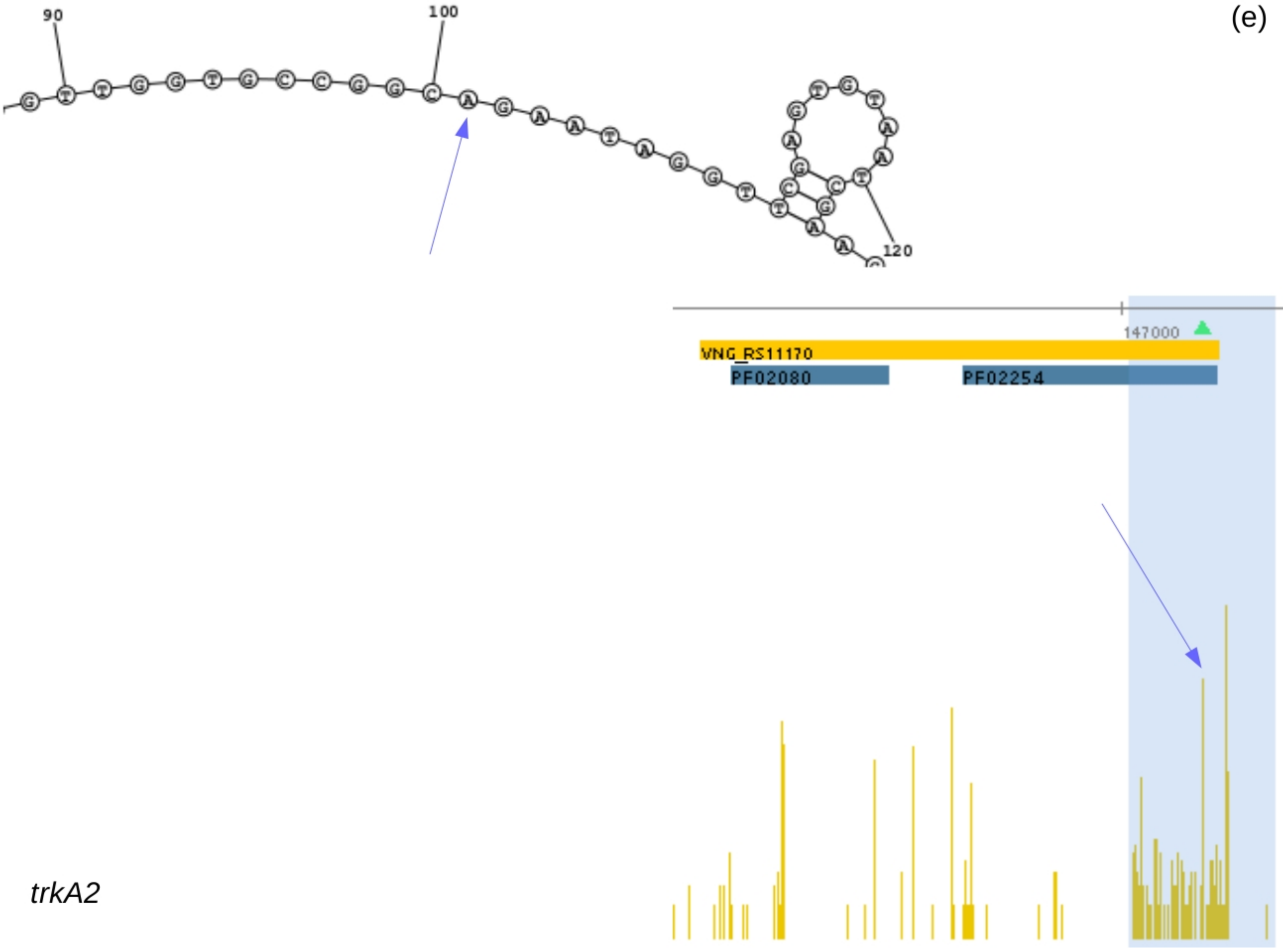
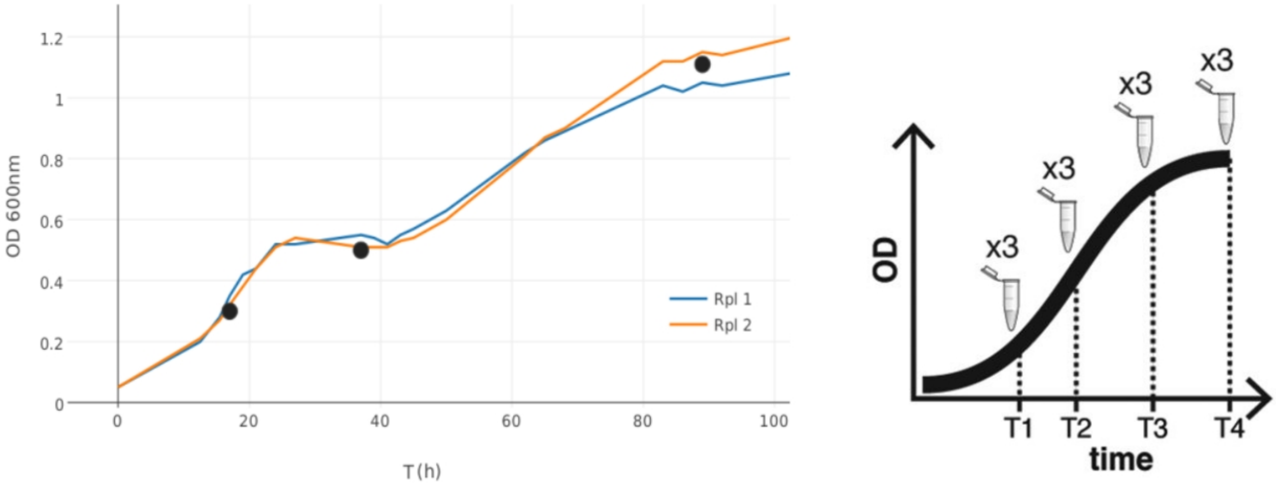


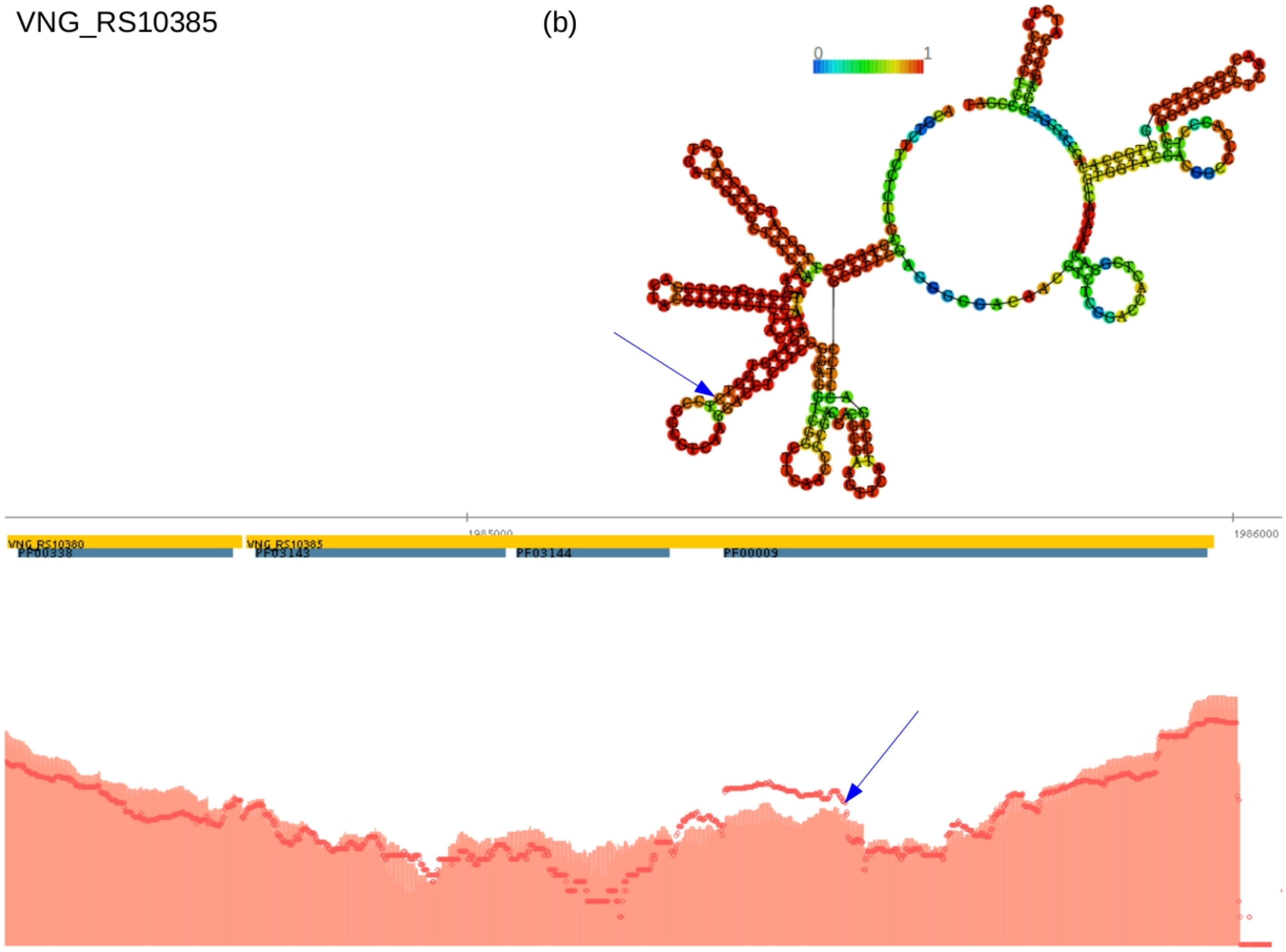
Figure S4 – Putative signature found in dis-regulated genes after VNG2099C RNase deletion.
(a) Sequences from -100 to +100 around TPS (underlined bases) inside genes *bop*, *kdpQ*, *yhdG* and *trkA2* were used as alignment input. All genes presented a `CGGCCG` sequence (orange highlights) downstream of a strong stem-loop. The secondary structure predictions were filtered to report only base pairings with >0.99 probability. RNase-mediated phenotypic switching scheme was adapted directly (from Wurtmann et al. 2014). Zoom out of all four predicted structures are shown in (b) to (e). **(b)** Overview of *bop* transcript secondary structure prediction result. Sequences that do not base pair with anything were cropped for clarity. Arrows point to TPS position. Light blue highlights the actual 100+1+100 nt sequence used for structure prediction. Pfam domain annotation (blue rectangle) and coding sequences (yellow rectangle) are in forward strand, thus 5′→3′ direction is left to right. Distribution of relative numbers of aligned reads starting at a given genomic coordinate in TEX- libraries is shown in yellow (arbitrarily log-scaled normalized counts). Green triangles are TPS defined in the present work (see Supplemental File 1 to navigate them). Red triangle marks ChIP-seq based binding site of TfbD transcription factor co-localized with genes' TSS (from Wilbanks *et al*., 2012). **(c)** Overview of *kdpQ* gene, same data description as in (b). **(d)** Overview of *yhdG*, same data description as in (b). **(e)** Overview of *trkA2*, same data description as in (b), except that 5′→3′ direction is right to left.

(a)



VNG_RS10385

(b)



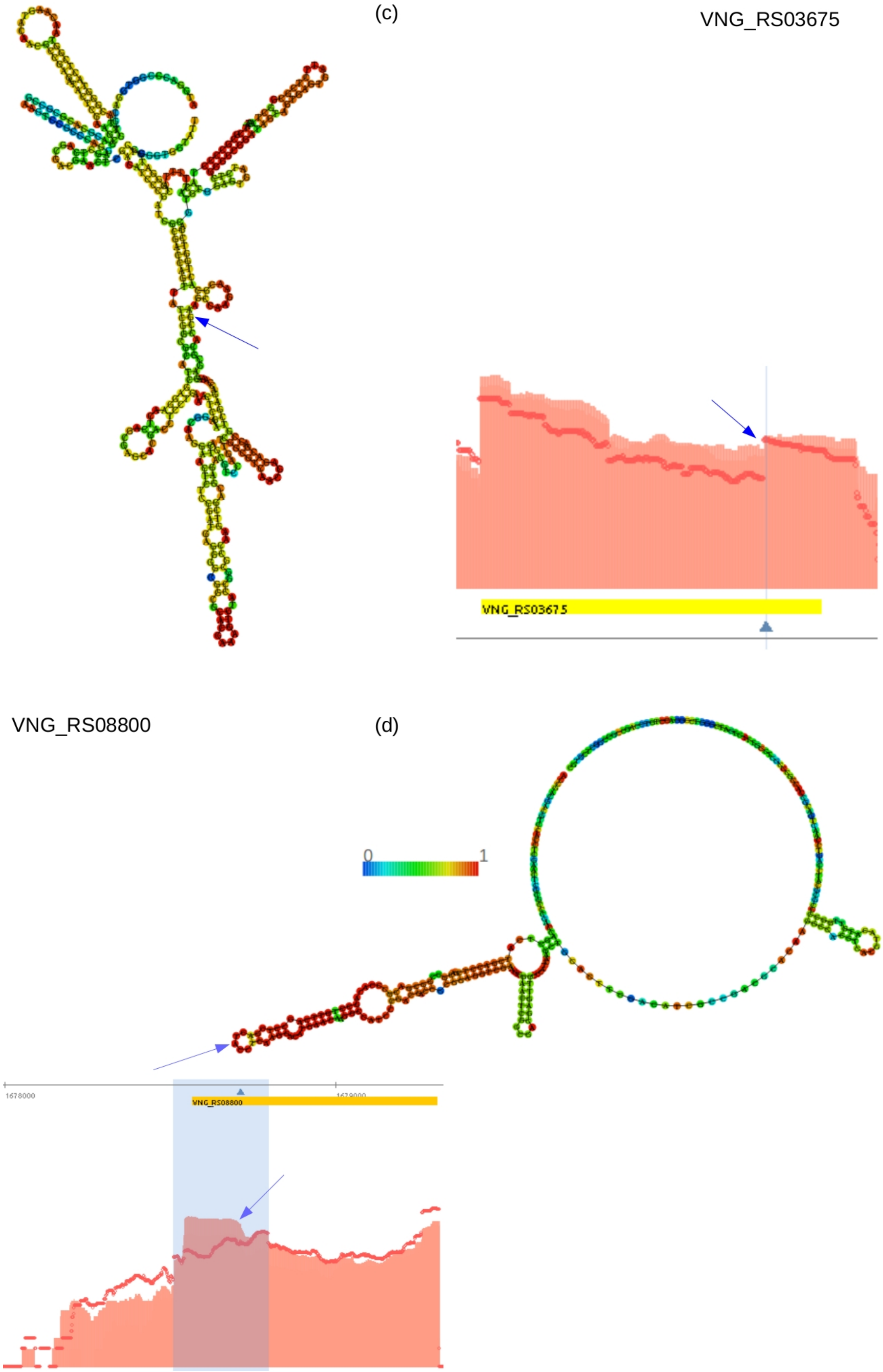


Figure S5 – Example of differential processing at TPS during *H. salinarum* growth.

(a) Growth curves from which original datasets were sampled. Left panel was addapted from Caten & Vêncio *et al*. (2018) Figure S1; dots show dRNA-seq data duplicate samples. Right panel was adapted from Lomana *et al*. (2020) Figure 7; lines show **(b)** *eEF1A*, encoding an elongation factor (TPS_16108_1, VNG_RS10385). Pfam domain annotation (blue rectangle) and coding sequences are in reverse strand (orange rectangle) thus 5′→3′ direction is right to left. Aligned reads coverage along genomic coordinates for TEX+ libraries at exponential and stationary phases are shown in light red (solid) and red (dots), respectively ($\log_2$ counts normalized and arbitrarily jointly scaled). Blue arrows point to conserved TPS. **(c)** a putative arsenic resistance operon repressor encoded at the VNG_RS03675 *locus* (TPS_2832_1). Coding sequence in forward strand (yellow rectangle) thus 5′→3′ is left to right. Transcriptome signal same as (b). **(d)** *pcn*, encoding a DNA polymerase III subunit (TPS_14733_1, VNG_RS08800). Coding sequence is in reverse strand (orange rectangle) thus 5′→3′ direction is right to left. Light blue highlight delimits the sub-sequence used for secondary structure prediction. Transcriptome signal same as (b).

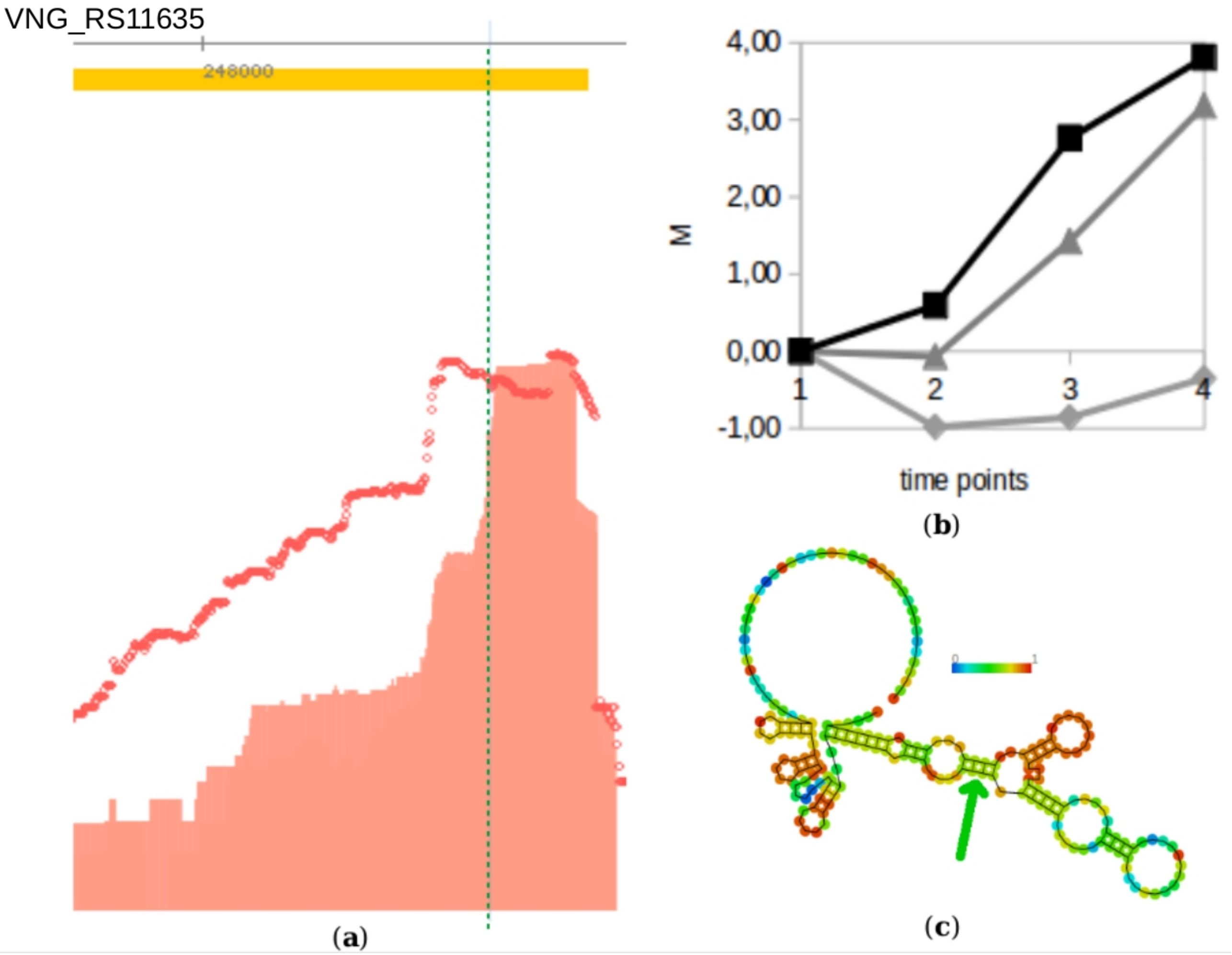


Figure S6 – Example of translation affecting probably due to TPS during *H. salinarum* growth.
(a) The TPS (TPS_20943_1, vertical dashed line) is near the start codon inside *arcA* gene (zoomed in VNG_RS11635 *locus*) which encodes an arginine deiminase pathway gene. Coding sequence is in reverse strand (orange rectangle) thus 5′→3′ direction is right to left and only first 600 bp are zoomed in out of ~1.5 kbp gene. Aligned reads coverage along genomic coordinates for TEX- libraries at exponential and stationary phases are shown in light red (solid) and red (dots), respectively ($\log_2$ counts normalized and arbitrarily jointly scaled). **(b)** *arcA* gene $\log_2$ fold-change (M) between published multi-modality measurements in different time-points relative to the early exponential phase from Lomana *et al.* (2020) and Lorenzetti *et al.* (2023) (Figure S5a, time-point 1: early exponential, 2: mid-exponential, 3: late exponential, 4: stationary, squares: RNA-seq data, triangles: Ribo-seq data). **(c)** Secondary structure prediction, color coded by pairing probabilities, using 100 bp around TPS (green arrow) as input.

(d)

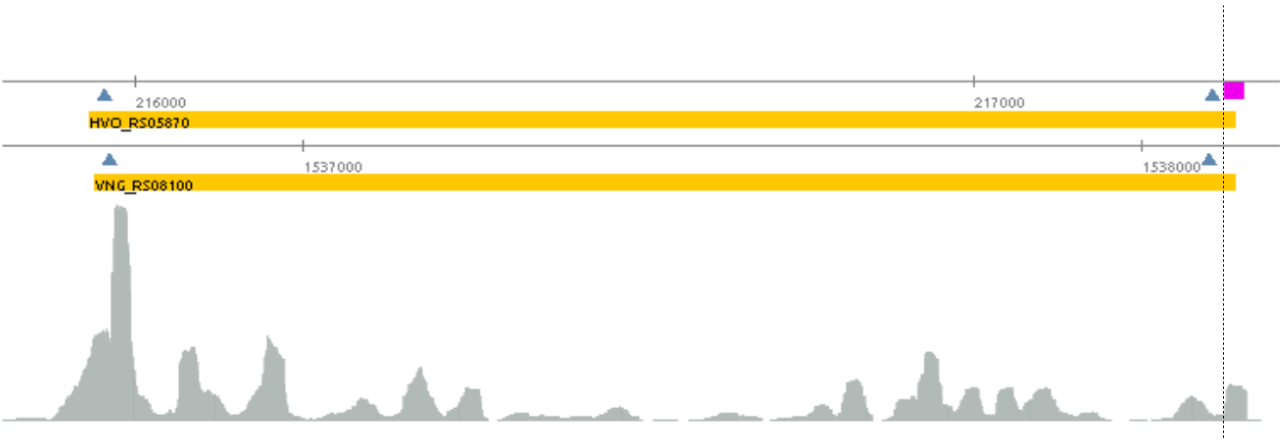


Figure S7 – Example of conserved TPS, Ribo-seq and rancRNA signal coincidence in *H. salinarum* and *H. volcanii*. Aligned reads coverage along genomic coordinates for ribosome footprint libraries (SRP119792, Lomana *et al*., 2020) are shown in gray (normalized counts arbitrarily scaled). Coding sequences are in reverse strand (5′→3′ direction is right to left) or forward strand (5′→3′ direction is left to right) if gene rectangles are orange or yellow, respectively. Pfam domain annotations are shown as blue rectangles with ID inside. Dark blue triangles point to conserved TPS. Magenta rectangles delimit published putative rancRNA *loci* in *H. volcanii* (Wyss *et al*., 2018). **(a)** *pan1* gene (VNG_RS01995 and HVO_RS08770 *loci*) which encodes PAN-A proteasome-activating nucleotidase. Vertical light blue highlight shows the difference between *pan1* known alternative transcripts (Chamieh *et al*., 2008). **(b)** VNG_RS04015 and HVO_RS20130 which encodes the putative archaeal translation factor aMBF1. **(c)** VNG_RS04165 and HVO_RS12050, which encodes an archaeosortase, a system-associated glycotransferase. **(d)** VNG_RS08100 and HVO_RS05870, which encodes a glutamine synthetase. **(e)** VNG_RS09610 and HVO_RS05290, which encodes the 54 kDa protein of the signal recognition particle ribonucleoprotein complex. **(f)** VNG_RS00020. **(g)** VNG_RS04995. Gene *sdo1*, which encodes for a ribosome maturation protein.

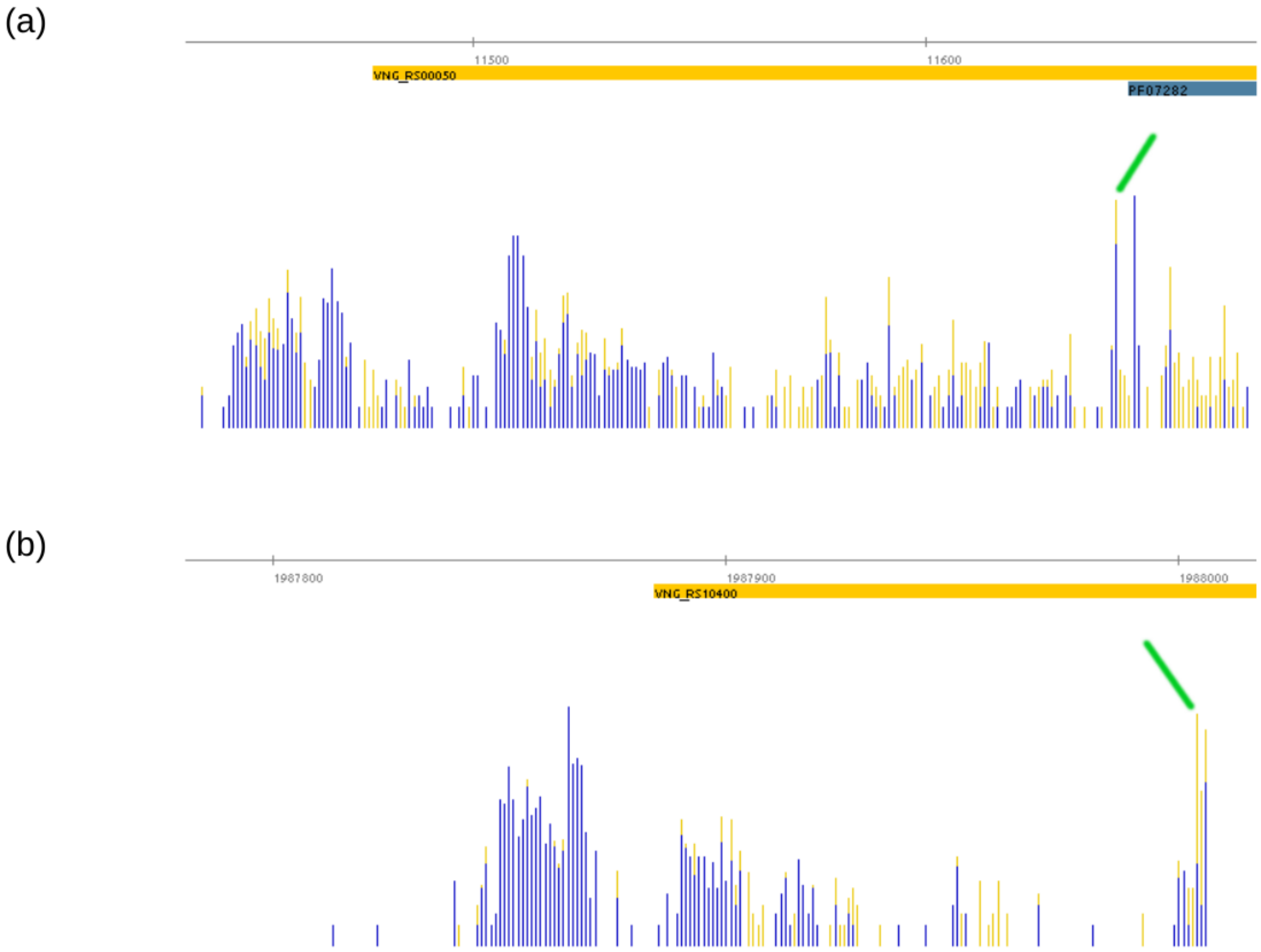


Figure S8 – Identification of processing site in sense overlapping transcripts VNG_sot0013 and VNG_sot2652. Distribution of aligned reads starting at a given genomic coordinate (horizontal ruler) for TEX+ (blue) and TEX- (yellow) datasets. Vertical bars in blue and yellow are superimposed signals (arbitrarily scaled $\log_2$ of normalized counts) at the same position. Zoom in of coding sequences are in reverse strand (orange rectangle, *locus* ID inside, 5′→3′ direction is right to left). Domain annotation (blue rectangle, PFAM ID inside). Green markers point to the TPS signatures which were not detected by the statistical significance finding methodology. **(a)** VNG_sot0013 and **(b)** VNG_sot2652 sotRNAs.

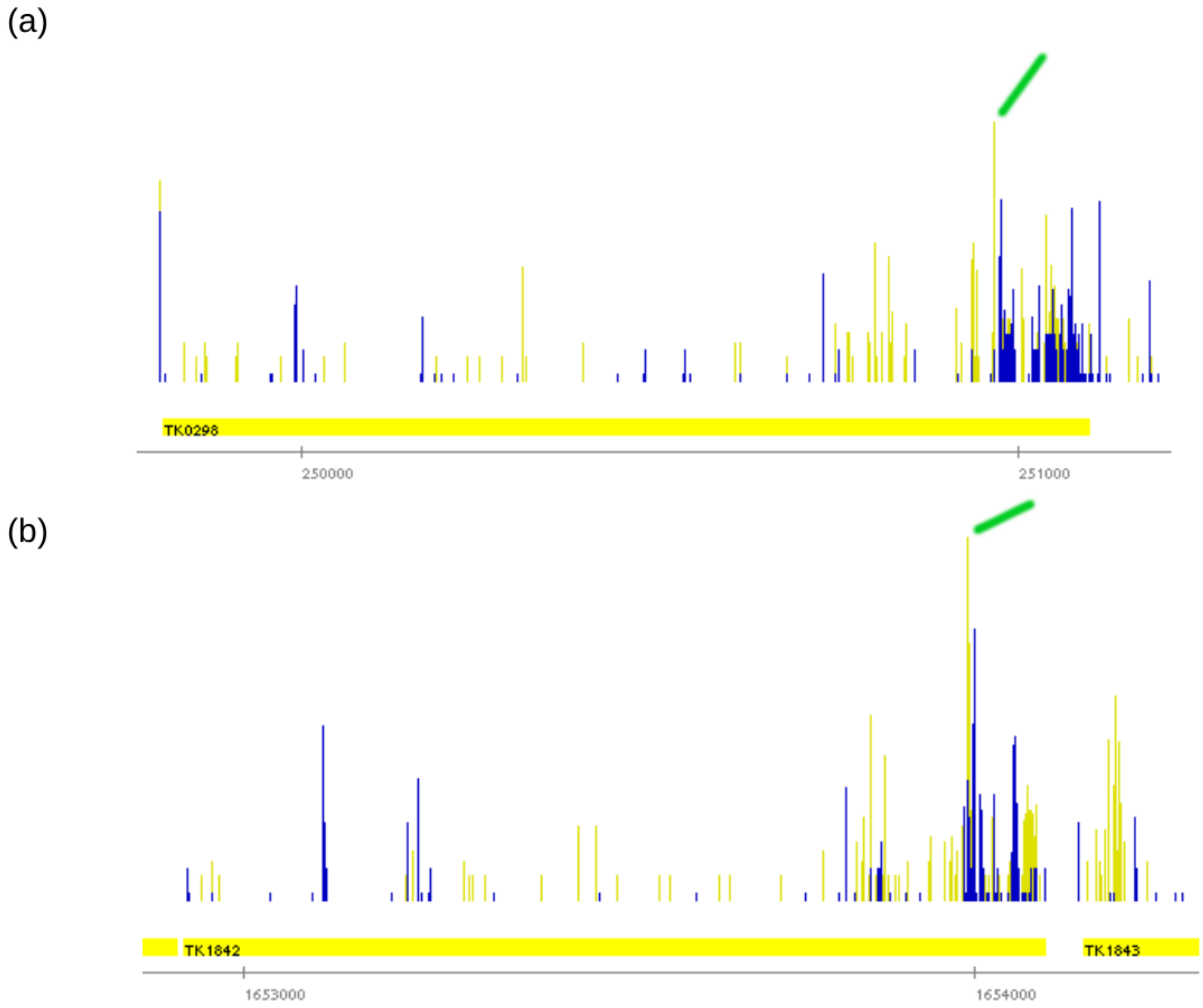
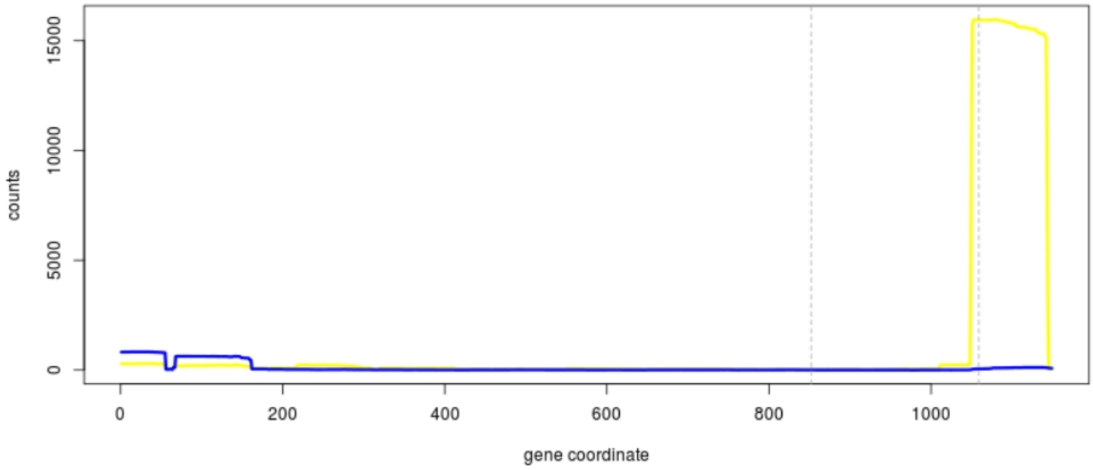
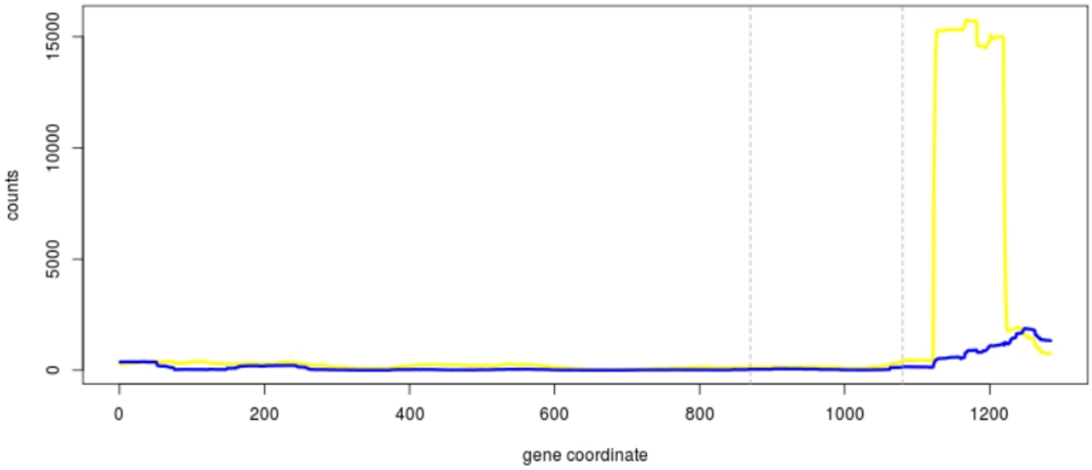


Figure S9 – Identification of processing site in sense overlapping transcripts in *T. kodakaraensis* IS605 insertion sequence family. Distribution of aligned reads starting at a given genomic coordinate (horizontal ruler) for TEX+ (blue) and TEX- (yellow) datasets. Vertical bars in blue and yellow are superimposed signals (arbitrarily scaled $\log_2$ of normalized counts) at the same position. Coding sequence is in forward strand (yellow rectangle, locus ID inside, 5′→3′ direction is left to right). Green markers point to putative TPS inside *loci* **(a)** TK0298 and **(b)** TK1842.

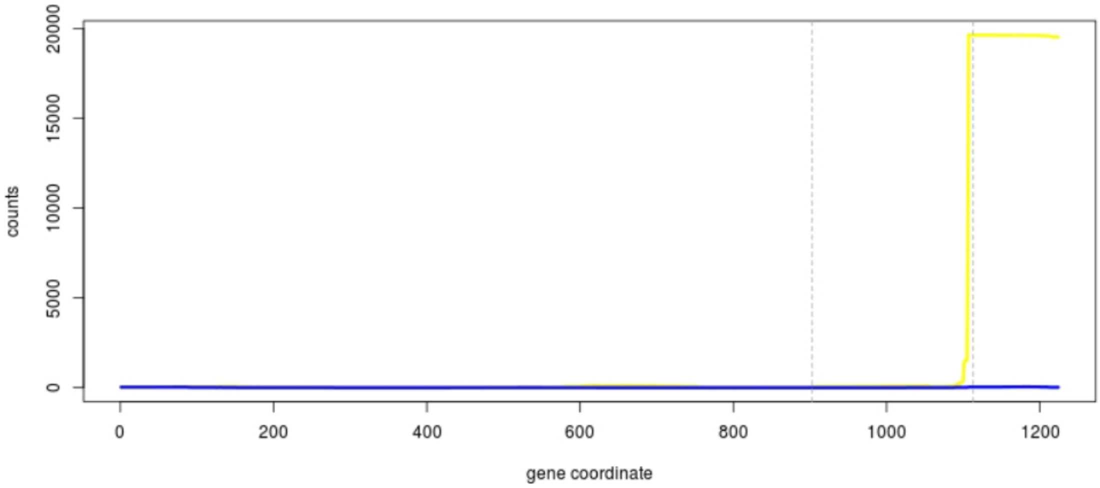
(a)


counts
0
5000
10000
15000
0
200
400
600
800
1000
gene coordinate

(b)


counts
0
5000
10000
15000
0
200
400
600
800
1000
1200
gene coordinate

(c)


counts
0
5000
10000
15000
20000
0
200
400
600
800
1000
1200
gene coordinate

(e)

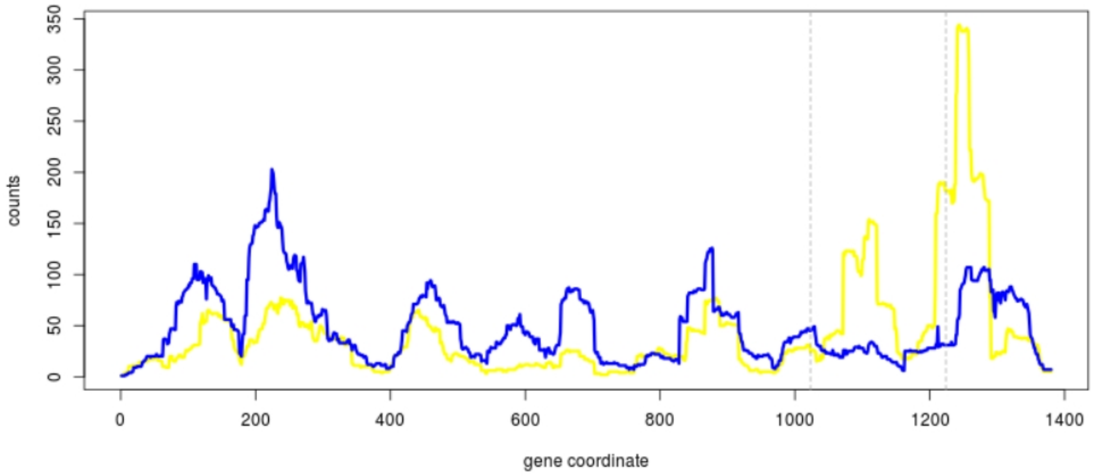


Figure S10 – Identification of processing site in IS associated sense overlapping transcripts in Bacteria. Aligned reads coverage signal along *tnpB* gene coordinates for TEX+ (blue) and TEX- (yellow) datasets in:
(a) *Escherichia coli* K-12 (*locus* b1432), **(b)** *Helicobacter pylori* 26695 (*locus* HP0989), **(c)** *Streptomyces coelicolor* M145 (*locus* SCO3714), **(d)** *Synechocystis* sp. PCC 6803 substr. GT-I (*locus* slr2062) and **(e)** *Mycobacterium tuberculosis* H37Rv (*locus* Rv2978c). Vertical dotted lines delimit the characteristic transposase DNA-binding protein domain OrfB_Zn_ribbon (PFAM database accession: PF07282).

Conclusion

5. Conclusion

Taking into account all the results obtained during this PhD research project and critically reflecting on the discussions that accompanied the presentation of these aforementioned results in the previous sections, we derive the following conclusions:

1) We established a reliable and reproducible pipeline to process and analyses dRNA-seq data using *H. salinarum* NRC-1 as our main focus but also tested its generality in other 5 other organisms. The modifications implemented in previously existing methods were not sufficient to justify an effort to disseminate the pipeline for a wider audience, but certainly equipped our group with the capacity to extract non-trivial information from dRNA-seq data: processing and secondary structure detection;

2) We were able to adapt bioinformatics protocols to extract RNA processing information from an experiment originally designed to only discover start sites;

3) All proposed specific objectives related to the discovery of new biology of *H. salinarum* NRC-1 were achieved: we mapped processing sites (TPS), we detected alternative start sites (alTSS) along the standard growth curve, and we identified structure-based termination sites (TTS);

4) The alTSS and TTS results are original and will be prepared for publication including comparisons between *H. salinarum* and *H. volcanii* that did not make into the PhD thesis;

5) The TPS results showed many processing sites and the results were published in a respectful peer-reviewed journal. Although our approach could not distinguish between the biological process that generated the processing sites (cleavage vs degradation, etc), we could raise mechanistic hypothesis on some specific cases, such an RNase studied by other groups. Also, the produced map was already used and cited by other researchers: (i) to create a post-transcriptional regulation atlas helping to explain differences they observed between RNA and protein levels; and (ii) by our own group in salt stress. We consider this article our main scientific contribution to the field.

Appendix

6- Appendix

1- A comparison between the results of several tools using a high-quality manually curated dataset of different organisms. In Venn diagrams.

2- Table of 91 genes that have alTSS in *Halobacterium salinarum* NRC-1 considering all time points from the growth curve into a single collapsed dataset.

3- Table of 292 genes that have alTSS in *Halobacterium salinarum* NRC-1 considering time points from the growth curve separately: 17h, 37h, 86h, reference condition replicate 1 (REF2014) and replicate 2 (REF2015).

4- Table of 30 genes that have alTSS in in *Caulobacter crescentus* NA1000.

5- Table of 934 genes that have alTSS in in *Haloferax volcanii* DS2.

6- Table of 250 genes with alTSS in *Methanocaldococcus jannaschii* DSM2661.

7- Table of 121 genes with alTSS in *Thermococcus onnurineus* NA1.

8- Table of 238 genes with alTSS in *Thermus thermophilus* HB8.

9- GGB files to allow proper visualization of different organisms in this study.

10- List of 96 genes with alTSS class internal in *Halobacterium salinarum* NRC-1.

11- List of putative TTS with MFE < -20 kcal/mol and distance between fTSS and coding sequence 3’ border less than 50 nucleotides.

12- Rainbow diagram representation of the predicted TTS in *Halobacterium salinarum* NRC-1.

13- Directory of the HTML files of Gene enrichment analysis for *Halobacterium salinarum* NRC-1 genes with alTSSs, including the GO Biological process complete, Molecular function complete, and Cellular component complete for both datasets of Bulk genes and genes in different Time points.

References

7. References

Aalberts, D. P., & Jannen, W. K. (2013). Visualizing RNA base-pairing probabilities with RNAbow diagrams. RNA (New York, N.Y.), 19(4), 475–478. https://doi.org/10.1261/rna.033365.112.

Allers, T., & Mevarech, M. (2005). Archaeal genetics - the third way. Nature Reviews. Genetics, 6(January), 58–73. https://doi.org/10.1038/nrg1504.

Amman, F., D'Halluin, A., Antoine, R., Huot, L., Bibova, I., Keidel, K., Slupek, S., Bouquet, P., Coutte, L., Caboche, S., Locht, C., Vecerek, B., & Hot, D. (2018). Primary transcriptome analysis reveals importance of IS elements for the shaping of the transcriptional landscape of *Bordetella pertussis*. RNA Biology, 15(7), 967–975. https://doi.org/10.1080/15476286.2018.1462655.

Amman, F., Wolfinger, M. T., Lorenz, R., Hofacker, I. L., Stadler, P. F., & Findeiß, S. (2014). TSSAR: TSS annotation regime for dRNA-seq data. BMC Bioinformatics, 15(1), 89. https://doi.org/10.1186/1471-2105-15-89.

Araujo, P. R., Yoon, K., Ko, D., Smith, A. D., Qiao, M., Suresh, U., Burns, S. C., & Penalva, L. O. F. (2012). Before it gets started: Regulating translation at the 5'UTR. In Comparative and Functional Genomics (Vol. 2012). https://doi.org/10.1155/2012/475731.

Arraiano, C. M., Andrade, J. M., Domingues, S., Guinote, I. B., Malecki, M., Matos, R. G., Moreira, R. N., Pobre, V., Reis, F. P., Saramago, M., Silva, I. J., & Viegas, S. C. (2010). The critical role of RNA processing and degradation in the control of gene expression. FEMS Microbiology Reviews, 34(5), 883–923. https://doi.org/10.1111/j.1574-6976.2010.00242.x.

Bare, J. C., Koide, T., Reiss, D. J., Tenenbaum, D., & Baliga, N. S. (2010). Integration and visualization of systems biology data in context of the genome. BMC Bioinformatics, 11(1), 382. https://doi.org/10.1186/1471-2105-11-382.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Brooks, A. N., Reiss, D. J., Allard, A., Wu, W. J., Salvanha, D. M., Plaisier, C. L., ... & Baliga, N. S. (2014). A system-level model for the microbial regulatory genome. Molecular systems biology, 10(7), 740. https://doi.org/10.15252/msb.20145160.

Calvo, S. E., Pagliarini, D. J., & Mootha, V. K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proceedings of the National Academy of Sciences of the United States of America, 106(18), 7507–7512. https://doi.org/10.1073/pnas.0810916106.

Castelle, C. J., & Banfield, J. F. (2018). Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell, 172(6), 1181–1197. https://doi.org/10.1016/J.cell.2018.02.016.

Chan P. P., Andrew D. Holmes, Andrew M. Smith, D. T. and T. M. L. (2012). The UCSC Archaeal Genome Browser: 2012 update. Nucleic Acids Research, 40(November 2011), D646–D652. https://doi.org/10.1093/nar/gkr990.

Chung, B. Y., Simons, C., Firth, A. E., Brown, C. M., & Hellens, R. P. (2006). Effect of 5'UTR introns on gene expression in Arabidopsis thaliana. BMC Genomics, 7, 120. https://doi.org/1471-2164-7-120 [pii]\r10.1186/1471-2164-7-120.

Cohen, O., Doron, S., Wurtzel, O., Dar, D., Edelheit, S., Karunker, I., Mick, E., & Sorek, R. (2016). Comparative transcriptomics across the prokaryotic tree of life. Nucleic Acids Research, 44(W1), W46–W53. https://doi.org/10.1093/nar/gkw394.

Coker, J. A., DasSarma, P., Capes, M., Wallace, T., McGarrity, K., Gessler, R., Liu, J., Xiang, H., Tatusov, R., Berquist, B. R., & DasSarma, S. (2009). Multiple replication origins of *Halobacterium* sp. strain NRC-1: Properties of the conserved orc7-dependent oriC1. Journal of Bacteriology, 191(16), 5253–5261. https://doi.org/10.1128/JB.00210-09.

DasSarma, P., & Dassarma, S. (2008). On the origin of prokaryotic "species": the taxonomy of halophilic Archaea Historical background. 5, 1–5. https://doi.org/10.1186/1746-1448-4-5.

DasSarma, P., Negi, V. D., Balakrishnan, A., Karan, R., Barnes, S., Ekulona, F., Chakravortty, D., & DasSarma, S. (2014). Haloarchaeal gas vesicle nanoparticles displaying Salmonella SopB antigen reduce bacterial burden when administered with live attenuated bacteria. Vaccine, 32(35), 4543–4549. https://doi.org/10.1016/J.vaccine.2014.06.021.

DasSarma, S., Berquist, B. R., Coker, J. A., DasSarma, P., & Müller, J. A. (2006). Post-genomics of the model haloarchaeon *Halobacterium* sp. NRC-1. Saline Systems, 2(1), 3. https://doi.org/10.1186/1746-1448-2-3.

DasSarma, S. L., Capes, M. D., DasSarma, P., & DasSarma, S. (2010). HaloWeb: The haloarchaeal genomes database. *Saline Systems*, 6(1), 12. https://doi.org/10.1186/1746-1448-6-12.

de Klerk, E., & 't Hoen, P. A. C. (2015). Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing. In Trends in Genetics (Vol. 31, Issue 3, pp. 128–139). https://doi.org/10.1016/j.tig.2015.01.001.

Down, T. A., & Hubbard, T. J. P. (2016). Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA Methods 458 Genome Research. https://doi.org/10.1101/gr.216102.

Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., & Sharma, C. M. (2013). High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple Campylobacter jejuni Isolates. PLoS Genetics, 9(5), e1003495. https://doi.org/10.1371/journal.pgen.1003495.

Eme, L., Spang, A., Lombard, J., Stairs, C. W., & Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. Nature Reviews Microbiology, 15(12), 711–723. https://doi.org/10.1038/nrmicro.2017.133.

Ettwiller, L., Buswell, J., Yigit, E., & Schildkraut, I. (2016). A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. BMC Genomics, 17(1), 199. https://doi.org/10.1186/s12864-016-2539-z.

Furtwängler, K., Tarasov, V., Wende, A., Schwarz, C., & Oesterhelt, D. (2010). Regulation of phosphate uptake via Pst transporters in *Halobacterium salinarum* R1. Molecular Microbiology, 76(2), 378–392. https://doi.org/10.1111/j.1365-2958.2010.07101.x.

Gaba, S., Kumari, A., Medema, M., & Kaushik, R. (2020). Pan-genome analysis and ancestral state reconstruction of class halobacteria: probability of a new super-order. Scientific Reports 2020 10:1, 10(1), 1–16. https://doi.org/10.1038/s41598-020-77723-6.

Gehring, A. M., Walker, J. E., & Santangelo, T. J. (2016). Transcription regulation in archaea. Journal of Bacteriology, 198(14), 1906–1917. https://doi.org/10.1128/JB.00255-16.

Gomes-Filho, J. V., Zaramela, L. S., Italiani, V. C. da S., Baliga, N. S., Vêncio, R. Z. N., & Koide, T. (2015). Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea. RNA Biology, 12(5), 490–500. https://doi.org/10.1080/15476286.2015.1019998.

Gowda, M., Li, H., Alessi, J., Chen, F., Pratt, R., & Wang, G. L. (2006). Robust analysis of 5′-transcript ends (5′-RATE): A novel technique for transcriptome analysis and genome annotation. Nucleic Acids Research. https://doi.org/10.1093/nar/gkl522.

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zheng, C., Thibaud-Nissen, F., Geer, L. Y., … Pruitt, K. D. (2018). RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Research, 46(D1), D851–D860. https://doi.org/10.1093/nar/gkx1068.

Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S., & Matsushima, K. (2004). 5'-end SAGE for the analysis of transcriptional start sites. Nature Biotechnology, 22(9), 1146–1149. https://doi.org/10.1038/nbt998.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell. https://doi.org/10.1016/j.molcel.2010.05.004.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E., & Downing, K. H. (1990). Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. Journal of Molecular Biology, 213(4), 899–929. https://doi.org/10.1016/S0022-2836(05)80271-2.

Ibrahim, A. G. A. E. R., Vêncio, R. Z. N., Lorenzetti, A. P. R., & Koide, T. (2021). *Halobacterium salinarum* and *Haloferax volcanii* comparative transcriptomics reveals conserved transcriptional processing sites. Genes, 12(7), 1018. https://doi.org/10.3390/genes12071018.

Ingoldsby, L. M., Geoghegan, K. F., Hayden, B. M., & Engel, P. C. (2005). The discovery of four distinct glutamate dehydrogenase genes in a strain of *Halobacterium salinarum*. Gene, 349, 237–244. https://doi.org/10.1016/j.gene.2005.01.011.

Jorjani, H., & Zavolan, M. (2014). TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. Bioinformatics, 30(7), 971–974. https://doi.org/10.1093/bioinformatics/btt752.

Koide, T., Reiss, D. J., Bare, J. C., Pang, W. L., Facciotti, M. T., Schmid, A. K., Pan, M., Marzolf, B., Van, P. T., Lo, F. Y., Pratap, A., Deutsch, E. W., Peterson, A., Martin, D., & Baliga, N. S. (2009). Prevalence of transcription promoters within archaeal operons and coding sequences. Molecular Systems Biology, 5(1), 285. https://doi.org/10.1038/msb.2009.42.

Kramer, P., Gäbel, K., Pfeiffer, F., & Soppa, J. (2014). *Haloferax volcanii*, a prokaryotic species that does not use the shine dalgarno mechanism for translation initiation at 59-UTRs. PLoS ONE, 9(4). https://doi.org/10.1371/journal.pone.0094979

Kumar, M., Srinivas, V., & Patankar, S. (2015). Upstream AUGs and upstream ORFs can regulate the downstream ORF in Plasmodium falciparum. Malaria Journal, 14, 512. https://doi.org/10.1186/s12936-015-1040-5.

Li, J., Qi, L., Guo, Y., Yue, L., Li, Y., Ge, W., Wu, J., Shi, W., & Dong, X. (2015). Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon *Methanolobus psychrophilus*. Scientific Reports, 5, 9209. https://doi.org/10.1038/srep09209.

Linder, P., & Jankowsky, E. (2011). From unwinding to clamping ĝ€" the DEAD box RNA helicase family. In Nature Reviews Molecular Cell Biology (Vol. 12, Issue 8, pp. 505–516). Nature Publishing Group. https://doi.org/10.1038/nrm3154.

Lorenzetti, A. P. R. (2021). Systems investigation of SmAP1-mediated regulation of transposase-encoding transcripts in *Halobacterium salinarum* NRC-1 [Universidade de São Paulo]. https://doi.org/https://doi.org/10.11606/T.95.2021.tde-15092021-103306.

Lorenzetti, A. P. R., Kusebauch, U., Zaramela, L. S., Wu, W.-J., Almeida, J. P. P. de, Turkarslan, S., Lomana, A. L. G. de, Gomes-Filho, J. V., Vêncio, R. Z. N., Moritz, R. L., Koide, T., & Baliga, N. S. (2023). A Genome-Scale Atlas Reveals Complex Interplay of Transcription and Translation in an Archaeon. MSystems. https://doi.org/10.1128/MSYSTEMS.00816-22.

Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. Life (Basel, Switzerland), 5(1), 818–840. https://doi.org/10.3390/life5010818.

Metzker, M. L. (2010). Sequencing technologies — the next generation. Nature Reviews Genetics, 11(1), 31–46. https://doi.org/10.1038/nrg2626.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 5(7), 621–628. https://doi.org/10.1038/nmeth.1226.

Mortimer, S. A., Trapnell, C., Aviran, S., Pachter, L., & Lucks, J. B. (2012). SHAPE-Seq: High-Throughput RNA Structure Analysis. Current Protocols in Chemical Biology, 4(4), 275–297. https://doi.org/10.1002/9780470559277.CH120019.

Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., Lasky, S. R., Baliga, N. S., Thorsson, V., Sbrogna, J., Swartzell, S., Weir, D., Hall, J., Dahl, T. A., Welti, R., Goo, Y. A., Leithauser, B., Keller, K., Cruz, R., … DasSarma, S. (2000). Genome sequence of *Halobacterium* sp. NRC-1. Proceedings of the National Academy of Sciences of the United States of America, 97(22), 12176–12181. https://doi.org/10.1073/pnas.190337797.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189.

Onga, E. A., Vêncio, R. Z. N., & Koide, T. (2022). Low Salt Influences Archaellum-Based Motility, Glycerol Metabolism, and Gas Vesicles Biogenesis in *Halobacterium salinarum*. Microorganisms, 10(12), 2442. https://doi.org/10.3390/microorganisms10122442.

Perwez, T., & Kushner, S. R. (2006). RNase Z in Escherichia coli plays a significant role in mRNA decay. Molecular Microbiology, 60(3), 723–737. https://doi.org/10.1111/J.1365-2958.2006.05124.x.

Pfeifer, F. (2015). Haloarchaea and the formation of gas vesicles. In *Life* (Vol. 5, Issue 1, pp. 385–402). https://doi.org/10.3390/life5010385.

Pfeiffer, F., Broicher, A., Gillich, T., Klee, K., Mejía, J., Rampp, M., & Oesterhelt, D. (2008). Genome information management and integrated data analysis with HaloLex. Archives of Microbiology, 190(3), 281–299. https://doi.org/10.1007/s00203-008-0389-z.

Promworn, Y., Kaewprommal, P., Shaw, P. J., Intarapanich, A., Tongsima, S., & Piriyapongsa, J. (2017). ToNER: A tool for identifying nucleotide enrichment signals in feature-enriched RNA-seq data. *PLOS ONE*, *12*(5), e0178483. https://doi.org/10.1371/journal.pone.0178483.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 35(SUPPL. 1), 61–65. https://doi.org/10.1093/nar/gkl842.

Pun, P. P., Schuldt, S., & Pun, B. T. (2005). The Three Domains of Life: A Challenge to the concept of the Universal Cellular Ancestor?. Progress in Complexity, Information and Design, 4.

Reyes, A., & Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1165.

Sampford, M. R. (1955). The Truncated Negative Binomial Distribution. *Biometrika*, *42*(1–2), 58–69. https://doi.org/10.1093/biomet/42.1-2.58.

Santangelo, T. J., Cubonová, L., Skinner, K. M., & Reeve, J. N. (2009). Archaeal Intrinsic Transcription Termination In Vivo. Journal of Bacteriology, 191(22), 7102. https://doi.org/10.1128/JB.00982-09.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., & Vogel, J. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, *464*(7286), 250–255. https://doi.org/10.1038/nature08756.

Sharma, C. M., & Vogel, J. (2014). Differential RNA-seq: The approach behind and the biological insight gained. Current Opinion in Microbiology, 19(1), 97–105. https://doi.org/10.1016/j.mib.2014.06.010.

Somers, J., Pöyry, T., & Willis, A. E. (2013). A perspective on mammalian upstream open reading frame function. The International Journal of Biochemistry & Cell Biology, 45(8), 1690–1700. https://doi.org/10.1016/j.biocel.2013.04.020.

Takahashi, H., Kato, S., Murata, M., & Carninci, P. (2012). CAGE (Cap analysis of gene expression): A protocol for the detection of promoter and transcriptional networks. Methods in Molecular Biology. https://doi.org/10.1007/978-1-61779-292-2_11.

Ten-Caten, F. (2017). A transcrição pervasiva na archaea *Halobacterium salinarum* NRC-1 e a identificação de novos transcritos. [Universidade de São Paulo]. https://doi.org/10.11606/T.95.2019.tde-17042019-143232.

Ten-Caten, F., Vêncio, R. Z. N., Lorenzetti, A. P. R., Zaramela, L. S., Santana, A. C., & Koide, T. (2018). Internal RNAs overlapping coding sequences can drive the production of alternative proteins in archaea. *RNA Biology*, *15*(8), 1119–1132. https://doi.org/10.1080/15476286.2018.1509661.

Thomm, M., Hausner, W., & Hethke, C. (1993). Transcription Factors and Termination of Transcription in Methanococcus. Systematic and Applied Microbiology, 16(4), 648–655. https://doi.org/10.1016/s0723-2020(11)80336-x.

Whitehead, K., Pan, M., Masumura, K. I., Bonneau, R., & Baliga, N. S. (2009). Diurnally entrained anticipatory behavior in archaea. PLoS ONE, 4(5). https://doi.org/10.1371/journal.pone.0005485.

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M., Pham, G. M., Nicotra, A. B., Gregorio, G. B., Jagadish, S. V. K., Septiningsih, E. M., Bonneau, R., & Purugganan, M. (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments. The Plant Cell. https://doi.org/10.1105/tpc.16.00158.

Williams, T. A., Foster, P. G., Cox, C. J., & Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. In Nature (Vol. 504, Issue 7479, pp. 231–236). https://doi.org/10.1038/nature12779.

Woese, C. R. (1994). There must be a prokaryote somewhere: microbiology's search for itself. Microbiological Reviews, 58(1), 1–9. https://doi.org/0146-0749/94/$04.00+0

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences of the United States of America, 87(12), 4576–4579. https://doi.org/10.1073/pnas.87.12.4576.

Won, S. Y., Kwon, S.-J., Lee, T.-H., Jung, J.-A., Kim, J. S., Kang, S.-H., & Sohn, S.-H. (2017). Comparative transcriptome analysis reveals whole-genome duplications and gene selection patterns in cultivated and wild Chrysanthemum species. Plant Molecular Biology, 95(4–5), 451–461. https://doi.org/10.1007/s11103-017-0663-z.

Yu, S. H., Vogel, J., & Förstner, K. U. (2018). ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. GigaScience, 7(9), 1–11. https://doi.org/10.1093/gigascience/giy096.

Zaramela, L. S., Vêncio, R. Z. N., Ten-Caten, F., Baliga, N. S., & Koide, T. (2014). Transcription start site associated RNAs (TSSaRNAs) are ubiquitous in all domains of life. *PLoS ONE*, *9*(9), e107680. https://doi.org/10.1371/journal.pone.0107680.

Zhang, P., Dimont, E., Ha, T., Swanson, D. J., Hide, W., & Goldowitz, D. (2017). Relatively frequent switching of transcription start sites during cerebellar development. BMC Genomics. https://doi.org/10.1186/s12864-017-3834-z.

Zhang, R., & Zhang, C. T. (2003). Multiple replication origins of the archaeon *Halobacterium* sp. NRC-1. Biochemical and Biophysical Research Communications, 302(4), 728–734. https://doi.org/10.1016/S0006-291X(03)00252-3.

Ziegel, E. R., & Ross, S. (1998). A First Course in Probability. Technometrics, 40(3), 268. https://doi.org/10.2307/1271207.