UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

JOSÉ VICENTE GOMES FILHO

RNAs não-codificantes associados a IS200/605: Identificação e caracterização funcional na archaea
Halobacterium salinarum NRC-1

Ribeirão Preto

2017

JOSÉ VICENTE GOMES FILHO

RNAs não-codificantes associados a IS200/605: Identificação e caracterização funcional na archaea
Halobacterium salinarum NRC-1

Tese apresentada à Faculdade de Medicina de
Ribeirão Preto da Universidade de São Paulo para
obtenção do título de Doutor em Ciências

Área de Concentração: Bioquímica

Orientadora: Prof^a Dr^a Tie Koide

Ribeirão Preto
2017

Gomes-Filho, J. V. **RNAs não-codificantes associados a IS200/605: Identificação e caracterização funcional na archaea *Halobacterium salinarum* NRC-1.** Tese apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Aprovado em:

Banca Examinadora

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

*Dedico este trabalho a meus pais José Vicente Gomes e Marlene Rizzolli Gomes,
a minha irmã Monica Rizzolli Gomes e a minha parceira Vivian Paula da Costa.*

AGRADECIMENTOS

À minha família por todo o suporte durante o desenvolvimento deste projeto. Ao meu pai José Vicente Gomes e a minha mãe Marlene Rizzolli Gomes por desde cedo me incentivarem a questionar o mundo. A minha irmã Monica Rizzolli Gomes pela companhia e discussões críticas.

À minha namorada Vivian Paula da Costa por estar ao meu lado por mais de uma década, me incentivando e proporcionando momentos inesquecíveis.

À Prof^a. Tie pela oportunidade de integrar seu grupo e pelo incentivo dado durante estes anos. Agradeço pelos conselhos, críticas e por demonstrar que a ciência deve ser feita de maneira criteriosa. Essa convivência foi fundamental para minha formação profissional.

Ao Prof. Ricardo Vêncio pelos questionamentos e por sua empolgação em relação a ciência.

À todos os companheiros de laboratório, Diego Martinez Salvanha, Livia Soares Zaramela, Felipe ten Caten, Marjorie Cornejo Pontelli, André Medina, Andrea Aguiar, Victoria Maruyama, Rafael Silva Rocha, Gilvan Furtado, Alan Lorenzetti, João Paulo, Jadson, Junier Marrero e Victor que proporcionaram grandes momentos de diversão e discussões científicas de altíssima qualidade.

Às alunas de iniciação científica Stella Kyomen e Polyana Garcia, que tive o prazer de co-orientar e foram fundamentais para minha formação como futuro orientador.

À Silvinha pela grande amizade e pelo imenso suporte técnico. Sua dedicação e empenho foram peças fundamentais na realização de meu trabalho.

À todos membros do laboratório da Prof^a. Angela K. Cruz por nos recepcionarem e cederem sua infraestrutura para o desenvolvimento de etapas fundamentais deste projeto.

À todos membros do laboratório do Prof. Dr. Roberto do Nascimento Silva pela boa convivência e por cederem sua infraestrutura.

À FAPESP pela bolsa de Doutorado e por todo apoio financeiro.

Ao CNPq, CAPES e FAEPA pelo apoio financeiro.

Aos meus amigos Fábio Rodrigues, Sidnei Ziotti, Rudá Andrade e Otávio Beraldi por proporcionarem momentos de extrema alegria e diversão em nossas viagens.

A todos que de maneira direta ou indireta contribuíram para a realização deste trabalho.

Muito obrigado!

Este trabalho foi realizado com o suporte financeiro das seguintes entidades:

Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP

Processo Nº 2012/13520-0

Processo Nº 2013/21522-5

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES

Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq

Fundação de Apoio ao Ensino, Pesquisa e Assistência do Hospital das Clínicas da FMRP – USP

RESUMO

Gomes-Filho, J. V. **RNAs não-codificantes associados a IS200/605: Identificação e caracterização funcional na archaea *Halobacterium salinarum* NRC-1.** 2017, 118 f. Tese (Doutorado) - Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo, Ribeirão Preto, 2017.

Os elementos genéticos móveis (*mobile genetic elements* MGEs), são elementos extremamente importantes para plasticidade e evolução dos genomas. Uma das classes mais importantes de MGEs são as sequências de inserção (*insertion sequences* - IS). Estes elementos são encontrados em bactérias e archaea e apresentam grande diversidade de famílias e mecanismos de movimentação. Uma família interessante de IS é IS200/605, esta família encontra-se distribuída em bactérias, archaea e vírus e utiliza substratos de DNA de fita simples para seu processo de transposição. Neste trabalho, através da análise de dados públicos de transcrito identificamos RNAs sobrepostos ao 3' de genes *tnpB* de IS200/605 em archaea e bactérias, estes transcritos foram chamados de *sense overlapping transcripts* (sotRNAs). As extremidades 5' e 3' dos sotRNAs foram mapeadas através de dados de RNA-seq de pequenos RNAs (sRNA-seq) e validadas através da técnica de C-RACE. Análises de sequência e estrutura secundária demonstraram que estes RNAs apresentam um motivo conservado chamado de RE-like. Utilizando a sequência consenso deste motivo pudemos identificar RNAs intergênicos derivados de IS200/605 em *H. salinarum* NRC-1. Para caracterização funcional, construímos linhagens superexpressando os RNAs VNG_sot0042 e VNG_R0052, ambos contendo o motivo RE-like. Curvas de crescimento, utilizando as linhagens construídas demonstraram que a superexpressão destes RNAs aumenta o crescimento de *H. salinarum* demonstrando sua funcionalidade. Devido a presença do motivo RE-like, extremidades 5' e 3' determinadas e fenótipo visualizado em curva de crescimento padrão, o sotRNA VNG_sot0042 foi estudado mais a fundo. Realizamos experimentos de RNA-seq para avaliar o impacto desta superexpressão no transcrito de *H. salinarum* NRC-1, assim como experimentos de SILAC para identificação de proteínas parceiras em larga escala. Nestes ensaios identificamos proteínas e genes associados ao processo de adesão, geração de células persistentes e resistência a metais pesados. Ensaios de adesão e sobrevivência a metais pesados demonstraram que a linhagem de superexpressão apresenta maior capacidade de aderir a vidro e maior sobrevivência em diversas condições de estresse. Deste modo, podemos sugerir que ncRNAs derivados de IS200/605 são importantes moléculas regulatórias em *H. salinarum* NRC-1 e nos ajudam a compreender a manutenção de IS200/605 e seus derivados nos genomas de procariotos.

Palavras-chave: Archaea, RNAs não-codificantes, Elementos móveis, RNA-seq, Sequências de inserção

ABSTRACT

Gomes-Filho, J. V. **non-coding RNAs associated with IS200/605: Identification and functional characterizal in the archaea *Halobacterium salinarum* NRC-1.** 2017, 118 f. Tese (Doutorado) - Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo, Ribeirão Preto, 2017.

Mobile genetic elements (MGEs) are extremely important for plasticity and evolution of genomes. Their impacts are diverse and could be related to antibiotic resistance or symbiosis. One of the most important class of MGEs are insertion sequences (ISs). These elements are found widespread throughout bacteria and archaea, presenting a great diversity of families. An interesting family is the IS200/605, this family is found widespread in bacteria, archaea and viruses and is divided in three subgroups according to it's genetic composition: IS200 with *tnpA* gene alone, IS605 with *tnpA* and *tnpB* and IS1341 with *tnpB* only. Another interesting aspect is the utilization of single-stranded DNA as a substrate during the transposition process. In this work, through the analysis of public available transcriptomic data we identified transcripts that overlaps the 3' end of *tnpB* in IS200/605 in both bacteria and archaea. These transcripts were named sense overlapping transcripts (sotRNAs). Sequence and secondary structure analysis showed a conserved motif present in sotRNAs, the RE-like motif. Using the consensus sequence of this motif we identified novel intergenic ncRNAs containing this motif that are derived from IS200/605. For functional characterization we overexpressed a sotRNA (VNG_sot0042) and a intergenic (VNG_R0052), both containing the RE-like motif. Standard growth curves demonstrated that the overexpression of these ncRNAs improve *H. salinarum* growth showing that this RNA is functional. To further evaluate the impact of the overexpressions we prepared RNA-seq libraries of the strain overexpressing VNG_sot0042 and in parallel performed SILAC experiments to identify potential protein-RNA interaction partners. Differentially regulated genes and interacting proteins associated with adhesion and persistent cells generation were found. Adhesion and survival assays showed that the lineage overexpressing VNG_sot0042 has a better capability to adhere in glass surfaces and survive more in diverse stressful conditions.

Keywords: Archaea, non-coding RNAs, Mobile elements, RNA-seq, Insertion sequences

Sumário

1) Introdução.....	12
1.1) Elementos transponíveis (TEs).....	12
1.2) RNAs não-codificantes e suas relações com TEs.....	14
1.3) Sequências de Inserção e derivados não-autônomos.....	16
1.4) A família IS200/605.....	20
1.5) O domínio Archaea.....	24
1.6) <i>Halobacterium salinarum</i> NRC-1.....	27
2) Objetivo Geral.....	29
2.1) Objetivos Específicos.....	29
3) Materiais e Métodos.....	30
3.1) Linhagens e condições de crescimento.....	30
3.2) Oligonucleotídeos.....	31
3.3) Identificação de 5' de transcritos através de dados de sRNA-seq.....	32
3.4) Análise do transcriptoma de <i>H. salinarum</i> NRC-1 para identificação de sotRNAs e dados de Chip-ChIP.....	33
3.5) Extração de RNA e tratamento com DNase.....	33
3.6) <i>Northern Blot</i>	34
3.7) Mineração de dados de transcriptoma para identificação de sotRNAs em diferentes organismos.....	34
3.8) Análise da estrutura secundária de sotRNAs utilizando abordagens computacionais.....	35
3.9) Circularização de RNAs para identificação de extremidades 5' e 3'.....	35
3.10) Primer extension para determinação de extremidade 5' de RNAs.....	36
3.11) Detecção de homólogos do sotRNA VNG_sot0042 utilizando a ferramenta GlassGO.....	37
3.12) Predição de mRNAs alvos de VNG_sot0042 utilizando a ferramenta CopraRNA.....	37
3.13) Construção de linhagens de superexpressão.....	38
3.14) Análise de inícios de transcrição e sítios de processamento através de dRNA-seq.....	38
3.14.1) Depleção de RNAs processados.....	39
3.14.3) Sequenciamento e processamento dos dados.....	39
.....	40
3.15) Ensaios de caracterização fenotípica.....	40
3.15.1) Curvas de Crescimento.....	40
3.15.2) Ensaios de Sobrevivência.....	40
3.15.3) Ensaios de Adesão a Vidro.....	40
3.16) Estudos de interação proteína-RNA utilizando SILAC.....	41
3.17) Preparação de bibliotecas de RNA-seq e sequenciamento.....	42
3.18) Sequenciamento e processamento dos dados de RNA-seq.....	43
3.19) Cálculo de expressão diferencial de VNG_sot0042.....	44
4) Resultados e Discussão.....	45
4.1.1) Identificação de transcritos sobrepostos ao 3' de <i>tnpB</i> de IS200/605 em dados de transcriptoma de <i>H. salinarum</i> NRC-1.....	45
4.1.2) Biogênese de sotRNAs, produtos de processamento ou transcritos primários?.....	53
4.1.3) Identificação de sotRNAs em outros organismos.....	57
4.1.4) Identificação de motivo conservado RE-like em sotRNAs e PATE-RNAs.....	58
4.1.5) Superexpressão de RNAs contendo o motivo RE-like: identificação de fenótipo em curva de crescimento.....	62
4.2) VNG_sot0042: Identificação <i>in vitro</i> de proteínas parceiras, análise <i>in silico</i> de RNAs alvo e avaliação do impacto de sua superexpressão no transcriptoma.....	64
4.2.1) Identificação de proteínas que interagem com VNG_sot0042 através de SILAC.....	64
4.2.2) Identificação <i>in silico</i> de mRNAs alvo de interação com VNG_sot0042.....	73
4.2.3) Impacto da superexpressão de VNG_sot0042 no transcriptoma de <i>H. salinarum</i> NRC-176	

4.2.4) Ensaios fenotípicos complementares.....	86
5) Conclusão.....	95
6) Referências Bibliográficas.....	98
7) Anexos.....	109
ANEXO A- Homólogos do sotRNA VNG_sot0042 encontrados utilizando a ferramenta GlassGo	109
ANEXO B- Lista completa de genes diferencialmente expressos identificados por RNA-seq da linhagem VNG_sot0042.....	120
ANEXO C – Proteínas enriquecidas nas análises de SILAC.....	125
ANEXO D - GOMES-FILHO, J. V. et al. Sense overlapping transcripts in IS 1341 -type transposase genes are functional non-coding RNAs in archaea. RNA Biology, v. 12, n. 5, p. 490–500, 4 maio 2015.....	126

1) Introdução

1.1) Elementos transponíveis (TEs)

No decorrer do século passado, o entendimento sobre a estrutura e comportamento do genoma foi drasticamente modificado. Previamente, o genoma era tratado como uma estrutura estável que evoluía lentamente através do acúmulo de mutações randômicas de frequência constante (MILLER; CAPY, 2004). Esta visão começou a ser modificada por volta de 1940, quando Barbara McClintock estudava a quebra de cromossomos e sua relação com a coloração dos grãos de *Zea mays* e identificou a presença de elementos transponíveis (*transposable elements* - TE) (NANJUNDIAH, 1996). O modelo proposto por McClintock, levava em consideração a presença de “elementos controladores”, sendo um Elemento Ativador (EA) e um Elemento de Dissociação (ED) (MCCLINTOCK, 1941). Neste contexto, o EA seria responsável pelas quebras originadas na região onde o ED se encontrava. Cerca de 30 anos após o relato de McClintock, com o avanço das técnicas de biologia molecular, demonstrou-se que este sistema é composto por elementos autônomos e não-autônomos (EA e ED respectivamente), onde o elemento autônomo é responsável por codificar a transposase necessária para mobilização do DNA (FEDOROFF; WESSLER; SHURE, 1983).

Em procariotos, a história dos TEs começou mais tardiamente durante as décadas de 1960 e 1970, quando pesquisadores verificaram a atividade de um bacteriófago, chamado Mu, capaz de inserir seu prófago em diversas regiões do genoma de *Escherichia coli* e ocasionalmente gerar mutações (TAYLOR, 1963). Em seguida, o estudo de mutações polares do operon *gal* em *E. coli* demonstrou a presença de pequenas sequências de DNA exógeno, a estas sequências deu-se o nome de Sequências de Inserção (*insertion sequence* – IS) (CHACONAS, 2002). Em paralelo a descoberta de ISs e do fago Mu, a caracterização de TEs relacionados a resistência a antibióticos em bactérias é um dos exemplos mais importantes e bem documentados. Os elementos responsáveis por carregarem os genes relacionados a resistência foram chamados de Transposons (Tn), um exemplo clássico é o transposon Tn10 que carrega genes de resistência a tetraciclina (FOSTER et al., 1981).

Devido a variação na configuração genética e dos mecanismos utilizados para mobilização dos TEs, estes são divididos em duas classes: 1) Retrotransposons, que são capazes de se amplificar nos genomas e utilizam intermediários de RNA e 2) Transposons de DNA, que

podem saltar de região em região no genoma do hospedeiro através dos mecanismos de *cut-and-paste* ou *copy-and-paste* e não utilizam intermediários de RNA (SIGUIER et al., 2015a). Os retrotransposons (TEs de Classe I) ainda são subdivididos em: 1) Longos elementos nucleares intercalados (LINEs); 2) Curtos elementos nucleares intercalados (SINEs), os quais dependem dos LINEs para sua propagação; 3) Elementos endógenos semelhantes a retrovírus (*ERV-like*) contendo sequências terminais longas e repetidas (LTRs) (ERWIN; MARCHETTO; GAGE, 2014). Os transposons de DNA (TEs de Classe II), são descritos como sequências de DNA com terminações definidas e que podem ser mobilizados entre genomas ou em regiões diferentes de um mesmo genoma, independente de processos de recombinação homóloga (NEGI; RAI; SUPRASANNA, 2016).

Os TEs são ubíquos em todos os domínios da vida, porém sua abundância e atividade variam entre espécies (GIRARD; HANNON, 2008). Em *Saccharomyces cerevisiae*, cerca de 3% do genoma é composto por TEs, já em humanos e *Z. Mays* esta proporção chega a 50% e 80% do genoma respectivamente (FLAVELL et al., 1974; KIM et al., 1998; LANDER et al., 2001). A ampla distribuição das classes mais importantes de TEs nos três domínios da vida, demonstra que estes elementos tem se mantido por milhões e milhões de anos, apesar de todo processo de seleção (GIRARD; HANNON, 2008).

A colonização generalizada dos genomas por diversas classes de TEs levanta questões relacionadas ao impacto evolutivo causado por estes elementos. Um fenômeno que nos ajuda a compreender como os genomas hospedeiros lidam com a presença destes elementos é a exaptação. A exaptação é caracterizada pela utilização de um elemento em funções as quais não estão relacionadas com sua original (GOULD; VRBA, 1982). Um exemplo ilustrativo de exaptação é a presença de penas ao longo da história evolutiva dos pássaros. Fósseis de *Archaeopteryx*, datados do período jurássico, demonstraram que estes ancestrais não eram capacitados ao voo, porém possuíam uma grande quantidade de penas por todo corpo, sugerindo que o surgimento de penas estava relacionado principalmente ao controle de temperatura (GOULD; VRBA, 1982). Ao longo da evolução, as penas foram exaptadas para novas funções relacionadas ao acasalamento e enfim ao voo (OSTROM, 1974). A nível molecular, os TEs são alvos importantíssimos do fenômeno de exaptação. A presença de TEs pode afetar diretamente a regulação de genes próximos, tanto a nível transcricional como pós-transcricional. Análise de promotores, experimentalmente caracterizados em humanos,

demonstrou que aproximadamente 25% contém sequências derivadas de transposons, incluindo elementos *cis*-regulatórios empiricamente definidos (JORDAN et al., 2003). Mecanismos celulares complexos, como o sistema de degradação proteica proteossomo-ubiquitina, também apresentam elementos exaptados de TEs. Pelo menos três dos genes envolvidos neste sistema apresentam éxons derivados de TEs, originando isoformas a nível transcricional (YING et al., 2009).

1.2) RNAs não-codificantes e suas relações com TEs

O dogma central da biologia molecular, proposto por Francis Crick e James Watson em 1958, tratava as moléculas de RNA como intermediários de DNA e proteínas (WATSON; CRICK, 1953). Em meados de 1970, estudos demonstrando a presença de longos RNAs não-codificantes (*long noncoding RNAs* - lncRNAs) e RNAs com capacidade autocatalítica em Eucariotos, foram fundamentais para iniciar um questionamento ao dogma vigente (KAUFFMAN, 1971). Nas últimas décadas, com a possibilidade de sequenciamento de genomas verificou-se que apenas 2% de todo o genoma humano corresponde a genes que codificam proteínas (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2004). Os outros 98% poderiam então representar regiões de DNA “lixo” ou conter elementos regulatórios importantes ainda não caracterizados. A presença de extensas regiões conservadas, sem potencial codificante (STEPHEN et al., 2008) e de regiões transcionalmente ativas foram evidências importantes em favor da segunda hipótese (CLARK et al., 2011). O advento das tecnologias de sequenciamento de nova geração, e.g. RNA-seq e *tiling arrays*, permitiu verificar que a maioria do genoma é transcrita. A este fenômeno, presente em procariotos e eucariotos, foi dado o nome de transcrição pervasiva (CLARK et al., 2013). Em humanos, estudos demonstraram que mais de 75% do genoma está relacionado com a produção de transcritos, em *Saccharomyces cerevisiae* este número pode chegar a 85% (DAVID et al., 2006; DJEBALI et al., 2012). Já em bactérias, a presença de mais de 1000 transcritos antisenso a genes codificantes foi detectada em *E. coli* (DORNENBURG et al., 2010). Em Archaea, análises combinadas de sítios de interação de fatores de transcrição e de dados de expressão gênica de *Halobacterium salinarum* NRC-1, permitiram a identificação de subunidades de expressão em operons, assim como diversos RNAs intergênicos (KOIDE et al., 2009). Estudos em nosso laboratório apontam que 85% das

sequências codificantes (*coding sequences* – CDS) de *H. salinarum* NRC-1 apresentam transcrição na fita oposta, em pelo menos metade de sua extensão (de Almeida, Vêncio et al, em preparação).

Os ncRNAs apresentam papéis regulatórios de alta relevância em diversos processos como simbiose, interação patógeno-hospedeiro, diferenciação celular e metabolismo. Em procariotos a maior parte dos ncRNAs são classificados como pequenos RNAs (*small* RNAs – sRNAs), os quais apresentam tamanho menor do que 200 nt, são gerados através de transcrição primária ou processamento e interagem com seus alvos através de pareamento de bases (STORZ; ALTUVIA; WASSARMAN, 2005). Os sRNAs podem alterar a estabilidade do transcrito alvo ou ocluir sítios de ligação de ribossomo em mRNAs (GOTTESMAN; STORZ, 2011). Diversas classes de ncRNAs já foram identificadas (Figura 1): RNAs antisense (*antisense* RNAs – asRNA) - que são transcritos na fita oposta de genes e apresentam alto grau de complementaridade com seu alvo (DORNENBURG et al., 2010); RNAs associados ao início de transcrição (*transcription start site associated* RNA – TSSaRNA) (CSERZO et al., 2010; SEILA et al., 2008; ZARAMELA et al., 2014); RNAs circulares (*circular* RNAs – circRNA) - que podem atuar como esponjas moleculares para outros miRNAs (WANG et al., 2014); lncRNAs (*long non coding* RNAs) - transcritos que podem alcançar milhares de nucleotídeos e que regulam funções biológicas importantes como a inativação do cromossomo X (LEE; BARTOLOMEI, 2013); miRNAs (*micro*RNAs) - são pequenos RNAs (20-30 nts) que em associação com proteínas da família Argonauta (AGO) atuam no silenciamento de diversos mRNAs (HA; KIM, 2014).

Figura 1: Diversas classes de ncRNAs já descritas. O transcrito primário é representado pela barra azul, as barras verdes indicam os possíveis variantes gerados por splicing alternativo, início de transcrição alternativo ou processamento de RNA. Nos casos onde a orientação não é indicada pelas setas o transcrito foi gerado por processamento, quando a seta está ao lado de um sinal de interrogação a origem não é bem conhecida. As diferentes espécies de RNAs geradas em torno da região codificante são representadas pelas barras vermelhas. PALR, longos RNAs associados ao promotor; PASR, pequenos RNAs associados ao promotor; PROMPT, transcritos upstream ao promotor; snoRNA, pequenos RNAs nucleolares; TASR, pequenos RNAs associados ao término de transcrição; tiRNA, pequenos RNAs associados ao início de transcrição; uaRNA, RNAs derivados da região 3'-UTR; lincRNA, RNAs longos não codificantes intergênicos. Adaptado de Clark et al., 2013.

Diversos estudos tem demonstrado a relação de TEs com a origem de genes de ncRNAs. Utilizando dados de RNA-seq e do GENCODE (HARROW et al., 2012), estima-se que, em humanos, cerca de 41% dos nucleotídeos presentes em lncRNAs são derivados de TEs, além disso 83% dos lncRNAs apresentaram pelo menos um fragmento de TE (KANNAN et al., 2015). Estas características dão origem a um fator intrigante onde diversos lncRNAs maduros contém uma combinação de múltiplos fragmentos originados de TE apresentando uma organização similar a estrutura de domínios proteicos. De acordo com estas evidências (JOHNSON; GUIGÓ, 2014) propuseram a hipótese dos chamados *Repeat Insertion Domains of lncRNAs* (RIDL), a qual sugere que TEs são exaptados e utilizados como domínios funcionais, responsáveis por favorecer interações lncRNA-RNA, lncRNA-proteína e lncRNA-DNA. Além da hipótese descrita, TEs são fontes importantes de miRNAs, como MER135 que é derivado da sequência de uma família de elementos não-autônomos chamados de MITEs (*miniature inverted-repeat transposable elements*) (PIRIYAPONGSA; MARIÑO-RAMÍREZ; JORDAN, 2007). No archaeon *Pyrococcus abyssi*, loci repetitivos encontrados em regiões intergênicas e que apresentam alta similaridade com a sequência da região 3' de *tnpB* são fontes de asRNAs que atuam em *trans* (PHOK et al., 2011).

1.3) Sequências de Inserção e derivados não-autônomos

Sequências de inserção (ISs) são elementos extremamente profusos nos genomas de procariotos e suas versões completas são compostas por um segmento de DNA relativamente pequeno (<2,5kb) e delimitados por uma repetição invertida (*inverted repeats* - IR) em cada extremidade (Figura 2).

Figura 2: Visão geral da organização de ISs. Os triângulos azuis demonstram as inversões repetidas e as caixas vermelhas as repetições diretas das ISs. As *orfs* das transposases estão representadas por setas pretas. No caso de ISs que utilizam intermediários de fita simples (neste caso IS608) as estruturas subterminais são representadas em vermelho e azul. De cima para baixo: ISs típicas das famílias IS4/IS5/IS256 apresentando apenas um quadro de leitura de transposase; Uma IS3 contendo dois quadros abertos de leitura (A e B) e que utiliza de *frameshifting* para sua expressão; Organização de um membro (IS608) típico da família IS200/605. Adaptado de Siguier et al, 2015.

Nas últimas décadas, diversos esforços para anotação e classificação acurada de ISs culminaram com a criação do banco de dados ISfinder (SIGUIER et al., 2006). Este banco é atualmente o maior repositório de IS contendo mais de 4000 IS, completas e parciais (ISs sem IRs definidos, *orfs* incompletas, etc), anotadas e divididas em mais de 25 famílias (Figura 3) (VARANI et al., 2011).

Figura 3: Distribuição das família de IS no banco de dados ISfinder. O histograma demonstra o número de IS de cada família depositada. Os subgrupos de cada família são representados por tonalidades diferentes de suas cores referentes. As ISs são divididas de acordo com a química de sua transposase. Azul: DDE. Roxo: Não determinada. Verde: DEDD. Vermelho: HUH. Laranja: Serina. Adaptado de Siguier et al, 2015.

As IS são classificadas, em primeira instância, de acordo com a química utilizada por sua transposase. Além da química de suas transposases, considera-se também a composição dos IR, tamanho das repetições diretas que estão flanqueando a IS, organização dos quadros de leitura (*open reading frames* - ORFs) e a natureza de suas sequências alvo (SIGUIER et al., 2006). As transposases podem ser classificadas em: 1) Transposases do tipo DDE: possuem uma tríade de aminoácidos (Asp-Asp-Glu) que coordena íons de metal e utilizam OH como nucleófilo numa reação de transesterificação, sem a formação de intermediário transposase-DNA (SIGUIER; GOURBEYRE; CHANDLER, 2014); 2) Transposases do tipo DEDD: Possuem um quarteto de aminoácidos (Asp-Glu-Asp-Asp) e são relacionadas com a resolvase RuvC (MAHILLON; CHANDLER, 1998); 3) Transposases do tipo HUH: possuem dois resíduos de histidina separados por um aminoácido hidrofóbico e utilizam tirosina como nucleófilo para gerar um intermediário 5' tirosina-DNA ligado covalentemente (HE et al.,

2013); 4) Serina transposases: Estão presentes apenas na família IS607 apresentando semelhança com uma serina recombinase sítio específica (BAO; JURKA, 2013); 4) Tirosina transposases: Até o momento não se conhece ISs que utilizam este tipo de enzima, acredita-se que esta transposase utiliza uma tirosina catalítica para a formação de intermediários 3' fosfo-tirosina entre a enzima e o DNA (SIGUIER et al., 2015b).

A inserção de uma IS sempre causará algum tipo de alteração no genoma hospedeiro, enquanto sua excisão pode restaurar o genoma a conformação original ou originar uma mutação (SIGUIER; GOURBEYRE; CHANDLER, 2014). Os impactos da mobilização de uma IS são variáveis de acordo com o sítio alvo do processo de inserção. Caso a região de inserção esteja localizada no meio de um gene ou em suas sequências reguladoras, como seu promotor, a inativação do gene pode ocorrer. Em *Acinetobacter baumannii*, a inserção de ISAba1 *upstream* ao gene *ampC* resultou na superexpressão da betalactamase ADC (*Acinetobacter-derived cephalosporinase*), conferindo assim resistência a um espectro estendido de cefalosporinas (CORVEC et al., 2003). Além de interferir na expressão de genes, pares de IS da mesma família podem atuar como transposons compostos, mobilizando material genético através da transferência lateral de genes (MUGNIER; POIREL; NORDMANN, 2009).

A regulação da expressão de transposases é realizada de maneira extremamente controlada, com elementos atuando desde a transcrição da transposase até as etapas de transposição (NAGY; CHANDLER, 2004). No controle da transcrição, a presença de promotores fracos é um fator comum, porém não universal. Em IS21, por exemplo, a expressão *in vivo* do promotor endógeno é indetectável (REIMMANN et al., 1989). Outros fatores importantes para regulação da expressão de transposons são: a presença de terminadores de transcrição ao longo do gene (MAHILLON; CHANDLER, 1998), utilização de promotores alternativos (WEINREICH; REZNIKOFF, 1992) e estabilidade do mRNA (SAWERS, 2005). A nível de tradução, a presença de RNAs antisense é bem descrita para diversas IS. Em IS10, a presença de um promotor (pOUT) localizado dentro da *orf* da transposase, porém no sentido oposto, resulta na produção de um pequeno RNA (70 nt) chamado de RNA-OUT (MA, CHARLES AND SIMONS, 1990). Este RNA é extremamente estável, estruturado e complementar a região 5' do mRNA da transposase. A interação de RNA-OUT com o mRNA resulta no sequestro dos fatores relacionados ao início de tradução e

consequente diminuição da concentração da proteína (ROSS; WARDLE; HANIFORD, 2010). Além da presença de asRNAs, a regulação da tradução pode ocorrer pela mudança programada do quadro de leitura (CHANDLER; FAYET, 1993), sítios alternativos de início de tradução (REIMMANN et al., 1989), término de tradução (CHANDLER; MAHILLON, 2002) e controle por temperatura (SHIGA et al., 2001).

Além das ISs completas, os genomas procarióticos são infestados por pequenos fragmentos de ISs que podem ser mobilizados em *trans* por transposases originadas do elemento completo (SIGUIER; GOURBEYRE; CHANDLER, 2014). Estes elementos foram inicialmente identificados em plantas e chamados de *miniature inverted terminal repeats* (MITEs) e estão relacionados com transposases do tipo DDE (HE et al., 2015a). Em *Haloquadratum walsbyi*, elementos similares derivados de IS200/605, chamados de *palindrome associated transposable elements* (PATEs), foram encontrados e análises comparativas do genoma de duas linhagens distintas evidenciaram que PATEs são originados de eventos de deleção da parte central de IS200/605 (Figura 4) (DYALL-SMITH et al., 2011).

Figura 4: Origem de PATEs através da deleção da parte central de uma IS200/605 (setas cinza e preta). O resultado da deleção corresponde a fusão das extremidades da IS (*hairpins* azul e vermelho). Adaptado de Siguier et al, 2015.

1.4) A família IS200/605

Dentre as diversas famílias de ISs, a família IS200/605 encontra-se altamente dispersa em diversas espécies de Bacteria e Archaea e pode ser composta por dois genes, *tnpA* (Transposase HUH) e/ou *tnpB* (não essencial ao processo de transposição) (SIGUIER et al., 2015a). Análises filogenéticas demonstraram que esta família é a mais antiga já detectada para o domínio Archaea (FILÉE; SIGUIER; CHANDLER, 2007).

Esta família pode ser dividida em três subgrupos, organizados de acordo com a composição genética do elemento: 1) IS200: Grupo que contém somente *tnpA*; 2) IS605: Grupo que

contém ambos genes *tnpA* e *tnpB*; 3) IS1341: Grupo que contém somente *tnpB* e é considerado elemento não-autônomo (Figura 5) (SIGUIER; GOURBEYRE; CHANDLER, 2014).

Figura 5: Divisão da família IS200/605. I) Membros do grupo IS200 contendo apenas *tnpA*. II, III, IV) Membros do grupo IS605 contendo *tnpA* e *tnpB*. V) Membros do grupo IS1341, contendo apenas *tnpB*. (Adaptado de Siguier et al, 2014)

Características peculiares desta família são: o mecanismo de transposição utilizado, a presença de palíndromos imperfeitos nas duas extremidades do elemento e a presença das menores transposases já identificadas (<150 aminoácidos) (HO, 2009). Para a movimentação da IS, tanto o DNA que será clivado quanto o que receberá a IS devem ser fita simples. Este mecanismo, chamado de *peel-and-paste*, foi bem descrito e caracterizado para as bactérias *Helicobacter pylori* e *Deinococcus radiodurans* (HE et al., 2015b). Durante o processo de transposição, dois monômeros da proteína TnpA se ligam, um no motivo LE (*left end*), localizado na região 5' e outro no motivo RE (*Right end*) localizado na região 3' (Figura 6A). A presença destes motivos, que formam estruturas secundárias conservadas, é importante para o reconhecimento correto da região onde a proteína realizará a clivagem (RONNING et al., 2005). Após a ligação nos motivos RE e LE, ocorre a clivagem e posterior dimerização das TnpA envolvidas no processo (Figura 6B). Após a excisão da fita ocorre o religamento do DNA que flanqueava a IS, gerando um sítio capaz de receber a inserção de outras IS200/605

(Figura 6 B). A inserção da fita que foi excisada, depende da presença de sequências alvo específicas, sendo estas tetra ou pentanucleotídeos (Figura 6B).

Figura 6: Modelo de transposição de elementos da família IS200/605. **A)** Organização dos motivos encontrados no DNA importantes para o reconhecimento e início do processo de transposição. **B)** Esquematização da movimentação da transposase durante o processo de transposição. (i) A transposase TnpA formando dímero (verde e laranja claro) com seus sítios ativos na forma *trans*. Estes sítios são compostos por um motivo HUH de um dos monômeros e um resíduo de tirosina do outro monômero (verde e laranja escuro). (ii) O intermediário transposase-DNA de IS200/605 carregando os motivos LE e RE (vermelho e azul, respectivamente). (iii) Complexo transposase-DNA ligado covalentemente após clivagem na presença de Mg^{2+} . (iv) Rotação das hélices dos monômeros. (v) Formação de um sítio ativo em cis (vi). Formação da junção RE-LE-transposon. Adaptado de He et al, 2013.

Enquanto os mecanismos de ação de TnpA são extremamente bem detalhados, as funções de TnpB ainda estão em estágio inicial de conhecimento. TnpB é uma proteína de aproximadamente 400 aminoácidos e seu maior nível de conservação é encontrado na região C-terminal, onde um domínio 4-Cys *Zinc Finger* é localizado (BAO; JURKA, 2013). Os domínios *Zinc Finger* são encontrados em diversas proteínas e funcionam como módulos de interação a diversos ligantes como proteínas, pequenas moléculas e ácidos nucleicos (LAITY;

LEE; WRIGHT, 2001). Além do domínio *Zinc Finger*, estas proteínas podem apresentar um domínio *RuvC-like* (Figura 7) (KAPITONOV; MAKAROVA; KOONIN, 2016).

Figura 7: Alinhamentos múltiplos de sequências de representantes de TnpB e RuvC. Resíduos catalíticos: destaques em preto. Resíduos hidrofóbicos conservados: destaques em amarelo. Pequenos resíduos conservados: destaques em verde. Resíduos carregados positivamente: destaques em vermelho. As letras em verde representam a estrutura predita para a região E para folhas beta e H para alfa-hélice. Adaptado de Kapitonov et al, 2016.

A principal função da proteína RuvC em bactérias está relacionada a resolução de junções de Holliday durante o processo de recombinação e reparo de quebras de DNA (GARCIA et al., 2000). Os eventos de transposição podem levar a um aumento no número de quebras no DNA, elevando a ocorrência de eventos deletérios prejudiciais tanto para o genoma hospedeiro quanto para o transposon. Deste modo, a co-evolução de ISs contendo TnpB seria mais provável e bem sucedida devido a minimização de danos causados pelos eventos de transposição (PASTERNAK et al., 2013). Estudos demonstraram que TnpB está envolvida com a repressão tanto do processo de excisão quanto inserção de IS200/605 em *D. radiodurans*, reforçando a hipótese de contenção de danos no genoma hospedeiro (PASTERNAK et al., 2013). Um fato intrigante foi o estabelecimento de que proteínas Cas9, pertencentes aos sistema CRISPR-Cas de imunidade adaptativa em procariotos, são homólogas de TnpB. Ambas contém o domínio *RuvC-like*, sugerindo que o sistema CRISPR-Cas do tipo 2 evoluiu através da inserção de um transposon contendo um ancestral de Cas9 em um *locus* de CRISPR-Cas do tipo I (KAPITONOV; MAKAROVA; KOONIN, 2016). Outro fator interessante é a prevalência nos domínios Bacteria e Archaea, fato que se torna surpreendente quando associado a não necessidade desta proteína para o processo de transposição de IS200/605 (FILÉE; SIGUIER; CHANDLER, 2007).

Nos últimos anos, ncRNAs associados a IS200/605 foram identificados, gerando outro

ponto de discussão sobre o caráter evolutivo e funcional destas IS. Em *Salmonella typhimurium*, o ncRNA STnc490 apresenta-se antisense a IS200, sobrepondo-se 90 nucleotídeos da região 5' UTR do mRNA da transposase TnpA (KRÖGER et al., 2012). O mesmo ncRNA foi encontrado em outro elemento chamado IS1541 de *Yersinia pestis*, sugerindo que este RNA é conservado (YAN et al., 2013). Estudos recentes evidenciaram que este RNA, agora renomeado para art200 (*antisense regulator of transposase* IS200), liga-se ao mRNA de TnpA e atua como um repressor extremamente eficiente da expressão da transposase (ELLIS; TRUSSLER; HANIFORD, 2015).

Estas características reforçam o caráter elusivo de ISs contendo o gene *tnpB*, principalmente em elementos não-autônomos como IS1341.

1.5) O domínio Archaea

A caracterização taxonômica clássica de microrganismos apresenta-se como um grande desafio desde a sua concepção. A utilização de características morfológicas, fenotípicas e metabólicas são extremamente deficientes quando utilizadas para o estabelecimento de relações evolutivas entre organismos (WOESE; FOX, 1977). No final da década de 1960, Carl Woese e colaboradores realizavam estudos focados na caracterização filogenética de bactérias e sua possível relação com o último ancestral comum (LUCA – *last universal common ancestor*) (KOONIN, 2014). Os principais marcadores moleculares utilizados para construção de relações filogenéticas eram proteínas (ZUCKERKANDL; PAULING, 1965). Uma das propostas inovadoras de Woese foi utilizar ácidos nucleicos, mais especificamente a subunidade menor do RNA ribossomal (rRNA), para reconstrução de relações filogenéticas (LUEHRSEN, 2014). Utilizando esta abordagem, alguns resultados inesperados foram encontrados. A presença de um grupo monofilético distinto a eucariotos e bactérias foi estabelecido (WOESE; FOX, 1977). A este novo domínio, foi proposto o nome de Archaea, em referência ao Éon Arqueano. O trabalho de Woese não foi prontamente aceito e apresentou grande resistência por parte dos taxonomistas clássicos. Inconsistências encontradas em filogenias realizadas com proteínas e a idéia de que Archaea eram limitadas a ambientes específicos da terra contribuíram para a demora na aceitação das propostas de Woese (CAVICCHIOLI, 2011). Conforme novos estudos relacionados a Archaea foram publicados, a árvore da vida com três domínios se estabeleceu e hoje é o modelo mais aceito pela

comunidade científica (Figura 8) (ALLERS; MEVARECH, 2005).

A divisão do domínio Archaea é composta por dois filos principais, sendo eles: 1) Euryarcheota: onde encontramos Archaea metanogênicas, halófilas, termófilas e psicrofílas; 2) Crenarchaeota: onde encontramos Archaea hipertermófilas, com temperaturas de crescimento acima de 80°C. Em relação às características moleculares que diferenciam estes filos, a presença de histonas muito similares a de eucariotos em Euryarcheota se contrapõem ao conjunto de proteínas ligantes de DNA presentes em Crenarchaeota (WHITE; BELL, 2002). Outros dois filos, Nanoarchaeota e Korarchaeota, apresentam poucos representantes e muitas espécies não cultiváveis, deste modo sua posição na árvore filogenética ainda é incerta (ALLERS; MEVARECH, 2005).

Figura 8: Quatro grupos do domínio Archaea. Euryarchaeota é o mais diversificado englobando organismos metanogênicos e halófilos, além de alguns psicrofílos e termófilos. O grupo Crenarchaeota apresenta os organismos hipertermófilos. Os grupos Nanoarchaeota e Korarchaeota apresentam poucas espécies e o posicionamento dos seus filos na árvore filogenética ainda não foi determinado (linhas tracejadas). Adaptado de Allers & Mevarech, 2005.

As Archaea apresentam diversas características compartilhadas com bactérias e eucariotos, formando um verdadeiro mosaico (Tabela 1). A organização genômica, apresentando poucas regiões intergênicas e a presença de diversos genes em operon é uma característica compartilhada por procariotos em geral (ALLERS; MEVARECH, 2005). Já em relação a maquinaria de transcrição, por exemplo, vemos maior semelhança com a utilizada por eucariotos, onde a RNA polimerase II utiliza fatores basais, proteínas de ligação a TATA box e fator de transcrição B, para ligação eficiente a regiões promotores além da presença de motivos BRE-TATA (HICKEY; DE MACARIO; MACARIO, 2002). Além das semelhanças, as Archaea apresentam diversas características exclusivas como a composição da bicamada lipídica da membrana e a capacidade de realizar metanogênese (CAVICCHIOLI, 2011). A membrana celular de Archaea apresenta uma cadeia lateral de isoprenóides que se encontram unidos por ligações glicerol-1-fosfato, já em bactérias e eucariotos esta união é realizada por ligações glicerol-3-fosfato (JAIN; CAFORIO; DRIESSEN, 2014). Esta configuração diferenciada das ligações tem sido relacionada com a capacidade adaptativa de Archaea em diversos ambientes inóspitos.

Tabela 1: Características apresentadas por bactérias, Archaea e eucariotos. Adaptado de Cavicchioli, 2010.

Características	Bactéria	Archaea	Eucariotos
Ligação dos carbonos nos lipídios	Éster	Éter	Éster
Fosfatos dos lipídios da membrana	Glicerol-3-fosfato	Glicerol-1-fosfato	Glicerol-3-fosfato
Metabolismo	Bacteriano	Similar ao bacteriano	Eucariótico
Maquinaria de transcrição	Bacteriano	Similar ao eucariótico	Eucariótico
Fatores de elongação da tradução	Bacteriano	Similar ao eucariótico	Eucariótico
Núcleo	Não	Não	Sim
Organelas	Não	Não	Sim
Metanogênese	Não	Sim	Não
Patógenos	Sim	Não	Sim

A presença de maquinarias moleculares extremamente adaptadas a ambientes extremos, tornam as Archaea organismos de altíssimo valor a indústria biotecnológica. A prospecção de enzimas termoresistentes com aplicação em biologia molecular, como *Pfu* de *Pyrococcus furiosus* amplamente utilizada em reações de PCR, tem sido bastante lucrativas. Outras

possíveis aplicações estão relacionadas a bioremediação de solos contaminados (OREN, 2015), extração de celulasas mais eficientes (XU et al., 2016) e biofilmes (ORELL; FRÖLS; ALBERS, 2013).

Além do interesse biotecnológico, as Archaea são organismos de grande interesse para os estudos de Astrobiologia. A capacidade de diversas espécies de Archaea em habitar ambientes extremos, similares a terra pré-biótica, sugere que estes organismos poderiam representar formas de vida encontradas em exoplanetas ou participar do processo de panspermia (CAVICCHIOLI, 2011).

Apesar dos membros mais estudados deste domínio serem extremófilos, cada vez mais a sua presença e importância em ambientes mais amenos tem sido demonstradas. Em águas oceânicas, por exemplo, as Archaea tem papel importante em ciclos biogeoquímicos (DELONG, 1998).

1.6) *Halobacterium salinarum* NRC-1

Dentro do domínio Archaea, uma das espécies que mais se destaca e que se estabeleceu como modelo é *Halobacterium salinarum* NRC-1. Este organismo, apesar de ser um halófilo obrigatório, é de fácil cultivo em laboratório apresentando condições ideais de crescimento a 37°C e 4~5M de NaCl (KAUR et al., 2006a). Apresenta morfologia de bastonete, com flagelos polares e uma camada de glicoproteínas que envolvem a membrana citoplasmática (Figura 9).

Figura 9: Da esquerda para direita: Microscopia eletrônica evidenciando a forma de bastonetes e a presença de flagelos. Cristais de sal contendo *H. salinarum* NRC-1. Microscopia eletrônica de corte transversal evidenciando a presença de vesículas de gás. Adaptado de DasSarma, 2012.

Em relação a estrutura do genoma, encontramos um genoma compacto de

aproximadamente 2.6Mb composto por três replicons, um cromossomo de cerca de 2Mb e dois plasmídeos, que também são chamados de minicromossomos por conterem genes essenciais, (pNRC100 e pNRC200) de 190Kb e 360Kb respectivamente (NG et al., 2000). O cromossomo apresenta conteúdo GC de 68%, já os plasmídeos apresentam valores menores pNRC100 58% e pNRC200 59%. Os plasmídeos apresentam uma região de identidade de 145kb, na qual existem 2 repetições invertidas de 33kb e 39kb importantes nos processos de inversão e isomerização no genoma. No total, 2629 genes foram anotados, destes 676 (25,7%) são hipotéticos. Em relação ao seu proteoma, encontramos proteínas ácidas com pI médio de 5,1, demonstrando um grande nível de adaptação aos ambientes de alta salinidade (NG et al., 2000). Além da linhagem NRC-1, a linhagem R1 também é bastante estudada. As duas linhagens apresentam grande semelhança no genoma e possivelmente surgiram do mesmo isolado (PFEIFFER et al., 2008). Em relação ao cromossomo, encontramos 12 alterações entre as duas linhagens, a maior delas sendo uma inserção de aproximadamente 10kb em NRC-1 originada de um evento de transposição (NG et al., 2000).

Para lidar com as condições extremas de seu habitat, *H. salinarum* NRC-1 apresenta a capacidade de obtenção de energia por diversas vias, além de possuir um sistema simples e eficiente de armazenamento de energia. A oxidação de compostos orgânicos, principalmente aminoácidos, pelo ciclo do ácido tricarboxílico (TCA) é realizada em conjunto a uma cadeia de transporte de elétrons utilizando oxigênio como aceptor final (GHOSH; SONAWAT, 1998; SCHÄFER; PURSCHKE; SCHMIDT, 1996). Uma característica de regiões hipersalinas é baixa solubilidade do oxigênio, ocasionando numa baixa disponibilidade do mesmo. Quando isto ocorre, *H. salinarum* NRC-1 pode utilizar a respiração anaeróbica usando outros aceptores de elétrons, como dimetil sulfóxido, nitrato e fumarato (OREN; TRUPER, 1990). Outra via de obtenção de energia é a fermentação de L-arginina através da via de arginina deiminase (RUEPP; SOPPA, 1996). Por fim, *H. salinarum* NRC-1 apresenta a capacidade de realizar fotossíntese através de proteínas retiniais chamadas de bacteriorodopsina (BR), a qual bombeia prótons do citosol para o meio extracelular, resultando em um gradiente que pode ser utilizado pela ATP sintase para produzir ATP (SCHÄFER; ENGELHARD; MÜLLER, 1999).

Por ser um organismo de fácil cultivo e manipulação *H. salinarum* NRC-1 se tornou um organismo modelo para estudos em Archaea. Deste modo, existe um grande compêndio de estudos que avaliam o comportamento deste microrganismo em diversas condições genéticas

e ambientais. Exemplos disso são as análises da dinâmica do transcrito em exposição a metais pesados e ao longo de uma curva de crescimento (KAUR et al., 2006b; KOIDE et al., 2009).

De acordo com o banco de dados ISfinder, atualmente encontram-se anotadas 114 IS (parciais e completas) no genoma de *H. salinarum* NRC-1. Em relação a IS200/605 encontramos 10 elementos anotados. Cerca de 1.7% do cromossomo é composto por ISs, enquanto os plasmídeos pNRC100 e pNRC200 apresentam, respectivamente, 20% e 14% evidenciando um grande potencial de plasticidade genômica (NG et al., 2000). O número elevado de IS, associado com a plethora de dados de expressão já disponíveis, tornam este organismo extremamente interessante para estudo da dinâmica de IS e de possíveis RNAs associados.

2) Objetivo Geral

Este trabalho teve como objetivo identificar e caracterizar uma nova classe de RNAs não-codificantes associados a *tnpB* de IS200/605 em *H. salinarum* NRC-1.

2.1) Objetivos Específicos

Identificação de RNAs sobrepostos a *tnpB* de IS200/605 em *H. salinarum* NRC-1 (*sense overlapping RNA* - sotRNAs)

- Identificação de transcritos sobrepostos a região 3' de *tnpB* utilizando dados de *tiling microarrays* e RNA-seq
- Mapeamento das extremidades 5' e 3' de transcritos selecionados utilizando C-RACE
- Validação da existência desses RNAs utilizando *northern blot* e *primer extension*
- Predição de estrutura secundária e identificação de possíveis alvos de regulação por métodos computacionais
- Análise de conservação em outros procariotos utilizando dados públicos e métodos computacionais

Caracterização de sotRNAs: integração de abordagens experimentais *in vivo*, *in vitro* e *in silico*

- Identificação de proteínas parceiras de interação através do método de SILAC
- Construção de linhagem superexpressando um sotRNA
- Avaliação do impacto da superexpressão de sotRNAs em *H. salinarum* NRC-1 através de ensaios de curva de crescimento, ensaios de sobrevivência e ensaios de adesão
- Avaliação do impacto da superexpressão de sotRNAs no transcrito de *H. salinarum* NRC-1

3) Materiais e Métodos

3.1) Linhagens e condições de crescimento

As linhagens de *H. salinarum* utilizadas foram: *H. salinarum* NRC-1, *H. salinarum* VNG_sot0042, *H. salinarum* VNG_R0052 e *H. salinarum* PMTF-0 (Tabela 2). O cultivo deste microrganismo foi realizado em meio nutriente CM (NaCl 250 g/l; MgSO₄. 7H₂O 20 g/l; KCl 2 g/L; citrato de sódio 3 g/L; peptona 10 g/L), com agitação podendo variar entre 125-250 RPM, sempre na presença de luz. Quando necessário, foi suplementado com o antibiótico mevinolina (20 µg/ml). As linhagens de *Escherichia coli* DH5α foram cultivadas a 37° C sob agitação, em meio LB (triptona 10 g/l; extrato de levedura 5 g/l; NaCl 10 g/l; pH 7,5), podendo ou não ser suplementadas com antibiótico carbenicilina (50 µg/ml).

Tabela 2: Linhagens utilizadas nos experimentos descritos neste documento.

Linhagem	Utilização
<i>H. salinarum</i> NRC-1	Linhagem selvagem
<i>H. salinarum</i> PMTF-0	Linhagem selvagem com vetor de superexpressão vazio
<i>H. salinarum</i> VNG_sot0042	Superexpressão de VNG_sot0042
<i>H. salinarum</i> VNG_R0052	Superexpressão de VNG_R0052
<i>E. coli</i> DH5 α	Linhagem utilizada para procedimentos de clonagem

3.2) Oligonucleotídeos

Tabela 3: Oligonucleotídeos utilizados no desenvolvimento do trabalho. Todos os oligos, com exceção dos marcados com fluoróforos foram sintetizados pela empresa Exxtenda. Os oligos marcados foram sintetizados pela IDT. Aplicações: E) Primer extension; PR) Oligos para construções utilizadas no método de SILAC; DN) Confirmação de tratamento com DNase; SS) Sequenciamento de sanger; N) Northern blot; C) Circularização; CL) Construção de vetores de superexpressão.

Nome	Marcação	Aplicação	Sequência 5' → 3'
VNG2652 – NED	NED	E	TAACACTCCCTGTTCCGCGAG
VNG2652 – VIC	VIC	E	TAACACTCCCTGTTCCGCGAG
sot0042_1_for	-	PR	CGTTAATACGACTCACTATAGGGACGTATCTCCGAGTCCCGTCACGGATGGCGGTGATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTCTGT
sot0042_2_for	-	PR	GGCACAGCCATCGACGTTCTGTGTTTGACAATGAAACTGGTGCGTTTCGCACCTCAAGAACAGGCTACGTCGTAAACCATAATATCCCAACGCG
sot0042_3wt_rev	-	PR	ACAGGATAGCTGAGACTGCAGCATTGACATCCTCCCCGCCCTGAAGGGCGAGGATTCGGACCGCGTTGGGATATTATGGTTTAC
sot0042_3dm_rev	-	PR	ACAGGATAGCTGAGACTGCAGCACCGACCGCGTTGGGATATTATGGTTTAC
sot0042_4_rev	-	PR	CATGGCCCCGGCCCGCGACTATCTTACGCACTTGATGATTCTGGTTCGGTCCCATGGATCCACAGGATAGCTGAGACTGCAGCA
Random206_1	-	PR	CGTTAATACGACTCACTATAGGGCGATTCAATAACCCCTCGCGATCATCATCGGTGGTAAATCGCCCGCAAGGATTAGCGATTCTTAGAGTAGTTAT
Random206_2	-	PR	ATTAGCGATTCTTAGAGTAGTTATCCTAGTCGCGGGGAGGGGGGTGCACGGCCACCTACGTCGAGGACTTCATACTCGAGCGTGGAGTCTCCG
Random206_3	-	PR	GAGACTGATGTGCCCACGTAGCGACCCAAAGGCATATGTTAGAATTGCACTCGGATTTGTCCGGAGACTCCACGCTCGAG
Random206_4	-	PR	CATGGCCCCGGCCCGCGACTATCTTACGCACTTGATGATTCTGGTTCGGTCCCATGGATCCGAGACTGATGTGCCACGCTAGCGAC
VNG_R0052wt_for	-	PR	CGTTAATACGACTCACTATAGGGCGGCCAAGGTGGACTGTTACGGAGAATCGAGAAGAAAGC CTCGCCCTTCAGGGCGGGGAGGATGTCAAAAGTAGAATCCACCCCAAGTTTAG
VNG_R0052dm_for	-	PR	CGTTAATACGACTCACTATAGGGCGGCCAAGGTGGACTGTTACGGAGAATCGAGAAGAAAGTAGAATCCACCCCAAGTTTAG
VNG_R0052both_rev	-	PR	CATGGCCCCGGCCCGCGACTATCTTACGCACTTGATGATTCTGGTTCGGTCCCATGGATCCTCAGTCTAAACTTCTGGGTGGATTCTACT
+19 pNRC200Fow	-	DN	AACCACCGTGACCCGGAAGC
+19 pNRC200Rev	-	DN	GCCGATCCGATGGGAGTGACT
16S Fow	-	DN	GTGAGGGACGAAAGCTAGGG
16S Rev	-	DN	AAGGGTACTGCTGGCAATTA
M13 Fow	-	SS	AGCGGATAACAATTCACACAG
M13 Rev	-	SS	CGCCAGGGTTTTCCAGTCACGAC
Northern_VNG0042	-	N	GACGTAGCCTGTTCTTGAGGTGCGAACGCACCAAGTTTCAT
Northern_VNG2652	-	N	CCCTGTTCCGCGAGGGAACGGGATGAGGTGGGCGGCTTA
CircVNG0042-F	-	C	ATG GTTTAC GAC GTA GCC TGTTC
CircVNG0042-R	-	C	GTT CCT GTT TGA CAA TGA AAC TG
0042AT – F	-	CL	GCC GGG CAT ATG ACG TAT CTC CGA GTC CCG TCA
0042AT – R	-	CL	GCG GGG GGA TCC ACA GGA TAG CTG AGA CTG CAG CAT
CircVNG0044-F	-	C	TGCATGGGAATACCGGTTAACTA
CircVNG0044-R	-	C	TCGGGTGCGGAACCCACTGCCG

3.3) Identificação de 5' de transcritos através de dados de sRNA-seq

Dados públicos de RNA-seq de pequenos RNAs (< 200 nt) (SRA: SRX433542) foram utilizados para identificar o 5' de ncRNAs com alta resolução (ZARAMELA et al., 2014).

Para avaliar possíveis inícios de transcritos foram gerados perfis de *start* que contabilizam apenas a primeira posição de cada *reads*. Esse perfil de *start* foi utilizado para mapear picos que representam possíveis inícios de transcritos e então avaliar a composição dos reads que compõem os picos. Como as bibliotecas foram construídas com RNAs não fragmentados, reads que apresentam tamanho menor que o limite do sequenciamento (151pb) representam potenciais transcritos completos. A mesma técnica foi aplicada para identificação de transcritos associados ao início de transcrição – TSSaRNAs (ZARAMELA et al., 2014).

Figura 10: Ilustração do procedimento adotado para identificação dos TSSaRNAs em (Zaramela et al., 2014) e utilizado neste trabalho para identificação do 5' de sotRNAs. Caixa amarela representa a CDS do gene cognato. Barras verticais verdes representam a contagem dos *starts* dos reads mapeados. O círculo cinza indica o pico de *start* mais enriquecido. O retângulo cinza ilustra o grupo de reads que possuem a primeira base mapeada na mesma coordenada genômica. O grupo é composto por duas populações de reads: reads idênticos que representam o TSSaRNA e outros reads que representam transcritos originados do gene cognato. Barras horizontais pretas representam as regiões utilizadas na busca pelo pico de *start* mais frequente, a partir do início do CDS. Adaptado de (Zaramela et al., 2014).

3.4) Análise do transcrito de *H. salinarum* NRC-1 para identificação de sotRNAs e dados de Chip-ChIP

Dados públicos de *tiling microarrays* foram utilizados para interrogar a dinâmica de expressão dos ncRNAs descritos neste trabalho. Dentre os dados de *tiling* temos: 13 pontos ao longo de uma curva de crescimento de *H. salinarum* NRC-1 (GSE12923), superexpressão do fator de transcrição TfbD (GSE15788), nocaute do fator de transcrição Tfb e superexpressão de versões sinteticamente melhoradas (GSE31308), nocaute da RNase VNG2099C (GSE45988), variação de condições extracelulares de salinidade (GSE53544). As razões de expressão normalizadas foram obtidas diretamente dos registros GEO. Os dados de curva de crescimento e superexpressão de TfbD foram visualizados utilizando Gaggly Genome Browser (GGB).

Os sítios de ligação de diversos fatores de transcrição de *H. salinarum* NRC-1 foram obtidos de dados previamente publicados (FACCIOTTI et al., 2007; KOIDE et al., 2009; PLAISIER et al., 2014) e visualizados utilizando GGB. Esse dado representa um compêndio dos sítios de interação de todos os TFBs (TFBa, TFBb, TFBc, TFBd, TFBe, TFBf, and TFBg), quatro TBPs (TBPb, TBPc, TBPe e TBPf) e três reguladores da transcrição (Trh2, Trh3, Trh4), totalizando 13981 posições no genoma de *H. salinarum* NRC-1.

3.5) Extração de RNA e tratamento com DNase

Para a extração dos RNAs, utilizamos o kit mirVana miRNA Isolation (Ambion). *Pellets* de *H. salinarum* NRC-1 passaram pelo processo de lise celular e posterior remoção dos resíduos celulares através de extração com fenol/Clorofórmio. Após a etapa de extração orgânica adicionou-se 1,25X volume de etanol 100% a temperatura ambiente. A etapa de adição de etanol é crítica pra determinar o grau de afinidade dos RNAs pelas colunas de extração, dessa forma a adição de etanol na quantidade descrita promoveu a ligação de todos os RNAs nas colunas. Após passagem na coluna, os RNAs foram lavados com os tampões fornecidos no kit e então eluídos em água DEPC a 95° C e quantificados em espectrofotômetro.

Adicionou-se 50 µg de RNA a um microtubo de 1,5 ml, 10 µl de Turbo DNase I (Fermentas), 15 µl de Turbo DNase I 10X Buffer, 0,5 µl de Ribolock e volume de água DEPC para completar 150 µl. A reação foi incubada a 37° C por 45'. Em seguida, foram

adicionados 150 µl de fenol-ácido a amostra, que foi vortexada e prontamente centrifugada a 10.000 RPM por 5'. A fase aquosa foi transferida para novo microtubo onde 1,5 ml e 2,5x o volume de Etanol 100% e 1/10 de NaOAC 3 M foram adicionados. A reação foi incubada a -20° C *overnight*. A amostra foi centrifugada a 14.000 RPM, 4° C por 30', o sobrenadante desprezado e o *pellet* lavado com 500µl de Etanol 70%, em seguida a amostra foi centrifugada por 5' a 10.000 RPM. A amostra foi seca a 65° C por 10' e ressuspendida em água MiliQ DEPC.

3.6) Northern Blot

Para o ensaio de *northern blot*, 40 µg de RNA total tratado com DNase foram separados em gel de poliacrilamida 8% a 100 V durante 1 hora. O gel foi corado com 5 µl de brometo de etídeo 10 mg/mL, por 20' e documentado através do foto-documentador Bio Rad ChemiDoc RX (BioRad). A transferência do gel para as membranas Hybond N+ (GE Healthcare) foi realizada a 10V por 60 minutos utilizando o sistema Trans-Blot SDS e miDryTransferCell (Biorad). O *crosslinking* foi realizado utilizando UVC 500 Ultraviolet Crosslinker (Ameshan Biosciences) aplicando 70000 µJ/cm² na membrana. Para a marcação das sondas com γP³²-ATP foi utilizada a enzima T4 Polynucleotide Kinase (NEB) a 37° C por 1 hora. A reação de marcação foi realizada com 150 ng de oligonucleotídeo, 1 µL de T4 Kinase, 4 µL γP³²-ATP, 5 µL tampão 10X. As membranas foram pré-hibridizadas com 10 ml de tampão de hibridização Rapid-hyb Buffer (GE Healthcare) a 42° C, com rotação em forno de hibridação MiniOven OV2 (Biometra) por 1 hora. Após a marcação, a sonda foi purificada utilizando o kit illustra MicroSpin G-25 (GE Healthcare), transferida para o tampão de hibridação e sua atividade foi registrada utilizando contador geiger. A hibridação foi realizada a 42° C por 2 horas. As lavagens foram feitas com 50 ml de tampão (2x SSC; 0,1%SDS) por 20' a temperatura ambiente, seguida de 50 ml de tampão (0,2x SSC; 0,1% SDS) por 15' a 65° C. As membranas foram expostas a filme Amersham Hyperfilm ECL por 72 horas.

3.7) Mineração de dados de transcrito para identificação de sotRNAs em diferentes organismos

sotRNAs para genes *tnpB* foram encontrados em diferentes procariotos utilizando a plataforma BLASTn-SRA (<https://blast.ncbi.nlm.nih.gov/Blast.cgi?>

[PROGRAM=blastn&PAGE_TYPE=BlastSearch&BLAST_SPEC=SRA&LINK_LOC=blastta](#)

b) . Esta plataforma possibilita buscar por sequências de interesse nos dados já depositados de sequenciamento de transcrito e genomas. Os dados de RNA-seq utilizados foram: *S. acidocalcarius* DSM639 (SRX327494), *Methanopyrus kandleri* AV19 (SRP019041), *Helicobacter pylori* 26695 (SRX014018, SRX014030 to SRX014033 and SRX014050) e *E. coli* K12 (SRX254762).

3.8) Análise da estrutura secundária de sotRNAs utilizando abordagens computacionais

As sequências dos ncRNAs de interesse foram utilizadas para predição de estrutura utilizando 5 diferentes métodos: RNAfold-MFE, que utiliza *Minimum Free Energy* (MFE); RNAfold, que utiliza termodinâmica e um algoritmo focado em centróides (GRUBER et al., 2008); ContextFold, que usa um grande número de parâmetros em uma abordagem de aprendizado de máquina (ZAKOV et al., 2011); Vsfold5, que permite a predição de pseudonós (DAWSON et al., 2006) e Cylofold que simula de maneira tridimensional o processo de dobramento de RNA (BINDEWALD; KLUTH; SHAPIRO, 2010).

3.9) Circularização de RNAs para identificação de extremidades 5' e 3'

Para propiciar a ligação intramolecular das extremidades 5' e 3', é necessário que a extremidade 5' esteja monofosfatada. Deste modo, uma alíquota de 8 µg de RNA livre de DNA foi tratada com 1 µl de *Tobacco Acid Pyrophosphatase* (TAP) (Epicentre), 3 µl de TAP Buffer 10X e volume ajustado para 30 µl com água MiliQ DEPC. O Tratamento foi realizado a 37° C durante 30', o RNA tratado foi purificado com o kit *MinElute* (QIAGEN) de acordo com instruções do fabricante.

Em seguida, o RNA purificado foi tratado com 4 µl de *T4 RNA ligase* (Thermo Scientific) em 10 µl de T4 Ligase Buffer 10X, 10 µl de BSA 10x, 10 µl ATP 10mM e água para 100 µl. A reação ocorreu a 17° C overnight em termociclador. A purificação do RNA foi realizada com kit *MinElute* (QIAGEN). A síntese de cDNA de RNAs circularizados foi obtida utilizando o kit *First Strand Synthesis* (Life Technologies), duas reações foram construídas, uma contendo hexâmeros randômicos e a outra oligonucleotídios específicos para o RNA alvo.

Após obtenção dos cDNAs, foi realizada uma reação de PCR utilizando oligonucleotídios divergentes. Estes oligonucleotídios têm como objetivo a amplificação da junção das

extremidades 5' e 3' do RNA (Figura 11). Os produtos de PCR foram corridos em gel de agarose 1% corado com *SYBR SAFE* (Life Technologies) e as bandas mais definidas foram excisadas e posteriormente purificadas. O processo de purificação foi realizado utilizando o kit *Illustra GFX PCR DNA and Gel Band Purification* (GE Healthcare) seguindo instruções do fabricante. O DNA purificado foi clonado em vetor *pGEM T-Easy* (Promega) de acordo com instruções do fabricante. Colônias brancas foram selecionadas e submetidas a PCR de colônia para confirmar a clonagem. Após a seleção das colônias positivas, um inóculo contendo 5 ml de meio LB e 2,5 µl de Carbenicilina 100 mg/ml foi incubado a 37° C, 225 RPM por 18 horas. Minipreps foram preparadas a partir dos inóculos utilizando *Wizard Plus SV Minipreps DNA Purification Kit* (Promega) de acordo com instruções do fabricante. Os vetores purificados foram sequenciados utilizando o *Big Dye Terminator v3.1* (Life Technologies). A corrida das amostras foi realizada em ABI 3500 XL localizado no Núcleo de Serviços em Biotecnologia vinculado ao Hemocentro de Ribeirão Preto.

Figura 11: Esquematização do processo de determinação de extremidades 5' e 3' através de circularização de RNAs.

3.10) Primer extension para determinação de extremidade 5' de RNAs

Primeiramente, foi feita uma reação de anelamento entre o oligonucleotídeo marcado com fluoróforos (VIC ou NED) (2 pmol) e RNA (10 µg) em volume total de 10 µl. A temperatura e tempo de anelamento podem variar de acordo com a abundância do RNA alvo e com a T_m do oligonucleotídeo. Em seguida, adicionou-se o mix de extensão da enzima SuperScript III (Invitrogen) (1 µl 0,1 M DTT, 2 µl SuperScript Buffer 10x, 1 µl 10 mM dNTP, 1 µl SuperScript 200 U/µl, 2,4 µl MgCl₂ 50 mM, Água DEPC para 10 µl). O tempo de extensão também variou de acordo com a abundância e tamanho do RNA alvo. As reações foram terminadas com a adição de 2 µl de EDTA 125mM, a precipitação foi obtida com a adição de 2 µl de NaOAC 3 M pH 5,2 e 50 µl de EtOH 100% gelado, seguindo para centrifugação a 2500G por 30 minutos. Após centrifugadas, o sobrenadante foi descartado e adicionou-se 70 µl de EtOH 70%, as amostras foram novamente centrifugadas a 2500G por 15 minutos. Por fim, o sobrenadante foi desprezado e a amostra seca em concentrador a vácuo.

A amostra de *primer extension* foi ressuspendida em 9,8 µl de HiDi Formamide (Life Technologies) e foi adicionado 0,2 µl de marcador de tamanho Liz-500. A corrida das amostras foi realizada em ABI 3500 XL localizado no Núcleo de Serviços em Biotecnologia vinculado ao Hemocentro de Ribeirão Preto.

Os parâmetros utilizados na corrida foram:

Capilar 50cm com POP 7, Voltagem: 15kV, Tempo de Injeção: 15s, Dye Set: G5, Size Standard Liz500, sendo que o kit de instalação é o DS-33 com Liz500, Size Standard v2.0 - 6-FAM, VIC, NED, PET dyes.

Os dados gerados após corrida em eletroforese capilar foram analisados utilizando o software GeneMapper (Applied Biosystems).

3.11) Detecção de homólogos do sotRNA VNG_sot0042 utilizando a ferramenta GlassGO

A ferramenta GLASSgo (GLobal Automated sRNA Search go) combina BLAST iterativo, filtros por identidade e agrupamento baseado em estrutura em uma pipeline automática para buscar por homólogos de sRNA no genoma de procariotos (WRIGHT et al., 2013a).

Os parâmetros utilizados para busca foram os padrões da ferramenta. O webserver encontra-se em <http://rna.informatik.uni-freiburg.de/GLASSgo/Input.jsp>.

3.12) Predição de mRNAs alvos de VNG_sot0042 utilizando a ferramenta CopraRNA

A ferramenta CopraRNA integra informações filogenéticas para prever alvos de sRNAs em escala genômica e identificar os domínios dos sRNAs responsáveis pelo reconhecimento e interação com seu alvo (WRIGHT et al., 2013b).

O primeiro passo da ferramenta é realizar uma predição em escala genômica para cada organismo considerado. As energias de interação são ajustadas para uma distribuição geral de valores extremos (distribuição de Gumbel) e transformadas em p-valor para normalizar de acordo com o conteúdo GC e dinucleotídico de cada organismo. Estes p-valores são combinados em um único valor para genes ortólogos. Os genes ortólogos são determinados através de suas respectivas sequências de aminoácido, genes que estão presentes em menos de 50% dos organismos estudados são descartados. Além disso, a distância das espécies baseadas nas sequências de 16S são calculadas e utilizadas para dar peso a predição de cada espécie.

Após isso, o *output* é representado por uma tabela contendo as informações do p-valor da predição, energia de interação, região de interação no mRNA, região de interação do sRNA e genes ortólogos encontrados. As predições que apresentarem p-valor < 0,01 são consideradas significativas.

3.13) Construção de linhagens de superexpressão

RNAs identificados por técnicas de transcriptoma tiveram suas extremidades 5' e 3' determinadas por experimentos de circularização de RNAs, alguns destes RNAs foram selecionados para clonagem em vetor de superexpressão. Os oligonucleotídeos utilizados apresentaram em seu 5' e 3' sítios de BamHI e NdeI respectivamente. Os produtos foram visualizados em gel de agarose 1% com SYBR SAFE (Life Technologies) e purificados com *GFX PCR DNA and Gel Extraction Kit* (GE Healthcare). Os fragmentos foram tratados com as enzimas BamHI e NdeI por 1 hora a 37° C e clonados em vetor pMTF previamente digerido com as mesmas enzimas. Este vetor contém o promotor forte Pdfx e é utilizado para superexpressão em *H. salinarum*. Células competentes de *E. coli* DH5a foram transformadas por choque térmico e plaqueadas em LB sólido contendo carbenicilina. Colônias foram selecionadas para verificação por PCR de colônia e as positivas foram inoculadas para preparação de miniprep. O DNA plasmidial foi extraído utilizando *Wizard Plus SV Miniprep Kit* (Promega) e sequenciado utilizando o kit BigDye Terminator v3.1 seguindo as orientações do fabricante. Os plasmídeos foram então transformados em *H. salinarum* NRC-1 como

descrito em (DYALL-SMITH, 2008). A seleção das células transformadas foi realizada em meio CM sólido suplementado com Mevinolina (20 µg/ml).

3.14) Análise de inícios de transcrição e sítios de processamento através de dRNA-seq

Os dados obtidos através do protocolo descrito abaixo foram obtidos pelo Dr. Felipe ten Caten e utilizados para análises complementares neste trabalho.

3.14.1) Depleção de RNAs processados

As amostras de RNA de cada réplica foram divididas em duas alíquotas. Uma delas foi tratada com 1U da enzima TEX (Terminator 5'-Phosphate-Dependent Exonuclease) (Epicentre TER51020) para cada 1µg de RNA (TEX+) enquanto a segunda alíquota foi incubada apenas na presença do tampão (TEX-). As reações foram incubadas a 30° C por 60 minutos e logo após foi realizada a purificação com o kit RNeasy MinElute Cleanup (Quiagen). Para verificar se houve degradação dos RNAs processados (rRNAs) 1µL das amostras TEX+ e TEX- foi analisada em eletroforese em gel de agarose 1% com tampão de corrida TAE. Após a confirmação da depleção dos RNA processados as amostras foram tratadas com a enzima TAP (Tobacco Acid Pyrophosphatase) (Epicentre T19100) na concentração de 1U para cada 1µg de RNA a 37 °C por 30 minutos e purificadas novamente com o kit RNeasy MinElute Cleanup (Quiagen).

3.14.3) Sequenciamento e processamento dos dados

As bibliotecas foram sequenciadas como descrito em Materiais e Métodos 3.18.

A ferramenta TSSAR, utilizada para identificação de inícios de transcrição, analisa o enriquecimento de contagem de início de reads na biblioteca TEX+ em comparação a biblioteca TEX-, levando em consideração o nível de expressão da região na qual a posição analisada se localiza (AMMAN et al., 2014). Essa comparação é realizada assumindo-se que a contagem de inícios de reads segue uma distribuição de Poisson com parâmetro λ definido pela média das contagens na região. Dado que a diferença entre duas distribuições de Poisson pode ser modelada de acordo com uma distribuição de Skellam (Skellam, 1946), todas as posições podem ser avaliadas verificando se a diferença de contagem entre as bibliotecas está ou não de acordo com a distribuição esperada para a região. Caso essa diferença seja maior do

que o esperado a posição é determinada como um TSS (Amman et al., 2014). Para definição de um início de transcrição (TSS) utilizamos valor de corte o p-valor de 0.001, um número mínimo de 10 reads por posição e o agrupamento de inícios de transcrição identificados a menos de 10 pares de base de distância.

Para análise de possíveis sítios de processamento, a mesma pipeline utilizada para identificação dos TSS foi utilizada, porém com as bibliotecas invertidas de modo que as amostras tratadas com TEX seriam o “controle” e as não-tratadas são o “tratamento”. Deste modo, podemos identificar sítios que apresentavam abundância de início de *reads* e após tratamento com TEX apresentaram depleção significativa no número de contagens. A ferramenta está disponível em <http://nibiru.tbi.univie.ac.at/TSSAR/>.

3.15) Ensaios de caracterização fenotípica

3.15.1) Curvas de Crescimento

As curvas de crescimento padrão foram realizadas em frascos de 500 ml contendo 200 ml de CM com suplementação de mevinolina (20 µg/ml) se necessário. A influência de diversos metais de transição foi testada em curvas de crescimento realizadas em placas de 96 poços, contendo 200 µl de cultura, 225 rpm de agitação e 37° C. Os metais e suas concentrações utilizadas foram: CoSO₄ 0,2 mM, 0,3 mM e 0,5 mM; CuSO₄ 0,4 mM, 0,8 mM, 1 mM; FeSO₄ 2 mM, 4mM, 5 mM; MnSO₄ 0.8 mM, 1 mM, 1,5 mM (Kaur et al, 2006). Todas as culturas iniciaram em OD600 0,05 e o crescimento das linhagens foi acompanhado por um período de 7 dias.

3.15.2) Ensaios de Sobrevivência

Para avaliar a viabilidade celular das linhagens selecionadas, após estresse drástico com metais pesados, foram realizados os ensaios de sobrevivência. As linhagens foram crescidas em condições referência até atingir OD600 0,5. Alíquotas de 4 mL de cada linhagem foram colocadas em tubos de ensaio de 10 mL e coletadas após 30 minutos, 1 hora, 3 horas, 6 horas e 27 horas após adição de sais de metal pesado: MnSO₄ 2 mM, FeSO₄ 9,8 mM e CoSO₄ 0,5 mM (Kaur et al, 2006). Para cada tempo, 20 µl foram coletados e utilizados para realização de diluições de 10⁻¹ a 10⁻⁴. Para cada diluição, uma alíquota de 5 µl foi utilizada para

plaqueamento em placas de CM suplementado com mevinolina. As placas foram incubadas por 7 dias a 37° C e fotografadas. A análise do número de unidades formadoras de colônia (UFC) foi realizada através do software ImageJ. A imagem foi convertida para preto e branco de modo que as UFCs estivessem em branco, após isso fizemos a contagem de pixels brancos na região de inóculo.

3.15.3) Ensaios de Adesão a Vidro

De modo a avaliar a capacidade de adesão ao vidro das linhagens selecionadas o seguinte método foi utilizado: pré-inóculos foram crescidos na condição referência até atingirem OD₆₀₀ 0,5, após esta etapa, 3 ml de cultura foram aliquotados em poços de placas de cultura de célula de 12 poços. Adicionou-se então uma lamínula de vidro (22 mm x 22 mm) dentro de cada poço, em ângulo de 90° e incubou-se por 24 horas. Após este período, as lamínulas foram retiradas e coradas da seguinte maneira: 1) 3 minutos em ácido acético 2%; 2) Remoção do excesso de solução em lenço de papel; 3) Secagem ao ar e imersão em solução de cristal violeta 0,1% p/v por 10 minutos; 4) Lavagem das lamínulas em H₂O e secagem ao ar (ESQUIVEL; XU; POHLSCHRODER, 2013).

Para cada lamínula 6 campos aleatórios foram fotografados utilizando microscópio óptico (Olympus BX50 equipado com RT3 slider 2MP digital camera) (40x). O software ImageJ foi utilizado para processamento das imagens e contagem das células (SCHINDELIN et al., 2015). Um macro dos comandos utilizados está disponível em: <https://www.dropbox.com/s/tns6i6b67yrdaa1/contagem.ijm?dl=0>.

3.16) Estudos de interação proteína-RNA utilizando SILAC

Estabelecemos colaboração com o grupo do Dr. Falk Butter do Institute for Molecular Biology, Mainz, Alemanha para a realização da técnica de SILAC (*stable isotope labeling by amino acids in cell culture*) (BUTTER et al., 2009). Esta técnica consiste em crescer o organismo de interesse em dois meios de cultura diferentes, um contendo ¹³C₆-Lys e ¹³C₆-Arg (*light*) e o outro contendo ¹⁴C₆-Lys e ¹⁴C₆-Arg (*heavy*). O RNA a ser testado é transcrito *in vitro* fusionado ao aptâmero S1, o que permite sua ligação a *beads* magnéticos de estreptavidina, este RNA é então ligado a uma coluna. Um passo importante é a construção de um RNA idêntico ao que será testado, sendo sua única diferença a deleção do motivo

estrutural que possa ser o responsável por estabelecer as interações proteína-RNA. Os RNAs sintetizados são então ligados a duas colunas diferentes e por essas colunas os extratos proteicos obtidos das culturas *light* e *heavy*, crescidas em condições padrão, são filtrados. Os extratos proteicos foram obtidos através da lise de *H. salinarum* NRC-1 em tampão Tris contendo 4.3M de NaCl. Um tampão com a mesma concentração de sal foi utilizado na etapa de ligação. Posterior análise por espectrometria de massa revelará quais são os parceiros de interação únicos a cada RNA testado (Figura 12). A detecção das proteínas enriquecidas encontra-se descrita em (BUTTER et al., 2009).

Figura 12: Pull-down de RNAs baseado em SILAC. As células são crescidas em meio contendo ou $^{13}\text{C}_6$ -Lys e $^{13}\text{C}_6$ -Arg (*light*) ou $^{14}\text{C}_6$ -Lys e $^{14}\text{C}_6$ -Arg (*heavy*). O RNA é imobilizado em beads de estreptavidina através do aptâmero S1 o qual é introduzido através de PCR e posterior transcrição in vitro. Os extratos proteicos *heavy* e *light* são incubados com a sequência alvo e controle, separadamente. Os complexos de interação são então purificados e identificados por espectrometria de massa quantitativa. Adaptado de Butter et al, 2009.

Para a transcrição *in vitro* dos RNAs testados, as sequências destes RNAs foram fusionadas a sequência do promotor T7 na região 5' e a sequência do aptâmero S1 na região 3'. Os controles escolhidos foram duas variações do sotRNA VNG_sot0042: 1) A sequência do RNA porém com o motivo UUCA deletado, VNG_sot0042dm; 2) Uma sequência aleatória com mesmo conteúdo GC e tamanho, chamada de *random206*. A construção dos RNAs

utilizados neste experimento foi realizada combinando técnicas de PCR *overlapping* e hibridização de oligonucleotídeos.

3.17) Preparação de bibliotecas de RNA-seq e sequenciamento

As amostras utilizadas na preparação das bibliotecas de RNA-seq foram quantificadas com o kit *Quant-iT RiboGreen RNA Assay* (Invitrogen). Por se tratar de uma quantificação através de uma marcação por fluorescência específica para moléculas de RNA essa metodologia minimiza a interferência por nucleotídeos livres ou contaminantes como proteínas, fornecendo uma quantificação mais precisa. Para preparação das bibliotecas utilizamos o kit *TruSeq Small RNA Sample Preparation* (Illumina) com uma quantidade inicial de 1µg de RNA. A primeira etapa consiste na ligação de adaptadores 3' e 5' às moléculas de RNA, seguida de uma etapa de transcrição reversa para converter as moléculas em cDNA e uma amplificação por PCR que adiciona um index único a cada biblioteca, além de aumentar o número de cópias das moléculas presentes na reação. A etapa de extensão das moléculas na amplificação por PCR foi aumentada de 15 segundos para 1 minuto. Essa modificação foi introduzida para que conseguíssemos amplificar moléculas maiores presentes na reação, e não apenas os pequenos RNAs. Em seguida as amostras foram corridas em gel 6% *Novex TBE PAGE* (Invitrogen) e a porção logo acima do marcador de 145 pb e abaixo do marcador de 500 pb foi excisada para purificação, selecionamos os fragmentos até o tamanho de 600 pb por se tratar do tamanho máximo em que as moléculas são capazes de formar ligações em bridge para obtenção dos *clusters* no *flowcell*. Cada porção do gel correspondente as amostras foi fragmentado independentemente e os fragmentos foram eluídos em H₂O MilliQ *overnight*.

Ao final as amostras eluídas foram purificadas em tubos com filtros de 5 µm. Para assegurar que as bibliotecas fossem sequenciadas em proporções semelhantes foi realizada uma quantificação final por qPCR utilizando o kit *KAPA Library Quant* (Kapa Biosystems), o qual possui primers específicos para amplificação das moléculas ligadas aos adaptadores. Dessa forma, podemos aferir a quantidade exata de moléculas aptas para sequenciamento e ajustar o volume necessário de cada biblioteca para que obtenhamos quantidades similares de sequências ao final do processo.

3.18) Sequenciamento e processamento dos dados de RNA-seq

As bibliotecas foram sequenciadas com o kit *MiSeq Reagent v2 300 cycles Paired-end* (Illumina) no sequenciador MiSeq. Os *reads* obtidos foram processados para remoção dos adaptadores utilizando a ferramenta FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) e alinhados ao genoma referência de *H. salinarum* NRC-1 (NC_002607, NC_002608, NC_001869) com o software *Bowtie* (Langmead et al., 2009) com o parâmetro -m igual a 1 para desconsiderar qualquer *read* que alinhe em mais de uma posição, os demais parâmetros foram mantidos no padrão.

Para identificar os genes diferencialmente expressos, duas etapas foram necessárias. A primeira foi realizada utilizando o software HTSeq (ANDERS; PYL; HUBER, 2015). Este software utiliza, em conjunto, os alinhamentos gerados pelo Bowtie, em formato SAM, e uma tabela de *features* anotadas no genoma do organismo interesse no formato GFF. Com estes dados, o software faz uma contagem de *reads* sobrepostos em cada *feature*, sem realizar normalizações, no fim teremos uma matriz de contagens para cada *feature*. O cálculo de expressão diferencial foi realizado utilizando o pacote DESeq2, disponível na linguagem R. Nessa análise foram considerados os genes que apresentaram *FoldChange* > 1,5 e p-value < 0,05 ou *FoldChange* < -1,5 e p-value < 0,05 (LOVE; HUBER; ANDERS, 2014).

Utilizamos a ferramenta KEGG *orthology* (KO) para associação dos genes às suas possíveis funções (AOKI-KINOSHITA; KANEHISA, 2007). Uma lista de genes diferencialmente expressos encontrados através de DESeq2 foi gerada e os nomes dos genes convertidos para os identificadores utilizados no KEGG (e.g. VNG2652H → VNG_2652H). Esta lista foi submetida para análise utilizando o webserver KEGG mapper (http://www.genome.jp/kegg/tool/map_brite1.html). Na ausência de um KO associado, dados provenientes do banco MicrobesOnline (<http://www.microbesonline.org/>) e a utilização da ferramenta BlastCD (MARCHLER-BAUER et al., 2016) para identificar domínios proteicos conservados, foram utilizados para associar funções.

3.19) Cálculo de expressão diferencial de VNG_sot0042

De maneira a avaliar a expressão de VNG_sot0042 na linhagem de superexpressão utilizamos a sequência do vetor de superexpressão PMTF-VNG_sot0042 como índice. A utilização do vetor de expressão como índice é importante pois uma fração dos *reads*

relacionados ao VNG_sot0042 não são alinhados ao genoma referência, devido a presença de sequências relacionadas ao vetor, como uma porção do promotor Pfdx e sítios de enzimas de restrição. Deste modo, podemos discernir quais transcritos são advindos do cromosso e quais são do plasmídeo de superexpressão. Para o cálculo de expressão relativa, primeiramente calculamos um fator de correção para todas as bibliotecas em relação a biblioteca de maior contagem (pMTF0-B). Após isso, fizemos a contagem dos reads alinhados a região de VNG_sot0042 e mais uma porção do vetor. As contagens foram normalizadas dividindo o número de reads alinhados em VNG_sot0042 pelo tamanho da região analisada e multiplicando pelo fator de correção de cada biblioteca. Calculamos a média das contagens normalizadas de cada biblioteca e então dividimos a média de pMTF0042 pela média de PMTF-0. Deste modo, verificamos que o sotRNA VNG_sot0042 se encontra duas vezes mais expresso na linhagem de superexpressão.

4) Resultados e Discussão

Nesta seção, serão apresentadas a identificação e caracterização funcional de uma nova classe de RNAs sobrepostos ao 3' de genes *tnpB* em *H. salinarum* NRC-1, chamados de *sense overlapping transcripts* (sotRNAs). Serão discutidas as abordagens utilizadas para identificação destes RNAs em dados de transcriptoma previamente publicados, obtidos em diversas condições genéticas e ambientais. Além disso, abordaremos a conservação deste transcrito em procariotos assim como suas características de estrutura secundária e sequência. A seguir, serão apresentados os esforços realizados para iniciar a caracterização funcional através da construção de linhagens de superexpressão e curvas de crescimento. Por fim, a avaliação do impacto da superexpressão deste RNA no transcriptoma de *H. salinarum* NRC-1 será discutido através da análise de dados de RNA-seq e de interação proteína-RNA (SILAC), complementados com novos ensaios fenotípicos.

4.1) *Sense overlapping RNAs*: Identificação, conservação e mapeamento de motivo estrutural

4.1.1) Identificação de transcritos sobrepostos ao 3' de *tnpB* de IS200/605 em dados de transcriptoma de *H. salinarum* NRC-1

Comparando as anotações de *H. salinarum* R1 e *H. salinarum* NRC-1 em conjunto com anotações provenientes do ISfinder identificamos 10 genes *tnpB* de IS200/605 ao longo do genoma de *H. salinarum* NRC-1 (Tabela 4). Destes, 5 encontram-se associados ao gene *tnpA*, 4 em sentido convergente e 1 em sentido divergente, formando portanto elementos IS605. Os outros genes, por se encontrarem sozinhos são classificados como IS1341. É interessante notar que não foram encontrados genes *tnpA* sozinhos, evidenciando a ausência de IS200 em *H. salinarum* NRC-1.

Tabela 4: IS200/605 presentes em *H. salinarum* NRC-1.

Gene	Classificação dentro de IS200/605	Replicon	<i>tnpB</i> start codon	<i>tnpB</i> stop codon	Fita
VNG_RS00050	IS1341	chr	12734	11478	-
VNG_RS00105	IS1341	chr	21789	20533	-
VNG_RS00160	IS1341	chr	35931	34651	-
VNG_RS00170	IS605 – convergente	chr	36715	37971	+
VNG_RS01145	IS605 – convergente	chr	229443	230597	+
VNG_RS10400	IS605 – divergente	chr	1989188	1987884	-
VNG_RS11205	IS605 - convergente	pNRC200	155603	154368	-
VNG_RS11320	IS605 - convergente	pNRC200	183054	181795	-
VNG_RS11800	IS1341	pNRC200	287504	288844	+
VNG_RS11945	IS1341	pNRC200	316931	315786	-

Através da análise de dados públicos de *tiling array*, obtidos ao longo de uma curva de crescimento de *H. salinarum* NRC-1 (KOIDE et al., 2009), encontramos uma assinatura transcricional comum a todos os genes *tnpB*. Esta assinatura é caracterizada por uma mudança brusca no sinal de expressão próximo a extremidade 3' de *tnpB*, indicando a possível existência de transcritos sobrepostos (*sense overlapping transcripts* - sotRNAs) nesta região (Figura 13). A (Figura 13A-I) ilustra o perfil transcricional encontrado para o gene VNG_RS00160 e seu sotRNA VNG_sot0042 na condição referência. Ao analisarmos a dinâmica de expressão destes transcritos ao longo da curva de crescimento, verificamos que os sotRNAs apresentam aumento nos níveis de expressão relativa, enquanto *tnpB* apresenta um perfil contrário (Figura 13A-II). A utilização de dados de *tiling array* obtidos de linhagens superexpressando o fator de transcrição TfbD evidenciaram um perfil de expressão oposto ao visualizado na curva de crescimento. Neste caso, o sotRNA apresenta menor expressão relativa enquanto *tnpB* apresenta maior expressão relativa (Figura 13A-III). Deste modo, a presença dessas assinaturas transcricionais associadas a informação dinâmica dos *tiling microarrays* reforçam a existência dos sotRNAs no 3' de *tnpB* em *H. salinarum* NRC-1.

Figura 13: Identificação de sotRNAs em *tnpB* de *H. salinarum* NRC-1 utilizando dados de *tiling array* (Koide et al, 2009). (A) (I) Sinal de *tiling array* na condição referência para VNG_RS00160 e VNG_sot0042. (II) perfil de expressão durante uma curva de crescimento (III) perfil de expressão em linhagem superexpressando o fator de transcrição TfbD. (B) Sinal de *tiling array* na condição referência e perfil de expressão dos outros genes *tnpB*. Genes estão indicados em amarelo (fita +) e laranja (fita -) e sotRNAs em azul.

A análise de enriquecimento de 5' de *reads* e inspeção manual da região 3' obtidos através de RNA-seq de pequenos RNAs (sRNA-seq) previamente publicados (ZARAMELA et al., 2014), possibilitaram determinar as extremidades dos sotRNAs, como exemplificado para VNG_sot0042 na (Figura 14). As posições dos outros sotRNAs encontram-se na (Tabela 5).

Figura 14: Bibliotecas de RNA-seq de pequenos RNAs (sRNA-seq) foram utilizadas para mapear o 5' de sotRNAs em *tnpB* de *H. salinarum* NRC-1. Os dados de sRNA-seq são visualizados como $\log_2$ do número de reads alinhados em cada posição genômica para VNG_RS00160 e VNG_sot0042. O enriquecimento do 5' dos reads é visualizado como picos abaixo dos dados de sRNA-seq (*start counts*). Seta azul: VNG_sot0042. Seta laranja: VNG_RS00160.

Tabela 5: sotRNAs identificados no 3' de *tnpB* de IS200/605 em *H. salinarum* NRC-1.

sotRNA	Replicon	Fita	5' sotRNA	3' sotRNA
VNG_sot0013	chr	-	11641	11428
VNG_sot0026	chr	-	20701	20498
VNG_sot0042	chr	-	34775	34570
VNG_sot0044	chr	+	37906	38027
VNG_sot0286	chr	+	230444	230629
VNG_sot2652	chr	-	1988005	1987819
VNG_sot6181	pNRC200	-	154508	154259
VNG_sot6221	pNRC200	-	182008	181724
VNG_sot6361	pNRC200	+	288696	288915
VNG_sot6406	pNRC200	-	315833	315631

Para validar as posições obtidas através das análises de sRNA-seq, realizamos a

circularização de dois sotRNAs, VNG_sot0042 e VNG_sot0044 (Figura 15), seguida de amplificação com oligos divergentes e sequenciamento. As extremidades encontradas para estes sotRNAs foram as mesmas determinadas previamente pela análise de sRNA-seq.

Figura 15: Determinação das extremidades 5' e 3' dos sotRNAs VNG_sot0042 e VNG_sot0044 através da técnica de C-RACE. Os produtos de PCR provenientes da amplificação de cDNA obtido a partir de RNAs circularizados *in vitro* foram sequenciados. Destacados em vermelho estão a região 5' e em azul a região 3'.

Para melhor entender os perfis de expressão de *tnpB* e seus sotRNAs, utilizamos em conjunto diversos dados de *tiling micorarray* previamente publicados (Materiais e Métodos 3.4). Todos os sotRNAs demonstraram diversos padrões de regulação, como esperado para um típico transcrito funcional. A maioria dos sotRNAs apresenta perfis de expressão diferentes de seu gene cognato nas diversas condições analisadas. A (Figura 16) demonstra este perfil para VNG_RS00160 e seu sotRNA VNG_sot0042. Deste modo, os perfis de expressão opostos dos sotRNAs e seus genes cognatos em diversas condições genéticas e ambientais, sugerem a existência de algum tipo de regulação entre estes elementos.

Figura 16: Perfis de expressão de VNG_sot0042 e VNG_RS00160 em diferentes condições genéticas e ambientais. As condições representadas são: A) Curva de Crescimento; B) Superexpressão do fator de transcrição TfbD; C) Nocaute do gene URA3; D) Nocaute da RNase VNG_RS08130 ; E) Concentrações diferentes de salinidade; F) Nocaute de diversos fatores de transcrição e superexpressão de fatores de transcrição sintéticos. (BEER et al., 2014; KOIDE et al., 2009; TURKARSLAN et al., 2011; WURTMANN et al., 2014).

Dos 10 sotRNAs identificados, apenas VNG_sot0286 se encontra antisenso a um gene já anotado. Este sotRNA está localizado antisenso ao gene VNG_RS01150 (DNase putativa), cobrindo uma porção interna do gene e sua região 3' UTR sugerindo uma possível interação *cis*-antisenso. Ao analisarmos o perfil de expressão de ambos VNG_sot0286 e VNG_RS01150 nos dados de *tiling array*, pudemos visualizar que em determinadas condições, como nos estágio iniciais da curva de crescimento e na superexpressão do fator de transcrição TfbD, os perfis de expressão são opostos, reforçando a hipótese proposta de regulação antisenso (Figura 17).

Figura 17: Perfis de expressão de VNG_sot0286 e VNG_RS01150 em diferentes condições genéticas e ambientais. As condições representadas são: A) Curva de Crescimento; B) Superexpressão do fator de transcrição TfbD; C) Nocaute do gene URA3; D) Nocaute da RNase VNG_RS08130; E) Concentrações diferentes de salinidade; F) Nocaute de diversos fatores de transcrição e superexpressão de fatores de transcrição sintéticos. (BEER et al., 2014; KOIDE et al., 2009; TURKARSLAN et al., 2011; WURTMANN et al., 2014).

Para nossa surpresa, a presença de sotRNAs também foi verificada em IS200/605 incompletas, especificamente em versões truncadas de *tnpB*. No pseudogene homólogo ao gene *OE5220R* de *H. salinarum* R1 este fenômeno pode ser verificado (Figura 18). Os pseudogenes por muito tempo foram tratados como formas inativas de genes funcionais, originados através de processos de transposição ou inativação da cópia após evento de duplicação do gene (AKAMA et al., 2009). Em humanos, dados de transcrito demonstraram que uma parcela significativa dos pseudogenes são expressos e produzem RNAs estáveis com potencial regulatório (GUO et al., 2014). Além disso, pseudogenes são possíveis fonte de RNAs antisense com atuação em *trans* (MURO; ANDRADE-NAVARRO, 2010). Deste modo, a presença de um sotRNA nesta forma inativa de *tnpB* pode sugerir que a preservação destes transcritos no genoma é independente da manutenção de *tnpB* intacto.

Figura 18: Região de *H. salinarum* NRC-1 contendo o pseudogene homólogo a *OE5220R* de *H. salinarum* R1. O destaque em azul demonstra a região contendo o pseudogene. Os pontos azuis representam dados de sRNA-seq e demonstram uma maior quantidade de *reads* na região 3' do pseudogene.

4.1.2) Biogênese de sotRNAs, produtos de processamento ou transcritos primários?

A biogênese de RNAs sobrepostos na mesma fita pode ocorrer através da transcrição a partir de promotores internos (BILUSIC et al., 2014) ou através do processamento de RNAs (GEBERT; ROSENKRANZ, 2015). Para elucidar a possível biogênese dos sotRNAs identificados, integramos os seguintes dados: 1) Dados de *differential* RNA-seq (dRNA-seq), para identificar inícios de transcrição primários (TSS); 2) Dados de interação proteína-DNA (ChIP-chip); 3) Análise de regiões promotoras.

Os dados de dRNA-seq foram obtidos durante uma curva de crescimento de *H. salinarum* NRC-1 pelo Dr. Felipe ten Caten. O dRNA-seq consiste no enriquecimento de RNAs primários através do tratamento com a enzima 5' terminator exonuclease (TEX), que degrada RNAs apresentando 5' monofosfato, ou seja, transcritos processados (Figura 19) (SHARMA et al., 2010).

Figura 19: dRNA-seq. A técnica de dRNA-seq consiste no tratamento da amostra de RNA com a enzima 5' Terminator Exonuclease (TEX). Essa enzima degrada RNAs que apresentam apenas um grupo fosfato na extremidade 5', ou seja, transcritos que tenham passado por alguma etapa de processamento. Com isso a amostra é enriquecida para RNAs primários. A análise do sequenciamento desses RNAs, em conjunto com dados de uma biblioteca controle sem enriquecimento, permite a identificação das Liberação Serifposições de início de transcrição (TSS). Adaptado de Sharma et al , 2010.

A análise do perfil de interação proteína-DNA de todos os *transcription factor B* (TFBs), quatro *tata binding proteins* (TBP) e três *transcriptional regulators* (TR) foi realizada através de dados de *chromatin immunoprecipitation* (ChIP-chip) previamente publicados (FACCIOTTI et al., 2007; KOIDE et al., 2009; PLAISIER et al., 2014). A busca por regiões

promotoras foi feita utilizando como base sequências clássicas de BRE-TATA localizadas *upstream* aos sotRNAs.

A inspeção manual dos dados de dRNA-seq demonstram enriquecimento de transcritos primários apenas para VNG_sot0013 e VNG_sot0026 (Tabela 6). Em *T. kodakarensis*, dados de dRNA-seq também demonstram que alguns sotRNAs são transcritos primários, porém este não é fenômeno geral para estes RNAs (JÄGER et al., 2014). No caso dos sotRNAs com TSS identificados em *H. salinarum* NRC-1, algumas ressalvas devem ser feitas. A região *upstream* a base determinada como início de transcrição é uma região repetida e portanto apresenta zero contagens utilizando alinhamento padrão do Bowtie. A identificação de TSS através da ferramenta TSSar (AMMAN et al., 2014) depende diretamente do número de contagens presentes nas vizinhanças da região analisada. Deste modo a região de zero contagens pode ser um fator de enviesamento e geração de falso positivo. Além da análise convencional de dRNA-seq, uma análise complementar utilizando os mesmos dados para buscar possíveis sítios de processamento foi realizada. Para esta análise utilizamos a ferramenta TSSar com algumas mudanças na submissão dos dados. As bibliotecas sem tratamento com TEX foram submetidas como tratamento e a tratada com TEX submetidas como controle. Deste modo, poderíamos visualizar sítios onde o início de *reads* era abundante e após tratamento com TEX foi depletado indicando possíveis posições de processamento. No caso de VNG_sot0013 e VNG_sot0026, que apresentavam pelo menos um TSS identificado, encontramos possíveis sítios de processamento muito próximos as posições de seus TSS. Para VNG_sot0013, por exemplo, o TSS identificado encontra-se na posição 11646, enquanto os sítios de processamento mais próximos se encontram nas posições 11642 e 11645. Com exceção dos sotRNAs VNG_sot6361, VNG_sot6406 e VNG_sot6221, encontramos possíveis sítios de processamento na posição previamente determinada por sRNA-seq/C-RACE ou em posições muito próximas (até 10 nucleotídeos *upstream* ou *downstream*). Outro fator interessante é que a posição de 5' determinada por C-RACE para os sotRNAs VNG_sot0042 e VNG_sot0044 é exatamente a mesma posição que se mostra enriquecida para possíveis sítios de processamento (Figura 20).

Figura 20: Dados de dRNA-seq apresentando inícios de transcrição e possíveis sítios de processamento para VNG_RS00160 e seu sotRNA VNG_sot0042. Os picos em vermelho representam os reads da biblioteca tratada com TEX, e portanto depletada para RNAs processados. Os picos em preto representam a biblioteca não-tratada com TEX, ou seja toda a população de RNAs. As setas verdes representam os inícios de transcrição identificados através da ferramenta TSSAR e sua direção em relação ao gene VNG_RS00160 (AMMAN et al., 2014). Os triângulos vermelhos representam sítios de processamento identificados através da ferramenta TSSAR com alguns parâmetros modificados (Materiais e métodos 3.14). A caixa amarela representa o gene *tnpB* e a caixa azul representa seu sotRNA correspondente.

Em relação aos promotores BRE-TATA, nenhum dos sotRNAs demonstrou a presença destas sequências em regiões *upstream* ao seu 5'. A análise dos dados de ChIP-chip demonstraram que apenas dois sotRNAs apresentam sítios de ligação de fatores de transcrição em regiões de até 50 pb *upstream*, sendo TbpB para VNG_sot2652 e Trh3 para VNG_sot6221 (Tabela 6).

Tabela 6: Associação dos dados de ChIP-chip, dRNA-seq e 5' determinados pela análise de sRNA-seq. Para cada sotRNA foram avaliados os fatores de transcrição encontrados em regiões de até 50 pb upstream de seu 5'. A distância entre o sítio de ligação dos fatores de transcrição e o 5' dos sotRNAs foi calculado tomando como base o 5' identificado pelos dados de sRNA-seq e C-RACE. Os possíveis TSS e sítios de processamento também foram avaliados. * - 5' validados por C-RACE.

sotRNA	Proteína	Região de interação (ChIP-chip)	5'	TSS identificado Por dRNA-seq	Possíveis sítios de processamento	Distância entre o sítio de ligação E o 5' do sotRNA
VNG_sot0013	-	-	11646	11646	11642 11645 11654	-
VNG_sot0026	-	-	20701	20701	20697 20707	-
VNG_sot0042	-	-	*34775	-	34775	-
VNG_sot0044	-	-	*37813	-	37813	-
VNG_sot0286	-	-	230444	-	230444	-
VNG_sot2652	TbpB	1988042	1988005	-	1988005	-32
VNG_sot6181	-	-	154508	-	154508	-
VNG_sot6221	Tth3	182015	182008	-	-	-7
VNG_sot6361	-	-	288696	-	288725	-
VNG_sot6406	-	-	315833	-	-	-

Além disso, experimentos de *northern blot* dos sotRNAs VNG_sot0042 e VNG_sot2652 (Figura 21) evidenciaram um perfil interessante. Os sotRNAs (~ 200) foram detectados em ambos os casos. Por outro lado, a presença de bandas maiores do que os sotRNAs, porém não correspondentes ao transcrito completo de *tnpB* (esperado ~1.200 nts), sugerem a presença de possíveis intermediários de processamento.

Figura 21: Experimentos de *northern blot* realizados para os sotRNAs VNG_sot0042 (A) e VNG_sot2652 (B). As setas vermelhas representam as sondas utilizadas no experimento. Além das bandas referentes aos sotRNAs (~ 200 nt), bandas maiores, porém não correspondentes ao transcrito completo de *tnpB* foram encontradas, sendo assim possíveis intermediários de processamento.

Deste modo, a presença de TSS para os sotRNAs VNG_sot0013 e VNG_sot0026 sugerem que eles são produtos de transcrição primária. A ausência de TSS identificados por dRNA-seq para os outros 8 sotRNAs, em conjunto com a ausência de sequências BRE-TATA, ligação de fatores de transcrição e presença de possíveis sítios de processamento e o perfil de bandas detectado por *northern blot*, sugerem que estes RNAs são produzidos pela maquinaria de processamento de RNAs de *H. salinarum* NRC-1, provavelmente através da ação de endonucleases.

4.1.3) Identificação de sotRNAs em outros organismos

Utilizando dados de transcrito previamente publicados em conjunto da ferramenta GlassGo, a qual integra a realização de BLAST iterativo e *clustering* baseado na predição de estruturas para encontrar homólogos de sRNAs (WRIGHT et al., 2013a), verificamos que estes transcritos sobrepostos ao 3' de *tnpB* são mais profusos do que previamente esperado. Através da busca de sinais de transcrição similares aos visualizados em *H. salinarum* NRC-1, fomos capazes de identificar sotRNAs nas Archaea *Methanopyrus kandleri* AV19, *Sulfolobus acidocaldarius* DSM639 e *Pyrococcus furiosus* DSM3638. Em bactérias, a presença de sotRNAs foi identificada em *E. coli* K12 e *H. pylori* MG26695 (Figura 22).

Em paralelo, realizamos uma outra análise complementar utilizando a ferramenta GlassGO para buscar homólogos dos sotRNAs, identificados em *H. salinarum* NRC-1, em todos organismos presentes na base RefSeq do NCBI. Os homólogos identificados, são prevalentemente encontrados em Haloarchaea (ANEXO A). Deste modo, podemos sugerir que sotRNAs são elementos conservados nos genomas de procariotos.

Figura 22: Identificação de sotRNAs em outros organismos. A) Dados de *tiling array* de *Pyrococcus furiosus*. As linhas azuis e vermelhas indicam, respectivamente, 5' e 3' do sotRNA encontrado no gene PF2024 (*tnpB*). B) Dados de RNA-seq de *Sulfolobus acidocaldarius*. Os traços verdes representam o acúmulo de *reads* alinhados a região 3' do gene Saci_0487 (*tnpB*). As setas amarelas representam genes anotados na fita *forward*.

4.1.4) Identificação de motivo conservado RE-like em sotRNAs e PATE-RNAs

Analisando as sequências dos sotRNAs presentes em *H. salinarum* NRC-1, pudemos identificar um motivo de estrutura secundária, um *hairpin*, extremamente estável. Utilizando 5 métodos diferentes de predição de estrutura secundária, encontramos as seguintes características: 1) um motivo *tetraloop* contendo uma base (*stem*) de 8 pares de base e um *loop* de 4 bases (UUCA) nos sotRNAs VNG_sot0042, VNG_sot0044, VNG_sot6181 e VNG_sot6221; 2) variantes contendo diferença em apenas um nucleotídeo, VNG_sot0286 (UUUA), VNG_sot0013 (CUCA) e VNG_sot0026 (GUCA); 3) Outros *hairpins* originados de sequências diferentes em VNG_sot2652, VNG_sot6406 e VNG_sot6361 (Figura 23).

Figura 23: Motivo *tetraloop* encontrado na maioria dos sotRNAs em *H. salinarum* NRC-1. As cores representam os grupos de pareamento e a saturação corresponde a conservação, parênteses indicam pareamento e pontos indicam as regiões não-pareadas. Blocos cinzas abaixo da sequência representam os nucleotídeos mais conservados.

É importante ressaltar que os algoritmos utilizados para predição de estrutura secundária apresentam grandes diferenças em sua metodologia, reforçando a significância destes resultados. Uma das características principais de elementos IS200/605 é a presença de palíndromos imperfeitos localizados tanto em seu *left end* (LE) e *right end* (RE) (Figura 6) que formam *hairpins* a nível de DNA. A proteína TnpA, responsável pela transposição de IS200/605, liga-se aos motivos RE e LE, no DNA, para realizar o processo de excisão de fita (HE et al., 2013). Todos os sotRNAs apresentam-se sobrepostos as regiões que contém o motivo RE, deste modo o motivo UUCA *tetraloop* identificado, foi chamado de motivo RE-like. A presença do motivo RE-like nos sotRNAs gerou a hipótese de que estes RNAs ao se

estruturarem podem atuar como "iscas" da proteína TnpA, evitando sua ligação ao DNA e consequentemente o processo de transposição. Este caso reforçaria a hipótese de manutenção de *tnpB* nos genomas hospedeiros como um mecanismo adicional no controle de transposição.

Além disso, o alinhamento de dois ncRNAs intergênicos, sRK48 e sRK52, de *P. abyssi* e outras Archaea termófilas, demonstraram a presença de uma estrutura estável e conservada chamada de *loop* P1 (PHOK et al., 2011), este *loop* corresponde ao motivo RE-like aqui descritos, reforçando a profusão deste motivo em Archaea.

Figura 24: Bibliotecas de RNA-seq de pequenos RNAs (sRNA-seq) foram utilizadas para mapear o 5' de novos ncRNAs intergênicos contendo o motivo RE-like em *H. salinarum* NRC-1. Os dados de sRNA-seq são visualizados como $\log_2$ do número de reads alinhados em cada posição genômica para VNG_R0052. O enriquecimento do 5' dos reads é visualizado como picos abaixo dos dados de sRNA-seq (start counts). Seta azul: ncRNA identificado.

Utilizando a sequência consenso do motivo RE-like, realizamos uma busca por similaridade no genoma de *H. salinarum* NRC-1. Além dos sotRNAs, pudemos identificar três regiões intergênicas contendo este motivo. Das três, apenas duas apresentaram cobertura suficiente nos dados de RNA-seq para serem estudadas mais a fundo. Através da análise do enriquecimento do 5' de *reads* em dados de sRNA-seq, pudemos anotar dois novos RNAs intergênicos nestas regiões, VNG_R0052 e VNG_R6334 (Figura 24) (Tabela 7).

Tabela 7: ncRNAs identificados através da busca por regiões intergênicas contendo o motivo RE-like.

ncRNA	5'	3'	Replicon	Fita
VNG_R0052	14397	14299	chr	-
VNG_R6334	268030	268304	pNRC200	+

Análises dos dados de *tiling array* evidenciaram que estes RNAs são diferencialmente regulados e apresentam um perfil de expressão similar ao de sotRNAs (Figura 25).

Figura 25: Perfis de expressão de VNG_R0052 em diversas condições ambientais e genéticas. A) Curva de Crescimento; B) Superexpressão do fator de transcrição TfbD; C) Nocaute do gene URA3; D) Nocaute da RNase VNG_RS08130; E) Concentrações diferentes de salinidade; F) Nocaute de diversos fatores de transcrição e expressão de fatores de transcrição sintéticos. (BEER et al., 2014; KOIDE et al., 2009; TURKARSLAN et al., 2011; WURTMANN et al., 2014).

Em relação ao ncRNA VNG_R0052, além do motivo RE-like pudemos encontrar parte da sequência referente ao LE de IS200/605, fusionada na região 5' deste RNA. Curiosamente, no *locus* onde este RNA se encontra expresso encontra-se um PATE. Esta classe de elementos de transposição não-autônomos, encontrada através da comparação do genoma de duas linhagens diferentes de *Haloquadratum walsbyi* é caracterizada por se originar através de eventos de deleção da região central de IS200/605 (Figura 4) (DYALL-SMITH et al., 2011). Além disso, elementos similares a PATEs, chamados de REPINS (REP doublets forming hairpins) foram encontrados em bactérias (BERTELS; RAINEY, 2011). Estes elementos, quando inseridos em regiões próximas (~15 nts) ao 3' de genes codificantes, podem ser co-transcritos com o mRNA

do gene cognato e atuar como importantes reguladores de tradução (LIANG; DEUTSCHER, 2016). É interessante ressaltar que apesar da associação de ncRNAs com outros derivados não-autônomos, como MITEs, já ter sido descrita, este é o primeiro caso descrito para um PATE. De acordo com a semelhança de PATEs com REPINs, podemos sugerir que estes elementos também possam ser transcritos de maneira independente e serem importantes fontes de ncRNAs de papel regulatório.

Outro fator interessante é que o motivo RE-like está presente em uma família de RNAs chamada HgcC, depositada no banco de dados RFAM (RF00062) (Figura 26). A família HgcC foi inicialmente descoberta em *Methanococcus jannaschii* através de análises *in silico* focadas na identificação de regiões de alto conteúdo G+C, outros membros desta família foram identificados em Archaea, porém suas funções ainda são desconhecidas. Deste modo, a presença do motivo RE-like dentro de RNAs desta família pode ajudar novos estudos focados em elucidar suas funções.

Figura 26: Estrutura de RNAs da família HgcC (RF00062). Destacado em vermelho encontra-se o motivo RE-like tetraloop e em azul o motivo GNRA tetraloop (RM00008).

4.1.5) Superexpressão de RNAs contendo o motivo RE-like: identificação de fenótipo em curva de crescimento

Após a caracterização inicial dos sotRNAs e identificação de ncRNAs intergênicos contendo o motivo RE-like, procedemos para estudos focados em compreender o impacto e importância destes transcritos no fenótipo e na rede de regulação gênica de *H. salinarum* NRC-1.

Construímos duas linhagens de superexpressão: VNG_sot0042, superexpressando o sotRNA VNG_sot0042 e VNG_R0052, superexpressando o PATE-RNA VNG_R0052, ambos possuindo o motivo RE-like. Realizamos curva de crescimento padrão, comparando as linhagens de superexpressão com a linhagem controle PMTF-0 e a linhagem selvagem NRC-1. Neste ensaio, ambas linhagens demonstraram maior crescimento quando comparada com os controles (Figura 27), demonstrando que estes ncRNAs são funcionais e devem impactar a rede de regulação gênica deste organismo. O fenótipo semelhante pode indicar também a importância do motivo RE-like presente em ambas linhagens.

Figura 27: Curva de crescimento de 4 linhagens de *H. salinarum* NRC-1. Linhagens superexpressando RNAs (VNG_R0052 e VNG_sot0042) que contêm o motivo *tetraloop* apresentam maior crescimento quando comparadas com as linhagens controle (NRC-1, selvagem e PMTF-0, com o vetor de superexpressão vazio).

O papel exato de *tnpB* ainda é uma questão em aberto, porém os recentes relatos indicam um possível efeito inibitório nos processos de excisão e inserção de IS200/605 (PASTERNAK *et al.*, 2013). Dado que todos transcritos identificados neste trabalho contém o motivo RE-*like* podemos especular que estas moléculas estão relacionadas neste processo. Uma possibilidade é a de que este motivo poderia atuar como competidor pela ligação de TnpA, evitando sua ligação em regiões de DNA de fita simples. Portanto, IS200/605 poderia atuar como reservatório importante de motivos estruturais (neste caso *stem loops*), como sugerido anteriormente por (DELIHAS, 2009)

4.2) VNG_sot0042: Identificação *in vitro* de proteínas parceiras, análise *in silico* de RNAs alvo e avaliação do impacto de sua superexpressão no transcrito

Nesta seção apresentarei os esforços realizados para iniciar a caracterização funcional de VNG_sot0042, um dos sotRNAs identificados, contendo o motivo RE-like. Este RNA teve suas extremidades 5' e 3' mapeadas por C-RACE, foi confirmado por *northern blot*, apresenta perfil de expressão anticorrelacionado com seu gene cognato em diversas condições e ainda sua superexpressão acarretou em maior crescimento em relação a linhagem controle. Utilizamos três abordagens distintas: 1) SILAC para identificação de potenciais parceiros de proteínas parceiras ligantes a este RNA; 2) Análise *in silico* de potenciais mRNAs alvos e 3) RNA-seq para avaliar o impacto desta superexpressão no transcrito de *H. salinarum* NRC-1. Em seguida, os resultados gerados por estas técnicas foram utilizados como guias para novos ensaios fenotípicos.

4.2.1) Identificação de proteínas que interagem com VNG_sot0042 através de SILAC

A interação entre proteínas e RNAs tem se mostrado cada vez mais importante para o entendimento de redes de regulação gênica (LE NOVÈRE, 2015). Exemplos clássicos deste tipo de interação envolvem proteínas extremamente importantes como Hfq/LSm, que pode facilitar a ligação de ncRNAs com seus alvos ou recrutar a maquinaria de degradação de RNAs (SCHU et al., 2015). Um fator de extrema relevância neste tipo de interação é a presença de motivos estruturais e de sequência tanto nos RNAs alvo como em seus pares proteicos (JANKOWSKY; HARRIS, 2015). Como demonstrado anteriormente, sotRNAs e o PATE-derived RNA identificados por nosso grupo apresentam um motivo conservado chamado RE-like que compreende um UUCA *tetraloop* (GOMES-FILHO et al., 2015). Este motivo poderia, por exemplo, interagir com a proteína TnpA e inibir o processo de transposição de IS200/605 em *H. salinarum* NRC-1. Para testar esta hipótese, decidimos realizar uma identificação *in vitro* em larga escala de proteínas que interagem com VNG_sot0042.

Para a identificação das proteínas enriquecidas, nossos colaboradores do grupo do Prof. Falk Butter, IMB Mainz - Alemanha, realizaram a técnica de SILAC (BUTTER et al., 2009). Esta técnica consiste em crescer o organismo de interesse em dois meios de cultura diferentes, um contendo $^{13}\text{C}_6$ -Lys e $^{13}\text{C}_6$ -Arg (*light*) e o outro contendo $^{14}\text{C}_6$ -Lys e $^{14}\text{C}_6$ -Arg (*heavy*). Após

isso, os extratos proteicos passam por colunas contendo construções diferentes do RNA de interesse. Para este ensaio, três construções diferentes foram utilizadas: 1) Corresponde a sequência nativa do sotRNA (VNG_sot0042); 2) Sequência nativa do sotRNA VNG_sot0042, porém com o motivo *RE-like* deletado (VNG_sot0042dm) e 3) Sequência randômica de mesmo tamanho e conteúdo dinucleotídico de VNG_sot0042 (random206) (Figura 28).

Figura 28: Esquematização dos RNAs sintetizados para o experimento de SILAC. VNG_sot0042 representa a sequência nativa do sotRNA testado. VNG_sot0042dm é uma construção onde o motivo *RE-like* foi deletado e random206 é um RNA de sequência randômica de mesmo tamanho e conteúdo dinucleotídico que VNG_sot0042. Em azul claro encontra-se o RNA de interesse, em azul escuro o aptâmero S1 e em vermelho um proteína hipotética interagindo com o motivo de interesse, neste caso *RE-like*.

Um fato interessante é que mesmo após a deleção do motivo *RE-like* de VNG_sot0042, algumas estruturas secundárias se mantiveram estáveis (Figura 29). Nas análises convencionais de SILAC, a busca por proteínas enriquecidas é realizada sempre comparando o RNA de interesse com a sua versão contendo deleção do motivo estudado ((BUTTER et al., 2009). Em nosso caso, a utilização do RNA random206 e a estabilidade de estruturas secundárias adicionais ao motivo *RE-like* encontradas em VNG_sot0042 WT e DM, permitiram a realização de uma análise complementar para determinar quais proteínas interagem de maneira dependente e não-dependente de *RE-like*.

Figura 29: Estruturas secundárias obtidas através do software RNAfold (GRUBER et al., 2008). A) VNG_sot0042 em sua versão nativa; B) VNG_sot0042 contendo a deleção do motivo RE-*like*. O motivo RE-*like* encontra-se destacado em azul. Os quadros vermelhos demonstram as regiões estruturadas que se mantém estáveis mesmo após a deleção do motivo RE-*like*.

As corridas são realizadas em quadruplicata para cada construção. Em seguida, as proteínas retidas na coluna são enviadas para análise por espectrometria de massa e as proteínas que apresentarem pelo menos dois peptídeos únicos em cada corrida terão seus $\log_2$ da razão de abundância calculados e relacionados com o p-valor obtido após comparação das quadruplicatas (Figura 30).

Figura 30: Volcano plot das proteínas enriquecidas quando comparamos o pull-down de sot0042WT e sot0042dm. No eixo y temos o $-\log_{10}$ dos p-valores de cada proteína, de modo que as proteínas com significância estatística se encontrarão mais acima no gráfico. No eixo x temos o $\log_2$ da razão da abundância das proteínas nas condições analisadas. As proteínas encontradas a direita e acima da linha listrada, nos quadrantes de sot0042WT, representam as proteínas enriquecidas na presença do motivo RE-like. As proteínas encontradas a esquerda e acima da linha listrada, nos quadrantes de sot0042dm, são proteínas que interagem de maneira não dependente do motivo RE-like. Em verde: Proteínas não-enriquecidas. Em laranja: Proteínas enriquecidas. Os volcano plots restantes podem ser encontrados no ANEXO C.

A lista das proteínas enriquecidas para cada construção/relação pode ser encontrada na (Tabela 8).

Tabela 8: Proteínas enriquecidas em cada relação estabelecida no experimento de SILAC.

Grupo I VNG_sot0042dm vs random206			Grupo II VNG_sot0042WT vs random206		
Uniprot	Gene	Função	Uniprot	Gene	Função
Q9HQA3	VNG_RS04915	Hipotética	Q9HQA3	VNG_RS04915	Hipotética
Q9HPW6	VNG_RS05610	Tradução de sinal (Htr12)	Q9HQP0	VNG_RS04175	Glicerol desidrogenase
Q9HNG6	VNG_RS07720	glicose-6-p isomerase	Q9HNG6	VNG_RS07720	glicose-6-p isomerase
Q9HNG2	VNG_RS08205	orotato ribosiltransferase	Q9HQY0	VNG_RS03715	Hipotética
Q9HQY0	VNG_RS03715	Hipotética	Q9HQG1	VNG_RS04605	Regulador de transcrição
Q9HP99	VNG_RS06740	Hipotética	Q9HPY3	VNG_RS05530	Hipotética
O51989	VNG_RS10955	Transposase ISNCY-ISH7B	Q9HP58	VNG_RS06950	Hipotética
Q9HQT5	VNG_RS03955	Tradução de sinal (Htr13)	Q9V2W0	VNG_RS04435	Proteína ribossomal
P24574	VNG_RS10630	Vesículas de gás	Q9HQ93	VNG_RS04970	Hipotética
P24574	VNG_RS11400	Vesículas de gás	Q9HQ60	VNG_RS05135	Hipotética
Q9HPY3	VNG_RS05530	Hipotética	Q9HR04	VNG_RS03600	Transporte de ferro
Q9HPB1	VNG_RS06670	Translocação de proteínas	Q9HS21	VNG_RS01720	Hipotética
Q9HQ60	VNG_RS05135	Hipotética	Grupo III VNG_sot0042WT vs VNG_sot0042dm		
Q9HQG1	VNG_RS04605	Regulador de transcrição	Uniprot	Gene	Função
Q9HP58	VNG_RS06950	Adesão	Q9HQP0	VNG_RS04175	Glicerol desidrogenase
Q9HQL8	VNG_RS04290	Hipotética	Q9HMX6	VNG_RS09170	Hipotética
O52003	VNG_RS11035	RNA helicase	Q9HQY8	VNG_RS03670	Quimiotaxia
Q9HR04	VNG_RS03600	Transporte de ferro	Q9HS89	VNG_RS01380	Hipotética
Q9HMT4	VNG_RS09380	Transporte de ferro			
Q9HS21	VNG_RS01720	Hipotética			

De acordo com as construções descritas na (Figura 29) e as relações apresentadas na (Tabela 8), dividimos as proteínas em dois grupos para análise: 1) Proteínas que interagem com VNG_sot0042 de maneira dependente do motivo *RE-like*, ou seja proteínas que aparecem exclusivamente nos Grupos II e/ou III ou em ambos; 2) Proteínas que interagem de maneira independente do motivo *RE-like*, ou seja, proteínas que foram identificadas nos grupos I e II. Deste modo, identificamos 6 proteínas que interagem de maneira dependente do motivo *RE-like* e 20 de maneira independente (Figura 31).

Figura 31: Diagrama de Venn representando a sobreposição das proteínas encontradas em cada relação feita nos dados de espectrometria de massa. As proteínas presentes exclusivamente em WT x 206 (Grupo II), WTxDm (Grupo III) e em ambos grupos II e III, são proteínas que interagem de maneira dependente ao motivo RE-like. As proteínas presentes em Dm x 206 (Grupo I) demonstram interação não dependente do motivo RE-like.

Como discutido anteriormente, uma de nossas hipóteses era que o motivo RE-like, presente nos sotRNAs e PATE-RNAs, poderia atuar como uma isca para a proteína TnpA inibindo sua ação no DNA hospedeiro. Das proteínas identificadas que interagem de maneira dependente do motivo RE-like (Tabela 9), não fomos capazes de identificar nenhuma TnpA. Hipóteses que poderiam justificar a ausência desta proteína no ensaio são, a reação de ligação ocorrer em condições *in vitro* e portanto algum outro fator encontrado somente *in vivo* não esteja presente, as condições de tampão não são as adequadas para esta interação ou TnpA não é abundante o suficiente para ser detectada. Futuros ensaios *in vivo* focados especificamente em TnpA interação poderão auxiliar a responder esta questão.

Tabela 9: Proteínas identificadas que interagem com o sotRNA VNG_sot0042 de maneira dependente ao motivo RE-like.

Uniprot	Gene	Função
Q9HQP0	VNG_RS04175	Glicerol-3-fosfato desidrogenase
Q9V2W0	VNG_RS04435	Proteína ribossomal 50S L18E
Q9HQ93	VNG_RS04970	Não determinada
Q9HMX6	VNG_RS09170	Não determinada
Q9HQY8	VNG_RS03670	CheW2 – acoplamento de CheA à transdutores de sinal
Q9HS89	VNG_RS01380	
		Não determinada

Apesar da ausência de TnpA, algumas das proteínas encontradas se demonstram bem interessantes como futuros alvos de validação. O gene VNG_RS04175 codifica a proteína Glicerol-3-fosfato desidrogenase, esta proteína apresenta papel fundamental na biossíntese de lipídeos, atuando na redução de diidroxiacetona fosfato em glicerol-3-fosfato. Em *Haloferax volcanii*, o nocaute deste gene torna este organismo incapaz de crescer utilizando glicerol (SHERWOOD; CANO; MAUPIN-FURLOW, 2009). Ademais, o gene VNG_RS03670 codifica a proteína CheW2, esta proteína está envolvida no processo de transdução de sinal de *H. salinarum* NRC-1 atuando no processo de acoplamento de CheA a diversos transdutores de sinal (GEGNER et al., 1992). Estudos de proteômica em *H. salinarum* R1 demonstraram que esta proteína interage com os transdutores Htr1, Htr2, BasT, Htr4, CosT, Htr6, Htr8, Htr15, Car e Htr13 (SCHLESNER et al., 2012). Por fim, a presença de VNG_RS04435, que codifica para um componente ribossomal, sugere que RNAs contendo o motivo RE-like podem estar envolvidos no processo de síntese proteica.

Em relação as proteínas que não interagem de maneira dependente ao motivo RE-like (Tabela 10), podemos destacar algumas que poderiam nos direcionar para novos ensaios fenotípicos.

Tabela 10: Proteínas que interagem com o sotRNA VNG_sot0042 de maneira não dependente ao motivo RE-like.

Uniprot	Gene	Função
Q9HS21	VNG_RS01720	Ferritina-like
Q9HR04	VNG_RS03600	Ibp – ligação a ferro
Q9HQY0	VNG_RS03715	Putativa proteína flagelar C
Q9HQT5	VNG_RS03955	Htr13 – Transdutor de sinal
Q9HQL8	VNG_RS04290	Não caracterizada
Q9HQG1	VNG_RS04605	Regulador de transcrição AsnC/Lrp
Q9HQA3	VNG_RS04915	Não caracterizada
Q9HQ60	VNG_RS05135	Não caracterizada
Q9HPY3	VNG_RS05530	Não caracterizada
Q9HPW6	VNG_RS05610	Htr12 – transdutor de sinal
Q9HPB1	VNG_RS06670	SecY – translocação de proteínas
Q9HP99	VNG_RS06740	Não caracterizada
Q9HP58	VNG_RS06950	YpjA – Adesão a superfícies sólidas
Q9HMQ6	VNG_RS07720	pgi – fosfoglicoisomerase
Q9HNG2	VNG_RS08205	pyrE
Q9HMT4	VNG_RS09380	cysT1
P24574	VNG_RS10630	gvpC1 – vesículas de gás
P24574	VNG_RS11400	gvpC2 – vesículas de gás
O52003	VNG_RS11035	hepA – RNA helicase ATP dependente
O51989	VNG_RS10955	Transposase da família ISNCY -ISH7B

A proteína codificada por VNG_RS03715 está relacionada com a montagem da maquinaria flagelar de *H. salinarum* NRC-1. Em *Haloferax vivax* estudos demonstraram que a maquinaria flagelar está relacionada com a capacidade de adesão (LEGERME et al., 2016). A proteína codificada por VNG_RS06950, por sua vez, apresenta similaridade com a proteína YpjA, encontrada em *E. coli*. A deleção do gene *ypjA*, em *E. coli*, não demonstrou efeito no perfil de adesão, porém a superexpressão ocasionou um aumento na adesão a superfícies sólidas (ROUX; BELOIN; GHIGO, 2005).

Além de proteínas relacionadas a adesão, identificamos duas proteínas relacionadas ao metabolismo de ferro. As proteínas codificadas pelos genes VNG_RS01720 e VNG_RS03600 são, respectivamente, uma ferritina-like associada ao armazenamento de ferro no citoplasma e um componente do transportador ABC responsável pelo importe de íons Fe^{3+} .

Além disso, o gene VNG_RS04605, que codifica um fator de transcrição do grupo AsnC/Lrp (*asparagine synthase C / leucine-responsive regulatory protein*) foi um achado curioso. O grupo AsnC/Lrp é o mais estudado em bactérias e regula até 75 genes em *E. coli*, a maioria relacionada ao metabolismo de aminoácidos (CHEN; CALVO, 2002). Apesar de

fatores de transcrição serem comumente vistos como proteínas ligantes de DNA, estudos demonstraram que diversos fatores de transcrição também ligam-se a RNAs, como descrito para TFIID de *Xenopus*, p53 em mamíferos e fator σ^{70} em *E. coli* (CASSIDAY, 2002). Deste modo, a interação de VNG_sot0042 apresenta mais um caso de dualidade de fatores de transcrição.

Ademais, dois genes associados ao processo de transdução de sinal foram identificados. Neste caso temos *htr12* (VNG_RS05610) e *htr13* (VNG_RS03955), ambos homólogos a *methyl-accepting chemotaxis proteins* (MCPs) de eucariotos. É interessante ressaltar que o transdutor Htr13 já foi demonstrado como parceiro de interação da proteína CheW2 em *H. salinarum* R1, a qual encontramos interagindo de maneira dependente ao motivo RE-like. O sinal interpretado por estes transdutores ainda é desconhecido.

Por fim, a presença do gene VNG_RS06670 que codifica para a proteína SecY, um componente importante do sistema de translocação de proteínas (COREY et al., 2016) sugere que VNG_sot0042 possa atuar durante o processo de exportação de proteínas.

4.2.2) Identificação *in silico* de mRNAs alvo de interação com VNG_sot0042

Na seção anterior, demonstramos que VNG_sot0042 interage com diversas proteínas envolvidas em processos celulares variados, sugerindo um possível papel regulatório. Uma outra maneira pela qual ncRNAs podem atuar como elementos regulatórios é através do pareamento com mRNAs (TANG et al., 2005). Este pareamento pode alterar características do mRNA alvo, como tempo de meia-vida, acessibilidade ao sítio de ligação do ribossomo, mudanças estruturais em seus UTR, entre outros (GOTTESMAN; STORZ, 2011).

Para avaliar os potenciais alvos de regulação de VNG_sot0042, realizamos análises utilizando a ferramenta CopraRNA (WRIGHT et al., 2014). Esta ferramenta utiliza como *input* no mínimo três homólogos do RNA de interesse encontrados em espécies diferentes, como apresentado anteriormente os homólogos de VNG_sot0042 foram encontrados utilizando a ferramenta GlassGO (ANEXO). Após isso, a ferramenta utiliza a plataforma IntaRNA (WRIGHT et al., 2014), que combina predições de pareamento de bases e estrutura secundária para encontrar os alvos em cada espécie analisada. Com as predições obtidas a ferramenta CopraRNA agrupa os genes homólogos a partir de suas sequências de aminoácido e então utiliza estes grupos para calcular o p-valor de cada agrupamento. As predições que apresentam p-valor < 0,05 são consideradas significativas para fins de análise.

Dos alvos com potencial de interação (Tabela 11), podemos destacar os genes *tnpB* presentes em *H. salinarum* NRC-1. A região de interação detectada inicia-se na posição 13 do mRNA e se estende até a posição 19. Em VNG_sot0042, a região de interação inicia-se no nucleotídeo 141 e pode se estender até o nucleotídeo 149 (Figura 32). É interessante ressaltar que o motivo RE-like, presente em VNG_sot0042, inicia-se no nucleotídeo 150 e se estende até o nucleotídeo 183. Portanto, a região de interação de VNG_sot0042 com os mRNAs alvos não é o motivo RE-like, porém pela proximidade podemos sugerir que a estrutura conservada deste motivo poderia atuar na estabilização das interações. Por fim, a associação dos perfis de expressão opostos e o alto potencial de interação, podemos sugerir que os sotRNAs atuam tanto em *cis* quanto em *trans* na regulação da expressão dos homólogos de *tnpB*.

Tabela 11: Possíveis alvos de interação de VNG_sot0042 preditos utilizando a ferramenta CopraRNA (WRIGHT et al., 2014) . As interações consideradas foram as que apresentaram p-valor < 0,05. É possível visualizar a Mínima Energia Livre em kcal/mol de cada interação e a região de interação entre o mRNA alvo e o sRNA. A possibilidade de interação com homólogos do gene alvo predito é relatada na coluna homólogos adicionais.

Rank	P-valor	Gene	Energia [kcal/mol]	Posição mRNA	Posição sRNA	Homólogos adicionais
1	1,75E-05	vng_rs001700 (<i>tnpB</i>)	-8,33	13 -- 19	141 -- 147	vng_rs10400 vng_rs11205 vng_rs11945 vng_rs11320 vng_rs00160 vng_rs11800 vng_rs00050 vng_rs00105 vng_rs01145
2	3,25E-05	vng_rs05360 (histidina quinase)	-6,64	13 -- 19	141 -- 147	
3	3,65E-05	vng_rs04700 (Flavina mononucleotídeo fosfatase)	-11,87	108 -- 116	141 -- 149

Figura 32: Esquemática da interação predita do sotRNA VNG_sot0042 com o mRNA de genes *tnpB* em *H. salinarum* NRC-1.

Outros dois genes apresentaram potencial de interação antisense, VNG_RS05360 e VNG_RS04700. A proteína codificada por VNG_RS05360 possui domínios associados a função de histidina quinase. As histidinas quinase atuam na etapa inicial da cascata do sistema de transdução de sinal de dois componentes. Este sistema está associado a respostas celulares a mudanças no ambiente extracelular (CAPRA; LAUB, 2012). O gene VNG_RS04700 codifica uma flavina mononucleotídeo (FMN) fosfatase chamada de YigB. Em *E. coli* demonstrou-se que esta proteína quando superexpressa aumenta o número de células persistentes, comumente associadas a resistência multidroga, através da depleção do cofator FMN (HANSEN; LEWIS; VULIĆ, 2008). Deste modo, os sotRNAs podem atuar como reguladores de pontos importantes na sobrevivência de *H. salinarum* NRC-1, seja regulando

componentes dos sistemas de transdução de sinal ou genes associados a manutenção de células persistentes.

4.2.3) Impacto da superexpressão de VNG_sot0042 no transcrito de *H. salinarum* NRC-1

A comparação de transcritomas de linhagens diferentes, de uma mesma espécie, é extremamente útil para a identificação de vias metabólicas e mecanismos de resposta a estresses diversos, podendo explicar diferenças fenotípicas (GERSTEIN et al., 2014). O fenótipo de maior crescimento da linhagem superexpressando o RNA VNG_sot0042 foi observado em condições de cultivo padrão (Figura 27). Deste modo, decidimos construir bibliotecas de RNA-seq a partir de culturas de *H. salinarum* VNG_sot0042 (superexpressão de VNG_sot0042) e *H. salinarum* PMTF-0 (linhagem controle) a OD600 0,5, cultivadas a 37° C, 225rpm e na presença de luz. Foram construídas 4 bibliotecas, sendo duas réplicas biológicas da linhagem VNG_sot0042 (pMTF0042-A e pMTF0042-B) e duas réplicas biológicas da linhagem controle PMTF-0 (pMTF0-A e pMTF0-B).

As 4 bibliotecas apresentaram um número balanceado tanto de *reads* sequenciados, processados para remoção de adaptadores, quanto de *reads* alinhados no genoma de *H. salinarum* NRC-1 (Tabela 12).

Tabela 12: Número de *reads* sequenciados, processados e alinhados.

Biblioteca	Reads Sequenciados	Reads Processados	Reads Alinhados
pMTF0-A	5.208.952	2.592.356	1.364.291 (52,63%)
pMTF0-B	5.222.176	2.578.788	1.355.381 (52,56%)
pMTF0042-A	5.109.720	2.546.063	1.380.240 (54,21%)
pMTF0042-B	4.891.124	2.435.961	1.297.769 (53,28%)

Para identificar os genes diferencialmente expressos, duas etapas foram necessárias. A primeira foi realizada utilizando o software HTSeq para contagem dos *reads* alinhados a cada gene (ANDERS; PYL; HUBER, 2015). Posteriormente, o cálculo de expressão diferencial foi realizado utilizando o pacote DESeq2. Nesta análise foram considerados os genes que apresentaram *FoldChange* > 1,5 e p-valor < 0,05 ou *FoldChange* < -1,5 e p-valor < 0,05 (LOVE; HUBER; ANDERS, 2014).

Foram identificados 220 genes, aproximadamente 8,5% de todos os genes anotados, diferencialmente expressos na linhagem *H. salinarum* VNG_sot0042. Destes, 94 se apresentaram mais expressos e 116 menos expressos. De maneira a identificar quais as vias

metabólicas e processos celulares se mostraram alterados na linhagem VNG_sot0042, utilizamos a ferramenta KEGG *orthology* (KO) para associação dos genes às suas possíveis funções (AOKI-KINOSHITA; KANEHISA, 2007). Na ausência de um KO associado, utilizamos dados provenientes do banco MicrobesOnline e a da ferramenta BlastCD para identificar domínios proteicos conservados.

Os KOs mais prevalentes foram de genes relacionados aos processo de tradução (25), metabolismo de carboidratos (17), metabolismo de aminoácidos (15) e metabolismo energético (12) (Figura 33).

Figura 33: Representatividade dos KOs identificados nos genes diferencialmente expressos.

Dos genes identificados como potenciais alvos de interação sotRNA-mRNA identificados através da ferramenta CopraRNA (Tabela 11), os genes *tnpB* não demonstraram alteração em sua expressão. Por outro lado, os genes VNG_RS05360 e VNG_RS04700 se encontram mais expressos. O gene VNG_RS05360 será abordado mais a fundo na seção (**Adesinas e componentes do sistema de transdução de sinal**).

Transportadores ABC e o sistema de translocação de proteínas Sec

Dentro dos genes diferencialmente expressos, identificamos um total de 15 genes associados ao transporte de membrana, dos quais 14 se encontram mais expressos e somente um menos expresso (Tabela 13). Podemos dividir estes genes em três grupos: 1) Transportadores ABC; 2) Sistema de translocação de proteínas - Sec; 3) Outros transportadores. É interessante ressaltar que destes genes, somente um deles (VNG_RS02960) se apresentou menos expresso.

Os transportadores ABC são proteínas de membrana que, de maneira geral, utilizam a hidrólise de ATP para realizar a translocação de solutos (íons, peptídeos, aminoácidos, vitaminas) através da membrana celular (REES; JOHNSON; LEWINSON, 2009). Estas proteínas são de extrema importância para manutenção da homeostase osmótica, captação de nutrientes, eventos de esporulação e outras funções celulares (TER BEEK; GUSKOV; SLOTBOOM, 2014). Em nosso caso, identificamos transportadores ABC associados ao transporte de cobalto, fosfato, sulfato, ferro e peptídeos (Tabela 13).

Tabela 13: Genes diferencialmente expressos, componentes dos sistemas de transporte ABC e translocação de proteínas SEC.

Gene	Log2 Fold Change	Nome
VNG_RS02960	-0,70	anion permease
VNG_RS02055	0,52	ABC-type transport Fe-S cluster assembly protein SufD
VNG_RS02045	0,53	ABC transporter ATP-binding protein
VNG_RS06670	0,58	preprotein translocase subunit SecY
VNG_RS01470	0,59	preprotein translocase subunit SecE
VNG_RS01775	0,59	phosphate ABC transporter substrate-binding protein
VNG_RS01755	0,67	phosphate transport system regulatory protein PhoU
VNG_RS07700	0,69	preprotein translocase subunit SecD
VNG_RS09190	0,71	ABC transporter ATP-binding protein oppD2
VNG_RS09905	0,75	peptide ABC transporter permease
VNG_RS06335	0,94	cobalt transporter CbiN
VNG_RS09910	0,94	peptide ABC transporter permease
VNG_RS09385	1,00	ABC sulfate transport system ATP-binding protein cysA
VNG_RS06325	1,08	cobalt transport ATP-binding protein cbiO2
VNG_RS06765	2,00	ZIP family metal transporter

Os genes VNG_RS09900 (*ddpA*), VNG_RS09905 (*dppB2*) e VNG_RS09910 (*dppC1*) pertencem ao operon *ddpAdppB2C1*, que codifica subunidades de um transportador ABC de dipeptídeos. Dos três genes, apenas *ddpA* não se mostra mais expresso. Análises de *tiling array*, obtidos a partir de uma curva de crescimento demonstraram que, com exceção de *ddpA* todos os genes deste operon são mais expressos em densidades celulares maiores (KOIDE et al., 2009). Este fato é bem curioso, pois nossos dados são referentes a OD600 0,5, correspondente a primeira fase exponencial de *H. salinarum* NRC-1. Além disso, a análise do transcrito de *H. salinarum* NRC-1 crescendo em baixas concentrações de sal (2,9M), que resulta em menor crescimento, demonstrou que componentes do sistema de transporte de cobalto (*cbiQ* e *cbiN*) e o gene *oppD2* encontram-se menos expressos (COKER et al., 2007), ao contrário do visualizado em nossos dados, onde estes genes são mais expressos. Deste modo, podemos sugerir que estes genes, quando na superexpressão de VNG_sot0042, se comportam de maneira semelhante a situações de maior crescimento e de maneira oposta a pelo menos uma condição de baixo crescimento.

Um dos sistemas mais importantes para translocação de proteínas através da membrana é o sistema Sec, o qual é conservado em todos os domínios da vida (COREY et al., 2016). Em *H. salinarum* NRC-1, este sistema é composto pelos genes *secD*, *secE*, *secF* e *secY* (BOLHUIS, 2002). Destes componentes, apenas o gene *secF* não se apresenta diferencialmente expresso em nossas análises. *H. salinarum* NRC-1 apresenta 28 proteínas contendo o sinal de secreção do tipo SEC (BOLHUIS, 2002). Dentre estas, podemos destacar as proteínas CbiN e VNG_RS06420 (proteína hipotética) que apresentam seus genes mais expressos. É importante ressaltar que nos ensaios de SILAC identificamos a proteína SecY como parceira de interação de VNG_sot0042, demonstrando que a relação deste sotRNA com o sistema de translocação Sec não está limitada a nível transcricional.

Adesinas e componentes do sistema de transdução de sinal

Em Archaea, a habilidade de se locomover e aderir a diversas superfícies são fatores crucial no ciclo de vida destes organismos, sendo importantes para buscar nutrientes e evitar toxinas presentes no meio, assim como estabelecer a formação de biofilmes em resposta a diversos fatores ambientais (LEGERME et al., 2016). Nos dados de SILAC, havíamos encontrado a proteína associada a adesão YpjA e os transdutores Htr12 e Htr13. Do conjunto

de genes diferencialmente expressos, encontramos um gene associado a adesão e dois transdutores de sinal (Tabela 14). O gene VNG_RS11470, um dos genes mais expressos em nosso conjunto, é um componente do sistema ZnuABC. Este sistema apresenta papel fundamental na absorção de zinco em *E.coli* OH157:H7 e é altamente induzido quando este organismo encontra-se aderido a células epiteliais em cultura, além disso o nocaute destes componentes prejudicaram a adesão (GABBIANELLI et al., 2011).

Tabela 14: Genes diferencialmente expressos anotados como adesina ou transdutores de sinal.

Gene	Log2 Fold Change	Nome
VNG_RS11470	1,15	adhesin
VNG_RS07200	0,57	Htr3
VNG_RS05920	-0,54	Htr8
VNG_RS05360	0,57	histidine kinase
VNG_RS07895	-0,85	histidine kinase

Os genes VNG_RS07200 e VNG_RS05920, *htr3* e *htr8* respectivamente, são importantes componentes do sistema de transdução de sinal de *H. salinarum* NRC-1 (DASSARMA; DASSARMA, 2012). *htr3* também é conhecido como BasT, um importante quimiotransdutor para cinco aminoácidos, leucina, isoleucina, valina, metionina e cisteína (KOKOEVA; OESTERHELT, 2000). *htr8*, por sua vez, é um transdutor de sinal responsável pela capacidade de aerotaxia de *H. salinarum* NRC-1 (BROOUN et al., 1998). A maior expressão de *htr3* e menor expressão de *htr8* sugerem uma preferência desta linhagem por condições de menor concentração de oxigênio.

Além de transdutores, pudemos identificar duas histidinas quinases diferencialmente reguladas, VNG_RS05360 e VNG_RS07985. Um fato intrigante em relação a VNG_RS05360 é que este gene foi identificado como potencial alvo de regulação *trans*-antisense por VNG_sot0042. A ligação de VNG_sot0042 ao mRNA de VNG_RS05360 poderia atuar como um fator de estabilização e proteção, aumentando a meia-vida do RNA alvo e portanto seu acúmulo na célula. Evento semelhante foi descrito em *Streptococcus* do grupo A, onde o RNA FasX se liga ao 5' do mRNA de uma estreptoquinase e atua como uma barreira contra nucleases (RAMIREZ-PEÑA et al., 2010).

Reguladores de transcrição, tRNAs e outros ncRNAs

Dois reguladores de transcrição da família AsnC/Lrp, discutida anteriormente, VNG_RS05270 e VNG_RS07480 apresentam-se mais expressos (Tabela 15).

Tabela 15: Reguladores de transcrição da família AsnC/Lrp diferencialmente expressos.

Gene	Log2 Fold Change	Nome
VNG_RS05270	0,60	AsnC/Lrp transcriptional regulator
VNG_RS07480	0,63	AsnC/Lrp transcriptional regulator

Em *H. salinarum* R1, a superexpressão e nocaute de dois genes *lrp*, demonstrou que estes reguladores induzem a expressão de sistemas de transporte de peptídeos, transporte de fosfato, glutamina sintetase, oxoglutarato redutase e reprimem a expressão de aspartato transaminase, fator de transcrição TfbB e glicerol desidrogenase (SCHWAIGER et al., 2010). Em nossos dados, verificamos um padrão de expressão semelhantes ao descrito (Tabela 16).

Tabela 16: Comparação do perfil de expressão dos genes afetados pelos reguladores de transcrição *lrp*. Em azul, genes menos expressos. Em vermelho, genes mais expressos. Os perfis de KO e superexpressão foram obtidos de (SCHWAIGER et al., 2010).

Sistema / gene	<i>H. salinarum</i> R1 KO <i>lrp</i>	<i>H. salinarum</i> R1 Superexpressão <i>lrp</i>	<i>H. salinarum</i> VNG_sot0042
Transporte de fosfato	-	+	+
Transporte de dipeptídeos	-	+	+
Síntese de glutamina	-	+	+
<i>gldA1</i>	+	-	-
<i>aspB3</i>	+	-	-
<i>tfbB</i>	+	-	-

Além disso, constatamos que 18 dos 47 tRNAs encontrados em *H. salinarum* encontram-se menos expressos (Tabela 17). Estudos realizados em *Saccharomyces cerevisiae* demonstraram que em estresses relacionados a oxidação (H₂O₂) e alquilação (MMS) 58 tRNAs apresentam padrões idênticos de expressão, enquanto os 18 restantes apresentam perfis opostos (PANG et al., 2014).

Tabela 17: tRNAs diferencialmente regulados em nossa análise. Em azul, genes menos expressos.

Gene	Log2 Fold Change	Nome
VNG_RS00490	-0,67	tRNA-Val
VNG_RS00565	-0,59	tRNA-Asn
VNG_RS00805	-0,76	tRNA-Tyr
VNG_RS02070	-0,87	tRNA-Gln
VNG_RS02185	-0,66	tRNA-Met
VNG_RS03100	-0,78	tRNA-Ala
VNG_RS03210	-0,82	tRNA-Gly
VNG_RS04755	-0,77	tRNA-Thr
VNG_RS04785	-0,56	tRNA-Ser
VNG_RS06400	-0,57	tRNA-Val
VNG_RS06545	-0,60	tRNA-Gly
VNG_RS06715	-0,90	tRNA-Pro
VNG_RS07440	-0,85	tRNA-Leu
VNG_RS07665	-0,70	tRNA-Pro
VNG_RS07860	-0,60	tRNA-Gly
VNG_RS08420	-0,78	tRNA-His
VNG_RS09280	-0,98	tRNA-Gln
VNG_RS09795	-0,68	tRNA-Ala

Além de tRNAs, outros ncRNAs se mostraram diferencialmente expressos na linhagem de superexpressão (Tabela 18). Estes RNAs ainda não apresentam funções descritas, sendo bons candidatos para futuras validações experimentais.

Tabela 18: ncRNAs diferencialmente expressos em pMTF-0042.

ncRNA	Log2 Fold Change
VNG0030nc	-1,07
VNG0031nc	0,93
VNG0025nc	0,68

Via de biossíntese de purinas

A biossíntese de purinas envolve dez reações enzimáticas que levarão a formação de Inosina Monofosfato (IMP). Os genes *purF*, *purD*, *purL*, *purM* e *purC* são encontrados em

todos organismos que contém esta via (ZHANG; MORAR; EALICK, 2008). Em nossas análises, verificamos que os genes *purB*, *purK*, *purE*, *purL*, *purC* e *purL2* encontram-se mais expressos. Em *Bacillus cereus*, demonstrou-se que o nocaute de genes *purA*, *purC* e *purL* afetam drasticamente a capacidade de formação de biofilmes e adesão deste organismo, além de torná-lo mais suscetível a ação de actinomicina D (VILAIN et al., 2009). Em *H. salinarum* R1, uma análise proteômica demonstrou que as proteínas envolvidas no metabolismo de purinas, encontram-se mais abundantes em biofilmes maduros (LOSENSKY et al., 2017).

O papel de DNA extracelular (eDNA) no processo de adesão de *H. salinarum* ainda é desconhecido, mas evidências levantadas em outros procariotos indicam que o eDNA tem papel importante atuando como adesina.

Tabela 19: Genes da via de biossíntese de purinas encontrados diferencialmente expressos.

Gene	Log2 Fold Change	Nome
VNG_RS01615	0,78	<i>purB</i>
VNG_RS02505	0,67	<i>purK</i>
VNG_RS02510	0,53	<i>purE</i>
VNG_RS03375	0,72	<i>purL</i>
VNG_RS07530	0,56	<i>purC</i>
VNG_RS07550	0,80	<i>purL2</i>

Resistência a metais pesados

Os íons de metal são elementos chave na manutenção da homeostase celular. Metais como manganês, ferro e cobre servem de micronutrientes atuando, por exemplo, como centros redox em metaloproteínas (SRIVASTAVA; KOWSHIK, 2013). Os organismos evitam a toxicidade destes metais através de transporte seletivo, efluxo e conversão em estados redox menos tóxicos (KAUR et al., 2006b). Em *H. salinarum* NRC-1 estudos demonstraram que em exposição a diversos metais pesados, este organismo apresenta menor taxa de crescimento porém com a manutenção de células viáveis (KENNEDY et al., 2001). Além disso, estudos da resposta transcricional de *H. salinarum* NRC-1 em diversos estresses sob metais pesados demonstraram quais as vias e reguladores importantes para resistência a metal pesado neste organismo ((KAUR et al., 2006a). A (Figura 34) apresenta as vias de resistência a metal pesado de *H. salinarum* NRC-1 e os genes que se apresentam diferencialmente regulados na

linhagem VNG_sot0042 (Tabela 20). Três genes importantes puderam ser identificados em nossa análise: 1) *zntA* (VNG_RS00610) que codifica uma P1 ATPase relacionada diretamente com resistência a Cu, Co, Ni e Zn; 2) *ycdH* (VNG_RS11470) que codifica para um importante transportador de Mn e Zn que também está associado a processos de adesão e 3) *dpsA* (VNG_RS09555) que codifica uma ferritina responsável pela quelação de Fe(II).

Deste modo, na superexpressão podemos visualizar que os genes fundamentais associados a resistência dos metais Zn, Cu, Ni, Co e Fe se apresentam menos expressos, sugerindo que a linhagem VNG_sot0042 possa apresentar algum tipo de deficiência quando exposta a estes metais. Já em relação a Mn, é possível que a linhagem VNG_sot0042 apresente maior resistência devido a maior expressão do gene *ycdH* (VNG_RS11470).

Figura 34: Vias de resistência a metais pesados descritas em *H. salinarum* NRC-1. O regulador de transcrição VNG_RS04605 (VNG1179C) é importante para regulação de uma das vias de resistência a cobre. O gene *zntA* (VNG_RS00610) codifica uma P1 ATPase relacionada diretamente com resistência a Cu, Co, Ni e Zn. O gene *ycdH* (VNG_RS11470), codifica para um importante transportador de Mn e Zn que também está associado a processos de adesão. O gene *dpsA* (VNG_RS09555) codifica uma ferritina responsável pela quelação de Fe(II). Destacados em roxo, interação identificada por SILAC. Destacados em vermelho, mais expressos em VNG_sot0042. Destacados em azul, menos expressos em VNG_sot0042. Adaptado de Kaur et al, 2006b.

Tabela 20: Genes diferencialmente expressos associados a resistência frente a metais pesados

Gene	Log2 Fold Change	Nome
VNG_RS11470	1,1	<i>ycdH</i> (adhesin)
VNG_RS00610	-0,8	<i>zntA</i>
VNG_RS09555	-0,5	<i>dpsA</i>

Por fim, o gene VNG_RS004605 (VNG1179C), identificado nas análises de SILAC, codifica um regulador de transcrição do grupo AsnC/Lrp e é ponto importante na regulação de uma das vias de resposta a Cu (KAUR et al., 2006). Este regulador de transcrição ativa a transcrição do gene *yvgX* e de duas outras chaperonas relacionadas ao estresse de Cu. Em nossos dados o gene VNG_RS004605 e seus alvos de regulação não se mostram diferencialmente regulados, reforçando a atuação de VNG_sot0042 somente a nível pós-traducional.

4.2.4) Ensaios fenotípicos complementares

Os dados de RNA-seq e SILAC nos permitiram identificar um grande conjunto de genes e proteínas associados a diversas funções celulares como adesão, translocação de proteínas e transporte de íons de metal. Para melhor avaliar o fenótipo da linhagem superexpressando VNG_sot0042 realizamos ensaios com os seguintes objetivos: 1) Avaliar o perfil de crescimento em diversas concentrações de metais pesados; 2) Avaliar a viabilidade celular; 3) Avaliar o perfil de adesão.

As condições testadas nas curvas de crescimento e nos ensaios de sobrevivência se encontram na (Tabela 21).

Tabela 21: Concentrações de metais pesados utilizadas nas curvas de crescimento e ensaios de sobrevivência.

	Curva de crescimento			Sobrevivência
Metal	Concentração I mM	Concentração II mM	Concentração III mM	Concentração mM
CoSO ₄	0,2	0,3	0,5	0,5
MnSO ₄	0,8	1,0	1,5	2,0
CuSO ₄	0,4	0,8	1,0	-
FeSO ₄	2,0	4,0	5,0	9,8

Sobrevivência em CM

Como visualizado anteriormente nas curvas de crescimento padrão (Figura 27), a linhagem VNG_sot0042 apresenta maior crescimento. Para verificar se este fenótipo estava relacionado com a presença de células viáveis, testamos a sobrevivência da linhagem de superexpressão e da linhagem em meio CM. Neste ensaio, pudemos visualizar que nas condições padrão de crescimento, a linhagem VNG_sot0042 apresenta maior viabilidade celular, de maneira que nas diluições de 10^{-4} temos mais de dez vezes mais unidades formadoras de colônia (UFC) (Figura 35). Isso evidencia que o maior crescimento visualizado na curva é de fato associado a células viáveis.

Figura 35: Ensaios de sobrevivência em CM utilizando as linhagens *H. salinarum* VNG_sot0042 e *H. salinarum* PMTF-0. A viabilidade celular foi quantificada através do número de pixels presentes em cada spot nas placas de cultivo.

Exposição a CoSO_4 e CuSO_4

Nas curvas realizadas utilizando CoSO_4 como agente estressante verificamos um fenótipo interessante (Figura 36). Na presença de diferentes concentrações de CoSO_4 (Figura 36BCD), ambas linhagens apresentam menor crescimento quando comparadas a condição referência, sem estresse (Figura 36A). É importante ressaltar que nas (Figura 36B), (Figura 36C) apesar da linhagem VNG_sot0042 apresentar menor crescimento nas fases iniciais da curva a OD600 final é muito semelhante entre as linhagens.

Figura 36: Curvas de crescimento realizadas utilizando *H. salinarum* VNG_sot0042 e *H. salinarum* PMTF-0. A) Curva de crescimento nas condições referência. B) Curva de crescimento na presença de 0,2 mM de CoSO₄. C) Curva de crescimento na presença de 0,3 mM de CoSO₄. D) Curva de crescimento na presença de 0,5 mM de CoSO₄.

Na concentração de 0,4mM de CuSO₄ (Figura 37B) temos um perfil muito semelhante ao visualizado nas curvas na exposição a CoSO₄ (Figura 36BC). As outras concentrações de CuSO₄ testadas demonstraram efeito de inibição completa do crescimento (Figura 37C) e (Figura 37D).

Figura 37: Curvas de crescimento realizadas utilizando *H. salinarum* VNG_sot0042 e *H. salinarum* PMTF-0. A) Curva de crescimento nas condições referência. B) Curva de crescimento na presença de 0,4 mM de CuSO₄. C) Curva de crescimento na presença de 0,8 mM de CuSO₄. D) Curva de crescimento na presença de 1 mM de CuSO₄.

Como demonstrado anteriormente, o gene *zntA* que codifica uma P1 ATPase responsável pelo efluxo de Co, Cu, Ni e Zn (KAUR et al., 2006b) encontra-se menos expresso na linhagem VNG_sot0042. Esta menor expressão pode justificar a maior deficiência de crescimento apresentada pela linhagem VNG_sot0042 na presença de CoSO₄ e CuSO₄. Além disso, a interação do sotRNA VNG_sot0042 com o regulador de transcrição VNG_RS05605, identificada por SILAC, demonstra que a influência deste ncRNA não se limita somente a alterações transcricionais dos genes presentes nestas vias.

H. salinarum NRC-1 utiliza da diminuição da taxa de crescimento e manutenção de células viáveis para lidar com estresses associados a presença de metais pesados. Para avaliar se o crescimento tardio apresentado pela linhagem VNG_sot0042 estava relacionado a maior capacidade de produzir células viáveis, realizamos ensaios de sobrevivência a cobalto. Nestes ensaios (Figura 38), podemos visualizar um padrão semelhante ao encontrado nos ensaios realizados somente na presença de CM (Figura 35). Em todos os tempos testados, a linhagem VNG_sot0042 apresenta maior capacidade de sobrevivência.

Figura 38: Ensaio de sobrevivência a 0,5 mM de CoSO_4 utilizando as linhagens *H. salinarum* VNG_sot0042 e *H. salinarum* PMTF-0. A viabilidade celular foi quantificada através do número de pixels presentes em cada spot nas placas de cultivo.

Exposição a MnSO_4

As curvas utilizando MnSO_4 0,8 mM e 1 mM não demonstraram mudanças significativas no fenótipo quando comparadas a curva referência (Figura 39A), de modo que a linhagem VNG_sot0042 ainda demonstra maior taxa de crescimento, principalmente na segunda fase exponencial (Figura 39B) e (Figura 39C). Curiosamente, na curva de crescimento com 1,5 mM vemos que ambas linhagens apresentam o mesmo perfil de crescimento, ou seja, nesta concentração de MnSO_4 o efeito positivo na taxa de crescimento foi mais impactante na linhagem controle do que na superexpressão (Figura 39D).

Figura 39: Curvas de crescimento realizadas utilizando *H. salinarum* VNG_sot0042 e *H. salinarum* PMTF-0. A) Curva de crescimento nas condições referência. B) Curva de crescimento na presença de 0,8 mM de MnSO_4 . C) Curva de crescimento na presença de 1 mM de MnSO_4 . D) Curva de crescimento na presença de 1,5 mM de MnSO_4 .

Nos ensaios de sobrevivência a MnSO_4 visualizamos fenótipo semelhante ao encontrado em CoSO_4 de maneira que VNG_sot0042 apresenta mais células viáveis mesmo na presença do agente estressante.

Figura 40: Ensaios de sobrevivência a 2 mM de MnSO_4 utilizando as linhagens *H. salinarum* VNG_sot0042 e *H. salinarum* PMTF-0. A viabilidade celular foi quantificada através do número de pixels presentes em cada spot nas placas de cultivo.

Exposição a FeSO_4

A curva realizada na presença de FeSO_4 não apresentou diferença significativa entre o controle e a linhagem VNG_sot0042 (Figura 41BCD), demonstrando assim um fenótipo diferente do visto na curva de crescimento referência, onde a linhagem VNG_sot0042

apresenta maior crescimento. Este fenótipo pode ser justificado pela menor expressão do gene *dpsA* que codifica uma ferritina importante para a quelação de Fe(II).

Figura 41: Curvas de crescimento realizadas utilizando *H. salinarum* VNG_sot0042 e *H. salinarum* PMTF-0. A) Curva de crescimento nas condições referência. B) Curva de crescimento na presença de 2 mM de FeSO₄ C) Curva de crescimento na presença de 4 mM de FeSO₄ D) Curva de crescimento na presença de 5 mM de FeSO₄.

Assim como nos ensaios de CoSO₄ e MnSO₄ decidimos avaliar a viabilidade celular de VNG_sot0042 frente a FeSO₄. De maneira semelhante aos outros metais, a linhagem VNG_sot0042 apresenta maior número de células viáveis (Figura 42).

Figura 42: Ensaios de sobrevivência a 9,8 mM de FeSO₄ utilizando as linhagens *H. salinarum* VNG_sot0042 e *H. salinarum* PMTF-0. A viabilidade celular foi quantificada através do número de pixels presentes em cada spot nas placas de cultivo.

Ensaio de adesão a vidro

Em ambos os dados de SILAC e RNA-seq pudemos identificar genes e proteínas associadas ao processo de adesão. A interação de VNG_sot0042 com YpjA e a maior expressão de uma adesina e do sistema de biossíntese de purinas sugerem a relação deste sotRNA com este processo celular. Deste modo, através de ensaios de adesão a vidro avaliamos o padrão de adesão da linhagem VNG_sot0042. Este ensaio consiste na incubação *overnight* de lamínulas de vidro, em ângulo de 90 graus, dentro de meio CM contendo as linhagens a OD600 0,5. Após isso as lamínulas são coradas com cristal violeta, fotografadas (Figura 43AB) e o número de células aderidas é contado utilizando o software ImageJ. Por fim, a média do número de células e o desvio padrão das réplicas são calculados (Figura 43C).

Como visualizado na (Figura 43), a linhagem VNG_sot0042 apresenta quase três vezes mais células aderidas do que a linhagem controle, demonstrando que a superexpressão de VNG_sot0042 é importante para este fenótipo.

Figura 43: Ensaio de adesão utilizando as linhagens VNG_sot0042 e PMTF-0. A) Lamínula referente a PMTF-0; B) Lamínula referente a VNG_sot0042; C) Contagem das células aderidas a lamínulas de vidro utilizando o software ImageJ.

VNG_sot0042: Evidências de funcionalidade e regulação de processos celulares importantes

Através das análises de SILAC pudemos identificar a interação de VNG_sot0042 com diversas proteínas, com destaque para o regulador de transcrição do grupo AsnC/Lrp codificado pelo gene VNG_RS004605 associado a resposta ao estresse por Cu. A identificação de outras proteínas associadas ao processo de adesão, motilidade, transdução de sinal e translocação de proteínas sugere que este sotRNA está envolvido na regulação pós-traducional de diversos processos celulares.

Outra maneira pela qual este sotRNA poderia atuar como um regulador seria através do pareamento com mRNAs. Utilizando a ferramenta CopraRNA pudemos identificar 12 potenciais alvos de interação, sendo eles: 10 genes de *tnpB*, o gene VNG_RS05360 que codifica uma histidina quinase, envolvida no processo de transdução de sinal e o gene VNG_RS04700 (*yigB*) que codifica para uma flavina mononucleotídeo fosfatase associada a resistência multidroga e produção de células persistentes em *E. coli*.

Análises do transcrito da linhagem superexpressando o sotRNA VNG_sot0042 evidenciaram diversos processos celulares que são regulados de maneira direta ou indireta por este ncRNA. Dentre os processos celulares que apresentaram genes mais expressos podemos ressaltar: 1) Transportadores ABC de dipeptídeos e cobalto; 2) Sistema de translocação de proteínas SEC; 3) Adesinas (*ycdH* – gene também associado a resistência frente a Mn) e 4) Via de biossíntese de purinas. Dos que se apresentaram menos expressos podemos destacar: 1) 18 dos 47 tRNAs presentes em *H. salinarum* NRC-1; 2) *zntA*, que codifica uma P1 ATPase responsável pelo efluxo de diversos metais pesados e 3) *dpsA*, que codifica uma ferritina fundamental na detoxificação de ferro.

Ademais, ensaios de curva de crescimento na presença de metais pesados evidenciaram que a superexpressão deste sotRNA impacta negativamente o crescimento de *H. salinarum* quando em exposição a Cu, Co, Fe. Estes fenótipos corroboram o perfil encontrado nos dados de transcrito onde os genes *dpsA* e *zntA* encontram-se menos expressos. Na curva realizada na presença de Mn não há diferença fenotípica, o que pode ser corroborado pela maior expressão do gene *ycdH*. Curiosamente, nos ensaios de sobrevivência aos metais pesados testados nas curvas, a linhagem VNG_sot0042 demonstrou maior número de células viáveis em todas condições testadas, sugerindo que o fenômeno de interrupção de crescimento como

maneira de proteção celular é mais intenso nesta linhagem. A maior expressão de *yigB* pode ajudar a corroborar o fenômeno de maior produção de células viáveis.

Além disso, a interação de VNG_sot0042 com YpjA, a maior expressão da adesina *ycdH* e dos genes associados a biossíntese de purinas indicaram que este sotRNA poderia afetar a capacidade de adesão de *H. salinarum*. Ensaios de adesão a vidro confirmaram nossas suspeitas demonstrando que a linhagem de superexpressão tem maior capacidade de adesão.

5) Conclusão

A prevalência de elementos móveis nos três domínios da vida levanta diversas questões relacionadas aos seus impactos evolutivos. Estes elementos já foram associados a diversos processos importantes, como transferência lateral de genes e em alguns casos se comportam como reservatório de ncRNAs (DE SOUZA; FRANCHINI; RUBINSTEIN, 2013). Nas últimas décadas, os ncRNAs derivados de TEs tem sido alvos de estudos sistemáticos e seus impactos na regulação da expressão gênica elucidados (MCCUE; SLOTKIN, 2012). Os mecanismos de ação destes ncRNAs são bem diversificados, podendo envolver a estabilização, degradação ou inibição da tradução de um mRNA alvo (ELLIS; HANIFORD, 2016). Além disso, a presença de ncRNAs derivados de TEs evidencia que suas interações com o genoma hospedeiro podem ser mais complexas do que previamente avaliadas. Uma das classes mais importantes de TEs presente em procariotos, são as sequências de inserção (IS). Estes elementos altamente diversificados são extremamente profusos tanto em Bacteria quanto em Archaea. Um organismo muito atrativo para estudos de análise e caracterização de novos transcritos associados a TEs é a archaea halofílica *H. salinarum*. Além de pertencer a um domínio da vida pouco caracterizado e ser de fácil manipulação em laboratório, esse organismo apresenta número elevado de ISs em seu genoma. Ademais, uma grande variedade de dados públicos de transcrito, análises de interação proteína-DNA e proteômica estão disponíveis.

Diferentemente de asRNAs, os quais são amplamente estudados e descritos na literatura, transcritos sobrepostos na mesma fita ainda apresentam-se como grande desafio para caracterização e detecção. Neste trabalho, através da análise integrada de dados públicos de transcrito identificamos 10 RNAs sobrepostos ao 3' de genes *tnpB* de IS200/605 em *H. salinarum* NRC-1, estes transcritos foram chamados de *sense overlapping transcripts* (sotRNAs).

Os sotRNAs apresentaram-se diferencialmente regulados em diversas condições ambientais e genéticas de *H. salinarum* NRC-1, argumentando a favor de sua funcionalidade. A análise de dados de transcrito de outros procariotos em conjunto com a busca por potenciais homólogos evidenciaram que sotRNAs são encontrados em diversas espécies de Archaea e Bacteria, reforçando sua conservação em procariotos. Outro fator importante dos sotRNAs é a presença de uma estrutura conservada chamada de UUCA *tetra*loop. Este motivo

foi chamado de *RE-like* devido a alta similaridade com o motivo RE encontrado em IS200/605. Utilizando a sequência consenso desse motivo para buscar no genoma de *H. salinarum* NRC-1, encontramos novos ncRNAs intergênicos derivados de IS200/605 contendo o motivo *RE-like*. Curiosamente, a região onde encontramos o ncRNA VNG_R0052 é originada através de um evento de deleção da região central de IS200/605 muito semelhante ao descrito em *Haloquadratum walsbyi* e que dá origem aos elementos chamados de PATEs (DYALL-SMITH et al., 2011).

Diferentemente de transcritos convencionais, técnicas clássicas como construção de linhagens nocaute não são aplicáveis para sotRNAs. A deleção de um destes elementos afeta diretamente a expressão do outro, tornando impossível a associação de fenótipos. Para contornar isso, realizamos experimentos de superexpressão de um dos dez sotRNAs (VNG_sot0042) identificados e um RNA derivado de PATE (VNG_R0052), ambos contendo o motivo *RE-like*. Em ambos casos, a superexpressão demonstrou ser impactante em *H. salinarum* NRC-1 ocasionando maior crescimento das linhagens de superexpressão durante uma curva de crescimento padrão, evidenciando a funcionalidade destes ncRNAs.

Para melhor elucidar o impacto da superexpressão de VNG_sot0042 e seus possíveis mecanismos de ação, fizemos a análise comparativa do transcrito das linhagens VNG_sot0042 e pMTF-0 e em paralelo a identificação de possíveis parceiros de interação proteína-RNA de VNG_sot0042. Pudemos identificar 220 genes diferencialmente expressos, 94 mais expressos e 116 menos expressos, além de 26 proteínas com potencial de interação. Integrando esses dados pudemos caracterizar fenótipos interessantes associados a superexpressão de VNG_sot0042.

Verificamos que genes importantes associados a resistência de *H. salinarum* NRC-1 frente aos metais Co, Cu, Zn, Ni e Fe encontram-se menos expressos na linhagem VNG_sot0042. Ensaios de curva de crescimento demonstraram que a presença de Fe, Co, Cu impactam negativamente o crescimento de VNG_sot0042 em relação ao controle PMTF-0. Em relação a resistência frente a Mn, encontramos genes mais expressos. Curvas de crescimento na presença deste metal demonstraram que o impacto de Mn na linhagem VNG_sot0042 não é significativo. Ensaios de sobrevivência sugerem que apesar da deficiência no crescimento apresentada nas curvas de crescimento, a linhagem VNG_sot0042 é capaz de produzir maior número de células viáveis mesmo em condições de estresse.

A identificação de membros do sistema ZnuABC, *purL* e *purC* no conjunto de genes mais expressos assim como a interação com a proteína YpjA sugeriram que o processo de adesão desta linhagem poderia estar alterado. Ensaios de adesão a vidro confirmaram nossas suspeitas, evidenciando que a superexpressão de VNG_sot0042 aumenta a capacidade de adesão de *H. salinarum* NRC-1.

Por fim, diversos estudos tem demonstrado a relação de TEs com a origem de genes de ncRNAs em eucariotos. Estima-se que em humanos, cerca de 41% dos nucleotídeos presentes em lncRNAs são derivados de TEs (KANNAN et al., 2015). Em procariotos, estudos que identificam e caracterizam ncRNAs derivados de TEs são escassos. Neste trabalho, a identificação de novas classes de RNAs (sotRNAs e PATE-derived RNAs) associados a IS200/605 em espécies de Bacteria e Archaea, em conjunto a caracterização funcional do sotRNA VNG_sot0042, podem ser fundamentais para compreender a manutenção de ISs e seus derivados nos genomas procarióticos.

6) Referências Bibliográficas

- AKAMA, T. et al. Whole-genome tiling array analysis of *Mycobacterium leprae* RNA Reveals High Expression of Pseudogenes and Noncoding Regions. **Journal of Bacteriology**, v. 191, n. 10, p. 3321–3327, 2009.
- ALLERS, T.; MEVARECH, M. Archaeal genetics - the third way. **Nature reviews. Genetics**, v. 6, n. 1, p. 58–73, 2005.
- AMMAN, F. et al. TSSAR: TSS annotation regime for dRNA-seq data. **BMC bioinformatics**, v. 15, p. 89, 2014.
- ANDERS, S.; PYL, P. T.; HUBER, W. HTSeq-A Python framework to work with high-throughput sequencing data. **Bioinformatics**, v. 31, n. 2, p. 166–169, 2015.
- AOKI-KINOSHITA, K. F.; KANEHISA, M. Gene annotation and pathway mapping in KEGG. **Methods in Molecular Biology**, v. 396, p. 71–91, 2007.
- BAO, W.; JURKA, J. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. **Mobile DNA**, v. 4, n. 1, p. 12, 2013.
- BEER, K. D. et al. Model organisms retain an “ecological memory” of complex ecologically relevant environmental variation. **Applied and Environmental Microbiology**, v. 80, n. 6, p. 1821–1831, 2014.
- BERTELS, F.; RAINEY, P. B. Within-Genome evolution of REPINs: A new family of miniature mobile DNA in bacteria. **PLoS Genetics**, v. 7, n. 6, 2011.
- BILUSIC, I. et al. Revisiting the coding potential of the *E. coli* genome through Hfq co-immunoprecipitation. **RNA biology**, v. 11, n. 5, p. 641–54, 2014.
- BINDEWALD, E.; KLUTH, T.; SHAPIRO, B. A. CyloFold: Secondary structure prediction including pseudoknots. **Nucleic Acids Research**, v. 38, n. SUPPL. 2, 2010.
- BOLHUIS, A. Protein transport in the halophilic archaeon *Halobacterium* sp. NRC-1: A major role for the twin-arginine translocation pathway? **Microbiology**, v. 148, n. 11, p. 3335–3346, 2002.
- BROOUN, A. et al. An archaeal aerotaxis transducer combines subunit I core structures of eukaryotic cytochrome c oxidase and eubacterial methyl-accepting chemotaxis proteins. **Journal of Bacteriology**, v. 180, n. 7, p. 1642–1646, 1998.
- BUTTER, F. et al. Unbiased RNA-protein interaction screen by quantitative proteomics. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 26, p. 10626–10631, 2009.

- CAPRA, E. J.; LAUB, M. T. Evolution of Two-Component Signal Transduction Systems. **Annual Review of Microbiology**, v. 66, n. 1, p. 325–347, 2012.
- CASSIDAY, L. A. Having it both ways: transcription factors that bind DNA and RNA. **Nucleic Acids Research**, v. 30, n. 19, p. 4118–4126, 2002.
- CAVICCHIOLI, R. Archaea-timeline of the third domain. **Nature reviews. Microbiology**, v. 9, n. 1, p. 51–61, 2011.
- CHACONAS, G. Mobile DNA II - Nancy L. Craig - Google Books. **Mobile DNA II**, 2002.
- CHANDLER, M.; FAYET, O. **Translational frameshifting in the control of transposition in bacteria** *Molecular Microbiology*, 1993.
- CHANDLER, M.; MAHILLON, J. Insertion Sequences Revisited. **Mobile DNA II**, v. 62, p. 305–366, 2002.
- CHEN, S.; CALVO, J. M. Leucine-induced dissociation of Escherichia coli Lrp hexadecamers to octamers. **Journal of Molecular Biology**, v. 318, n. 4, p. 1031–1042, 2002.
- CLARK, M. B. et al. The reality of pervasive transcription. **PLoS Biology**, v. 9, n. 7, 2011.
- CLARK, M. B. et al. The dark matter rises: the expanding world of regulatory RNAs. **Essays in biochemistry**, v. 54, p. 1–16, 2013.
- COKER, J. A et al. Transcriptional profiling of the model Archaeon Halobacterium sp. NRC-1: responses to changes in salinity and temperature. **Saline systems**, v. 3, n. February 2007, p. 6, 2007.
- COREY, R. A. et al. Unlocking the Bacterial SecY Translocon. **Structure**, v. 24, n. 4, p. 518–527, 2016.
- CORVEC, S. et al. AmpC cephalosporinase hyperproduction in Acinetobacter baumannii clinical strains. **Journal of Antimicrobial Chemotherapy**, v. 52, n. 4, p. 629–635, 2003.
- CSERZO, M. et al. Relating underrepresented genomic DNA patterns and tRNAs: the rule behind the observation and beyond. **Biology direct**, v. 5, p. 56, 2010.
- DASSARMA, S.; DASSARMA, P. Halophiles. **eLS**, 2012.
- DAVID, L. et al. A high-resolution map of transcription in the yeast genome. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 14, p. 5320–5, 2006.
- DAWSON, W. et al. A method for finding optimal rna secondary structures using a new entropy model (vsfold). **Nucleosides, nucleotides & nucleic acids**, v. 25, n. 2, p. 171–89, 2006.

- DE SOUZA, F. S. J.; FRANCHINI, L. F.; RUBINSTEIN, M. Exaptation of transposable elements into novel Cis-regulatory elements: Is the evidence always strong? **Molecular Biology and Evolution**, v. 30, n. 6, p. 1239–1251, 2013.
- DELIHAS, N. Stem loop sequences specific to transposable element IS605 are found linked to lipoprotein genes in *Borrelia* plasmids. **PLoS ONE**, v. 4, n. 11, 2009.
- DELONG, E. F. Everything in moderation: Archaea as “non-extremophiles”. **Current Opinion in Genetics and Development**, v. 8, n. 6, p. 649–654, 1998.
- DJEBALI, S. et al. Landscape of transcription in human cells. **Nature**, v. 489, n. 7414, p. 101–108, 2012.
- DORNENBURG, J. E. et al. Widespread antisense transcription in *Escherichia coli*. **mBio**, v. 1, n. 1, 2010.
- DYALL-SMITH, M. The Halohandbook: protocols for halobacterial genetics. **Mark Dyall-Smith, Martinsried, Germany**, n. March, p. 1–144, 2008.
- DYALL-SMITH, M. L. et al. *Haloquadratum walsbyi* : Limited Diversity in a Global Pond. **PLoS ONE**, v. 6, n. 6, p. e20968, 2011.
- ELLIS, M. J.; HANIFORD, D. B. Riboregulation of bacterial and archaeal transposition. **Wiley Interdisciplinary Reviews: RNA**, v. 7, n. 3, p. n/a-n/a, 2016.
- ELLIS, M. J.; TRUSSLER, R. S.; HANIFORD, D. B. A cis-encoded sRNA, Hfq and mRNA secondary structure act independently to suppress IS200 transposition. **Nucleic Acids Research**, v. 43, n. 13, p. 6511–6527, 2015.
- ERWIN, J. A; MARCHETTO, M. C.; GAGE, F. H. Mobile DNA elements in the generation of diversity and complexity in the brain. **Nature reviews. Neuroscience**, v. 15, n. 8, p. 497–506, 2014.
- ESQUIVEL, R. N.; XU, R.; POHLSCHRODER, M. Novel archaeal adhesion pilins with a conserved N terminus. **Journal of bacteriology**, v. 195, n. 17, p. 3808–18, 2013.
- FACCIOTTI, M. T. et al. General transcription factor specified global gene regulation in archaea. **Proceedings of the National Academy of Sciences of the United States of America**, v. 104, p. 4630–4635, 2007.
- FEDOROFF, N.; WESSLER, S.; SHURE, M. Isolation of the transposable maize controlling elements Ac and Ds. **Cell**, v. 35, n. 1, p. 235–242, 1983.
- FILÉE, J.; SIGUIER, P.; CHANDLER, M. Insertion sequence diversity in archaea. **Microbiology and molecular biology reviews : MMBR**, v. 71, n. 1, p. 121–157, 2007.
- FLAVELL, R. B. et al. Genome size and the proportion of repeated nucleotide sequence DNA

in plants. **Biochemical Genetics**, v. 12, n. 4, p. 257–269, 1974.

FOSTER, T. J. et al. Genetic organization of transposon Tn10. **Cell**, v. 23, n. 1, p. 201–213, 1981.

GABBIANELLI, R. et al. Role of ZnuABC and ZinT in Escherichia coli O157:H7 zinc acquisition and interaction with epithelial cells. **BMC microbiology**, v. 11, n. 1, p. 36, 2011.

GARCIA, A D. et al. Bacterial-type DNA holliday junction resolvases in eukaryotic viruses. **Proceedings of the National Academy of Sciences of the United States of America**, v. 97, n. 16, p. 8926–31, 2000.

GEBERT, D.; ROSENKRANZ, D. RNA-based regulation of transposon expression. **Wiley Interdisciplinary Reviews: RNA**, v. 6, n. December, 2015.

GEGNER, J. A. et al. Assembly of an MCP receptor, CheW, and kinase CheA complex in the bacterial chemotaxis signal transduction pathway. **Cell**, v. 70, n. 6, p. 975–982, 1992.

GERSTEIN, M. B. et al. Comparative analysis of the transcriptome across distant species. **Nature**, v. 512, n. 7515, p. 445–448, 2014.

GHOSH, M.; SONAWAT, H. M. Kreb's TCA cycle in Halobacterium salinarum investigated by ¹³C nuclear magnetic resonance spectroscopy. **Extremophiles**, v. 2, n. 4, p. 427–433, 1998.

GIRARD, A.; HANNON, G. J. Conserved themes in small-RNA-mediated transposon control. **Trends in Cell Biology**, v. 18, n. 3, p. 136–148, 2008.

GOMES-FILHO, J. V. et al. Sense overlapping transcripts in IS 1341 -type transposase genes are functional non-coding RNAs in archaea. **RNA Biology**, v. 12, n. 5, p. 490–500, 4 maio 2015.

GOTTESMAN, S.; STORZ, G. Bacterial small RNA regulators: Versatile roles and rapidly evolving variations. **Cold Spring Harbor Perspectives in Biology**, v. 3, n. 12, 2011.

GOULD, S.; VRBA, E. **Exaptation—a missing term on the science of form** *Paleobiology*, 1982.

GRUBER, A. R. et al. The Vienna RNA websuite. **Nucleic acids research**, v. 36, n. Web Server issue, 2008.

GUO, X. et al. Characterization of Human Pseudogene- Derived Non- Coding RNAs for Functional Potential. **PLOS ONE**, v. 9, n. 4, 2014.

HA, M.; KIM, V. N. Regulation of microRNA biogenesis. **Nature reviews. Molecular cell biology**, v. 15, n. 8, p. 509–524, 2014.

- HANSEN, S.; LEWIS, K.; VULIĆ, M. Role of global regulators and nucleotide metabolism in antibiotic tolerance in *Escherichia coli*. **Antimicrobial Agents and Chemotherapy**, v. 52, n. 8, p. 2718–2726, 2008.
- HARROW, J. et al. GENCODE: The reference human genome annotation for the ENCODE project. **Genome Research**, v. 22, n. 9, p. 1760–1774, 2012.
- HE, Q. et al. Identification, Diversity and Evolution of MITEs in the Genomes of Microsporidian *Nosema* Parasites. **Plos One**, v. 10, n. 4, p. e0123170, 2015a.
- HE, S. et al. IS200/IS605 family single-strand transposition: Mechanism of IS608 strand transfer. **Nucleic Acids Research**, v. 41, n. 5, p. 3302–3313, 2013.
- HE, S. et al. The IS200/IS605 Family and “Peel and Paste” Single-strand Transposition Mechanism. **Microbiology spectrum**, v. 3, n. 4, p. 1–21, 2015b.
- HICKEY, A. J.; DE MACARIO, E. C.; MACARIO, A. J. L. Transcription in the Archaea: Basal Factors, Regulation, and Stress Gene Expression. **Critical Reviews in Biochemistry and Molecular Biology**, v. 37, n. 4, p. 199–258, 2002.
- HO, P. S. Detailed mechanism for transposition by TnpA transposase involves DNA shape rather than direct protein-DNA recognition to generate an active nucleoprotein complex. **Biology Reports**, v. 13410, p. 37–1, 2009.
- INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. **Nature**, v. 431, p. 931–945, 2004.
- JÄGER, D. et al. Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis*. **BMC genomics**, v. 15, n. 1, p. 684, 2014.
- JAIN, S.; CAFORIO, A.; DRIESSEN, A. J. M. **Biosynthesis of archaeal membrane ether lipids** *Frontiers in Microbiology*, 2014.
- JANKOWSKY, E.; HARRIS, M. E. Specificity and nonspecificity in RNA–protein interactions. **Nature Reviews Molecular Cell Biology**, v. 16, n. 9, p. 533–544, 2015.
- JOHNSON, R.; GUIGÓ, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. **Rna**, v. 20, n. 7, p. 959–76, 2014.
- JORDAN, I. K. et al. Origin of a substantial fraction of human regulatory sequences from transposable elements. **Trends in Genetics**, v. 19, n. 2, p. 68–72, 2003.
- KANNAN, S. et al. Transposable Element Insertions in Long Intergenic Non-Coding RNA Genes. **Front Bioeng Biotechnol**, v. 3, n. June, p. 71, 2015.
- KAPITONOV, V. V.; MAKAROVA, K. S.; KOONIN, E. V. ISC, a novel group of bacterial

and archaeal DNA transposons that encode Cas9 homologs. **Journal of Bacteriology**, v. 198, n. 5, p. 797–807, 2016.

KAUFFMAN, S. A. {C}ellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. **{J}ournal of {C}ybernetics**, v. 1, n. 1, p. 71–96, 1971.

KAUR, A. et al. A systems view of haloarchaeal strategies to withstand stress from transition metals. p. 841–854, 2006a.

KAUR, A. et al. A systems view of haloarchaeal strategies to withstand stress from transition metals A systems view of haloarchaeal strategies to withstand stress from transition metals. **Genome Research**, v. 16, p. 841–854, 2006b.

KENNEDY, S. P. et al. Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. **Genome research**, v. 11, n. 10, p. 1641–1650, out. 2001.

KIM, J. M. et al. Transposable Elements and Genome Organization : A Comprehensive Survey of Retrotransposons Revealed by the Complete Saccharomyces cerevisiae Genome ? Sequence Transposable Elements and Genome Organization : A Comprehensive Survey of Retrotransposons Revea. n. 515, p. 464–478, 1998.

KOIDE, T. et al. Prevalence of transcription promoters within archaeal operons and coding sequences. **Molecular systems biology**, v. 5, n. 285, p. 285, 2009.

KOKOEVA, M. V.; OESTERHELT, D. BasT, a membrane-bound transducer protein for amino acid detection in Halobacterium salinarum. **Molecular Microbiology**, v. 35, n. 3, p. 647–656, 2000.

KOONIN, E. V. Carl Woese’s vision of cellular evolution and the domains of life. **RNA Biology**, v. 11, n. 3, p. 197–204, 16 mar. 2014.

KRÖGER, C. et al. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. **Proceedings of the National Academy of Sciences**, v. 109, n. 20, p. E1277–E1286, 2012.

LAITY, J. H.; LEE, B. M.; WRIGHT, P. E. **Zinc finger proteins: New insights into structural and functional diversity** *Current Opinion in Structural Biology*, 2001.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 2001.

LE NOVÈRE, N. Quantitative and logic modelling of molecular and gene networks. **Nature reviews. Genetics**, v. 16, n. 3, p. 146–58, 2015.

LEE, J. T.; BARTOLOMEI, M. S. **X-inactivation, imprinting, and long noncoding RNAs**

in health and diseaseCell, 2013.

LEGERME, G. et al. Screening of a Haloferax volcanii Transposon Library Reveals Novel Motility and Adhesion Mutants. **Life**, v. 6, n. 4, p. 41, 2016.

LIANG, W.; DEUTSCHER, M. P. REP sequences: Mediators of the environmental stress response? **RNA biology**, v. 13, n. 2, p. 152–6, 2016.

LOSENSKY, G. et al. Shedding light on biofilm formation of Halobacterium salinarum R1 by SWATH-LC/MS/MS analysis of planktonic and sessile cells. **Proteomics**, v. 17, n. 7, abr. 2017.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome biology**, v. 15, n. 12, p. 550, 2014.

LUEHRSEN, K. R. Remembering Carl Woese. **RNA Biology**, v. 11, n. 3, p. 217–219, 7 mar. 2014.

MA, CHARLES AND SIMONS, R. W. The IS10 antisense RNA blocks ribosome binding at the transposase translation initiation site. **The EMBO journal**, v. 9, n. 4, p. 1267–1274, 1990.

MAHILLON, J.; CHANDLER, M. Insertion sequences. **Microbiol Mol Biol Rev**, v. 62, n. 3, p. 725–774, 1998.

MARCHLER-BAUER, A. et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. **Nucleic acids research**, p. gkw1129, 2016.

MCCLINTOCK, B. The Stability of Broken Ends of Chromosomes in Zea Mays. **Genetics**, v. 26, n. 2, p. 234–282, 1941.

MCCUE, A. D.; SLOTKIN, R. K. **Transposable element small RNAs as regulators of gene expression**Trends in Genetics, 2012.

MILLER, W. J.; CAPY, P. as Natural Tools for Genome Evolution. **Elements**, v. 260, p. 1–20, 2004.

MUGNIER, P. D.; POIREL, L.; NORDMANN, P. Functional analysis of insertion sequence ISAbal, responsible for genomic plasticity of acinetobacter baumannii. **Journal of Bacteriology**, v. 191, n. 7, p. 2414–2418, 2009.

MURO, E. M.; ANDRADE-NAVARRO, M. A. Pseudogenes as an alternative source of natural antisense transcripts. **BMC evolutionary biology**, v. 10, n. 1, p. 338, 2010.

NAGY, Z.; CHANDLER, M. Regulation of transposition in bacteria. **Research in Microbiology**, v. 155, n. 5, p. 387–398, 2004.

NANJUNDIAH, V. Barbara McClintock and the discovery of jumping genes. **Resonance**, v.

1, n. 10, p. 56–62, 1996.

NEGI, P.; RAI, A. N.; SUPRASANNA, P. Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response. **Frontiers in Plant Science**, v. 7, n. October, p. 1448, 2016.

NG, W. V et al. Genome sequence of Halobacterium species NRC-1. **Proceedings of the National Academy of Sciences of the United States of America**, v. 97, n. 22, p. 12176–12181, 2000.

ORELL, A.; FRÖLS, S.; ALBERS, S.-V. Archaeal Biofilms: The Great Unexplored. **Annual Review of Microbiology**, v. 67, n. 1, p. 337–354, 2013.

OREN, A. Halophilic microbial communities and their environments. **Current Opinion in Biotechnology**, v. 33, n. Figure 1, p. 119–124, 2015.

OREN, A.; TRUPER, H. G. Anaerobic growth of halophilic archaeobacteria by reduction of dimethylsulfoxide and trimethylamine N-oxide. **FEMS Microbiology Letters**, v. 70, n. 1, p. 33–36, 15 jun. 1990.

OSTROM, J. H. ARCHAEOPTERYX AND THE ORIGIN OF FLIGHT. **Chicago Journals**, v. 49, n. 1, p. 27–47, 1974.

PANG, Y. L. J. et al. Diverse cell stresses induce unique patterns of tRNA up- and down-regulation: tRNA-seq for quantifying changes in tRNA copy number. **Nucleic Acids Research**, v. 42, n. 22, p. 1–10, 2014.

PASTERNAK, C. et al. ISDra2 transposition in *Deinococcus radiodurans* is downregulated by TnpB. **Molecular microbiology**, v. 88, n. 2, p. 443–55, 2013.

PFEIFFER, F. et al. Evolution in the laboratory: The genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1. **Genomics**, v. 91, n. 4, p. 335–346, 2008.

PHOK, K. et al. Identification of CRISPR and riboswitch related RNAs among novel noncoding RNAs of the euryarchaeon *Pyrococcus abyssi*. **BMC genomics**, v. 12, n. 1, p. 312, 2011.

PIRIYAPONGSA, J.; MARIÑO-RAMÍREZ, L.; JORDAN, I. K. Origin and evolution of human microRNAs from transposable elements. **Genetics**, v. 176, n. 2, p. 1323–1337, 2007.

PLAISIER, C. L. et al. Evolution of context dependent regulation by expansion of feast/famine regulatory proteins. **BMC systems biology**, v. 8, p. 122, 2014.

RAMIREZ-PEÑA, E. et al. The group A *Streptococcus* small regulatory RNA FasX enhances streptokinase activity by increasing the stability of the ska mRNA transcript. **Molecular Microbiology**, v. 78, n. 6, p. 1332–1347, 2010.

- REES, D. C.; JOHNSON, E.; LEWINSON, O. ABC transporters: the power to change. **Nature reviews. Molecular cell biology**, v. 10, n. 3, p. 218–27, 2009.
- REIMMANN, C. et al. Genetic structure, function and regulation of the transposable element IS21. **Molecular & general genetics : MGG**, v. 215, n. 3, p. 416–24, 1989.
- RONNING, D. R. et al. Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. **Molecular Cell**, v. 20, n. 1, p. 143–154, 2005.
- ROSS, J. A.; WARDLE, S. J.; HANIFORD, D. B. Tn10/IS10 transposition is downregulated at the level of transposase expression by the RNA-binding protein Hfq. **Molecular Microbiology**, v. 78, n. 3, p. 607–621, 2010.
- ROUX, A.; BELOIN, C.; GHIGO, J. Combined Inactivation and Expression Strategy To Study Gene Function under Physiological Conditions : Application to Identification of New Escherichia coli Adhesins. **Journal of Bacteriology**, v. 187, n. 3, p. 1001–1013, 2005.
- RUEPP, A.; SOPPA, J. Fermentative arginine degradation in Halobacterium salinarum (formerly Halobacterium halobium): Genes, gene products, and transcripts of the arcRACB gene cluster. **Journal of Bacteriology**, v. 178, n. 16, p. 4942–4947, 1996.
- SAWERS, R. G. Transcript analysis of Escherichia coli K-12 insertion element IS5. **FEMS Microbiology Letters**, v. 244, n. 2, p. 397–401, 2005.
- SCHÄFER, G.; ENGELHARD, M.; MÜLLER, V. Bioenergetics of the Archaea. **Microbiology and molecular biology reviews : MMBR**, v. 63, n. 3, p. 570–620, 1999.
- SCHÄFER, G.; PURSCHKE, W.; SCHMIDT, C. L. **On the origin of respiration: Electron transport proteins from archaea to man.** FEMS Microbiology Reviews. **Anais...**1996
- SCHINDELIN, J. et al. The ImageJ ecosystem: An open platform for biomedical image analysis. **Molecular Reproduction and Development**, v. 82, n. 7–8, p. 518–529, 2015.
- SCHLESNER, M. et al. The protein interaction network of a taxis signal transduction system in a halophilic archaeon. **BMC microbiology**, v. 12, p. 272, 2012.
- SCHU, D. J. et al. Alternative Hfq-sRNA interaction modes dictate alternative mRNA recognition. v. 34, n. 20, p. 2557–2573, 2015.
- SCHWAIGER, R. et al. Transcriptional control by two leucine-responsive regulatory proteins in Halobacterium salinarum R1. **BMC molecular biology**, v. 11, n. 1, p. 40, 2010.
- SEILA, A. C. et al. Divergent transcription from active promoters. **Science (New York, NY)**, v. 322, n. 5909, p. 1849–1851, 2008.
- SHARMA, C. M. et al. The primary transcriptome of the major human pathogen Helicobacter pylori. **Nature**, v. 464, n. 7286, p. 250–255, 2010.

- SHERWOOD, K. E.; CANO, D. J.; MAUPIN-FURLOW, J. A. Glycerol-mediated repression of glucose metabolism and glycerol kinase as the sole route of glycerol catabolism in the haloarchaeon *Haloferax volcanii*. **Journal of Bacteriology**, v. 191, n. 13, p. 4307–4315, 2009.
- SHIGA, Y. et al. Involvement of H-NS in transpositional recombination mediated by IS1. **Journal of Bacteriology**, v. 183, n. 8, p. 2476–2484, 2001.
- SIGUIER, P. et al. ISfinder: the reference centre for bacterial insertion sequences. **Nucleic acids research**, v. 34, n. Database issue, p. D32-6, 2006.
- SIGUIER, P. et al. Everyman's Guide to Bacterial Insertion Sequences. **Microbiology Spectrum**, p. 1–35, 2015a.
- SIGUIER, P. et al. Everyman's Guide to Bacterial Insertion Sequences. p. 1–35, 2015b.
- SIGUIER, P.; GOURBEYRE, E.; CHANDLER, M. Bacterial insertion sequences: Their genomic impact and diversity. **FEMS Microbiology Reviews**, v. 38, p. 865–891, 2014.
- SRIVASTAVA, P.; KOWSHIK, M. **Mechanisms of metal resistance and homeostasis in Haloarchaea**. *Archaea*, 2013.
- STEPHEN, S. et al. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. **Molecular Biology and Evolution**, v. 25, n. 2, p. 402–408, 2008.
- STORZ, G.; ALTUVIA, S.; WASSARMAN, K. M. An abundance of RNA regulators. **Annual review of biochemistry**, v. 74, p. 199–217, 2005.
- TANG, T. H. et al. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. **Molecular Microbiology**, v. 55, n. 2, p. 469–481, 2005.
- TAYLOR, A. L. Bacteriophage-induced mutation in *Escherichia coli*. **Proceedings of the National Academy of Sciences**, v. 50, p. 1043–1051, 1963.
- TER BEEK, J.; GUSKOV, A.; SLOTBOOM, D. J. Structural diversity of ABC transporters. **The Journal of general physiology**, v. 143, n. 4, p. 419–35, 2014.
- TURKARSLAN, S. et al. Niche adaptation by expansion and reprogramming of general transcription factors. **Molecular systems biology**, v. 7, n. 554, p. 554, 2011.
- VARANI, A. M. et al. ISSaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. **Genome biology**, v. 12, n. 3, p. R30, 2011.
- VILAIN, S. et al. DNA as an adhesin: *Bacillus cereus* requires extracellular DNA to form biofilms. **Applied and Environmental Microbiology**, v. 75, n. 9, p. 2861–2868, 2009.

- WANG, P. L. et al. Circular RNA Is Expressed across the Eukaryotic Tree of Life. **PLoS ONE**, v. 9, n. 3, p. e90859, 7 mar. 2014.
- WATSON, J. D.; CRICK, F. H. **Genetical implications of the structure of deoxyribonucleic acid.***Nature*, 1953.
- WEINREICH, M. D.; REZNIKOFF, W. S. Fis plays a role in Tn5 and IS50 transposition. **Journal of Bacteriology**, v. 174, n. 14, p. 4530–4537, 1992.
- WHITE, M. F.; BELL, S. D. **Holding it together: Chromatin in the Archaea***Trends in Genetics*, 2002.
- WOESE, C. R.; FOX, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 11, p. 5088–5090, 1977.
- WRIGHT, P. R. et al. **Freiburg RNA Tools webserver**. Poster, 3rd Conference on Regulating with RNA in Bacteria. **Anais...**2013a
- WRIGHT, P. R. et al. Comparative genomics boosts target prediction for bacterial small RNAs. **Proceedings of the National Academy of Sciences**, v. 110, n. 37, p. E3487–E3496, 2013b.
- WRIGHT, P. R. et al. CopraRNA and IntaRNA: Predicting small RNA targets, networks and interaction domains. **Nucleic Acids Research**, v. 42, n. W1, 2014.
- WURTMANN, E. J. et al. An evolutionarily conserved RNase-based mechanism for repression of transcriptional positive autoregulation. **Molecular Microbiology**, v. 92, n. 2, p. 369–382, 2014.
- XU, J. et al. Enhancement in ionic liquid tolerance of cellulase immobilized on PEGylated graphene oxide nanosheets: Application in saccharification of lignocellulose. **Bioresource Technology**, v. 200, p. 1060–1064, 2016.
- YAN, Y. et al. Determination of sRNA expressions by RNA-seq in *Yersinia pestis* grown in vitro and during infection. **PloS one**, v. 8, n. 9, p. e74495, 2013.
- YING, M. et al. Origin and evolution of ubiquitin-conjugating enzymes from *Guillardia theta* nucleomorph to hominoid. **Gene**, v. 447, n. 2, p. 72–85, 2009.
- ZAKOV, S. et al. **Rich parameterization improves RNA structure prediction**. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **Anais...**2011
- ZARAMELA, L. S. et al. Transcription Start Site Associated RNAs (TSSaRNAs) Are Ubiquitous in All Domains of Life. **PLoS ONE**, v. 9, n. 9, p. e107680, 2014.

ZHANG, Y.; MORAR, M.; EALICK, S. E. **Structural biology of the purine biosynthetic pathway** *Cellular and Molecular Life Sciences*, 2008.

ZUCKERKANDL, E.; PAULING, L. Molecules as Documents of Evolutionary History By E lile Zuckerkandl and Linus Pauling Gates and Crellin Uboratories of Chemistry California of Technology Contribution. **Gates and Crellin Laboratories of Chemistry Institute of Technology**, v. 8, n. 3041, p. 357–366, 1965.

7) Anexos

ANEXO A- Homólogos do sotRNA VNG_sot0042 encontrados utilizando a ferramenta GlassGo

>VNG_sot0042-Halobacterium sp NRC-1

ACGTATCTCCGAGTCCCGTCACGGATGGCGGTGATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTTGACAATG
AAACTGGTGCCTTCGCACCTCAAGAACAGGCTACGTCGTAAACCATAATATCCCAACGCGGTTCGGAATCCTCGCCCTTCAGGG
CGGGGAGGATGTCAATGCTGCAGTCTCAGCTATCCTGT

>gi|284017284|gb|CP001862.1|:c383031-382825 Haloterrigena turkmenica DSM 5511 plasmid pHTUR02, complete sequence-taxID:543526
GGTACTCCCGAATCTCGCCACGGATGGCGGCGATAGGGATAACGGCTGGTTGGCACAGCCAACGGTTCACCTGTTTCGACCGTA
GTGGAGGACGTTTCGCCCCACGAGAACAGGCCGTGAACCGCGAACCGTAATATCCCAACGCGGTTCGGAATCCTCGCCCTTT
AGGGCGGGGAGGATGTCAACAAAGCGTCCATCAACCCGTCA

>gi|631803440|gb|CP007551.1|:c446099-445892 Haloferax mediterranei ATCC 33500 genome-taxID:523841

TCAGAAGATAACTCCAAGTCTCATGGAGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGACATACTTGTTCGACCGCG
AGAGCGGGACGTTTCGCACGAGAGAACAGGTCTGTCTGTAGACCGGCAAATATCCTACCTGCGGTACGGAAGCCCCGTCGTT
TACGACGGGGAGGATGTACCCGTTGCTTGGAGCGGCGGTTGC

>gi|284017284|gb|CP001862.1|:188701-188901 Haloterrigena turkmenica DSM 5511 plasmid pHTUR02, complete sequence-taxID:543526
AAGAAAGATAACTCAGAGTCACCTAACCGGGGATATGAATAATGGTCGGTTGGCACGGCCAGTAGCCTACCTGTTCAACCAACC
CTCGGGGCGTTTTCACCGAGCGAACAGGCAGGTTGCAAACCATAATATCCCACTGCTTGGGATTCCCACGCTTCAAGCGCGG
GAGGAGGTCAATCAGGTAGAGAGAATTTGATCA

>gi|433304139|gb|CP003372.1|:1922935-1923133 Natrinema pellirubrum DSM 15624, complete genome-taxID:797303

TCGAAAGATAACTCAGAGTCTTCGACAGGGGATATGAGTAAAGTTGGTTGGTACAGCCCGGTGCTTCGTGTTTGACTGGGAG
AGTGGGACGTTCAAGACGAGAGAACAGGGAGTCTGCAACCTTACTATCCCAACGCTCGGAACTTCTCCGCCTTCAACCGCGCG
AGGATGTCAAAAAACATTCCTTGGTGTGAT

>gi|433304139|gb|CP003372.1|:1959946-1960149 Natrinema pellirubrum DSM 15624, complete genome-taxID:797303

TAACTCCGAGTCTGCACCCAGTTTGGGTGGAGATAGGAGTACCGGCTGGTTGGCACAGCCCGAAGTCTACCTTCATAGCCTCTC
CTGCGGATTCCAACCGCAGACAGAGGTGGTGGACTGCAAAACCCTAATATCCCAACCGCGGTTCGGGATTCTCCGCCTTCAGGC
GGAGGAGGATGTCAATCAGAGTAGTGGGTGAAAAAGC

>gi|909823395|gb|CP011947.1|:c2079643-2079439 Haloferax gibbonsii strain ARA6, complete genome-taxID:35746

TCGACAGAAAGTATCTCCGAGTCTTCCGAGGATAGGAGTAACGGCTGGTTGGCACAGCCTTCGACGTTGTTGTTGACAAGG
ACACTGGTGCCTTCGCACCTCAAGAACAGGTACGTCGTAAACCACAATATCCCAACTGCGGTGGGAATCCTCGCGCTTCAG
CGCGGGGAGGATGTCAATTGTCCGCTCTGGACTCTGGTA

>gi|222454053|gb|CP001366.1|:c214249-214048 Halorubrum lacusprofundi ATCC 49239 chromosome 2, complete sequence-taxID:416348

GATAACTCCGAGTCTCAGGAGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGGTACACTTGTTCGACCGCGAGAGCG
GGACGGTTCGCCCCGAGAGAACAGGTTGTATCGTAAACCAGTAAATATCCCACTGCGGTACGGAAGCCCCGGCGTTTACGCC
GGGGAGGATGTCACTGGGTGCTTTTCGGTGCCTCGA

>gi|222454053|gb|CP001366.1|:c229621-229414 Halorubrum lacusprofundi ATCC 49239 chromosome 2, complete sequence-taxID:416348

GTCAGAAGATAACTCCTATCCTCACGGTGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGACACACTTGTTCGATTGAGA
AAGTGGAGCGTTTCGCCCCGAGAACAGGTCTGTCTGTAGACCGGCAAATATCCCACTGCGGTACGGAAGCCCCGGCGTTTC
ACGCCGGGGAGGATGTACCTGATCGAGGAGTGCCCATG

>gi|573483907|gb|CP007059.1|:c8277-8079 Halostagnicola larsenii XH-48 plasmid, complete sequence-taxID:797299

AGTATCTCCGAGTCTCTCTGTGACGGGGGAGATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGGTCTTGTGTTGACACGG
AAACTGGTGCCTTCACACCTCAAGAACACATAACGTCGTAAACCACAATATCCCACTGCGGTACGGAAGCCCCGGCGTTTC
GAGGATGTCAACGGACGACCGTTACCGGAACGG

>gi|55229667|gb|AY596297.1|:c1398829-1398632 Haloarcula marismortui ATCC 43049 chromosome I, complete sequence-taxID:272569

AGTATCTCCGAATCCTCGTAGCGAGGATAGGAGTAACGGCTGGTTGGCACAGCCATCAACGCTCCTGTTTGACAAGGAACTG

GTGTATTTGCACCTCAAGAACAGGTAGTGTCTGTAACCCCTAATATCCCAAATCAGCGGTAGCGGTACCGTGGGATTCCCGCGCC
TTCAGGCGCTGGAGGATGTCAACTCGGTTGG

>gi|222454053|gb|CP001366.1|:205971-206172 Halorubrum lacusprofundi ATCC 49239 chromosome 2, complete sequence-taxID:416348
GATAACTCCGAGTCCTCACGGTGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGACATACTTGTTCGACCGCGAGAGCG
GGACGTTTCACACGAGAGAACAAGCCGTGTCGTAGACCAGCAAATATCCCACCTGCGGTACGGGAAGCCCCGGCGTTACAGCC
GGGGAGGATGTCACTCATACTGCGGCGTAGACGTGC

>gi|222454053|gb|CP001366.1|:238432-238635 Halorubrum lacusprofundi ATCC 49239 chromosome 2, complete sequence-taxID:416348
GATAACTCCGAGTCCTCACGGAGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGACATACTTGTTCGACCGCGAGATCG
GGACGTTTCACACGAGAGAACAAGCCGTGTCGTAGACCAGCAAATATCCCACCTGCGGTACGGGAAGCCCCGGCGTTTACGCC
GGGGAGGATGTCACTGCTGTAAGTCTCTTATGCCGAA

>gi|222451341|gb|CP001365.1|:c1684021-1683826 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-
taxID:416348
AGTATCTCCGAATCCCATGCCGGGATAGGAGTAACGGCTGATTGGCACAGCCAGTAGCCGCCCAATGTGGTAGCTACAAACCGT
AACCATCCCAACCCAGCGGTGCGGTGCCGTGGGAATCCTCGCGCTTCAGCGCGGGGAGGATGTCAAACAGTCTGGCGCTGTT
GAATTCTGTTTCATCAGACACAATTTCTGT

>gi|222451341|gb|CP001365.1|:c980403-980208 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-taxID:416348
AGTATCTCCGAATCCCATGCCGGGATAGGAGTAACGGCTGATTGGCACAGCCAGTAGCCGCCCAATGTGGTAGCTACAAACCGT
AAACATCCCAACCCAGCGGTGCGGTGCCGTGGGAATCCTCGCGCTTCAGCGCGGGGAGGATGTCAAGGGCTCGCGTCGTCGA
CTGCCGCGTCGAGTGCCTCGTAGTCGGGCT

>gi|222451341|gb|CP001365.1|:c2513640-2513439 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-
taxID:416348
GATAACTCCGAGTCCTCACGGAGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGGTACACCTGTTTCGACCGCGAGAGCG
GGACGGTTCGCCCCGAGAGAACAGGTTGTATCGTAAACCAGCAAATATCCCACCTGCGGTACGGGAAGCCCCGGCGTTTACGCC
GGGGAGGATGTCAACCCCGCGACTTCCCAGACCGT

>gi|222451341|gb|CP001365.1|:c597352-597149 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-taxID:416348
TAACTCCAAGTCTGCACCCGATTTGGGTGGAGATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTATCTTCACGACCTCTC
TCGAGGATTCCAACCTCGGACAGAGGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCCCGC
GTCTTCAGGCGCGGGAGGATGTCAAGGGAATTCTCG

>gi|222451341|gb|CP001365.1|:c1787571-1787376 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-
taxID:416348
AGTATCTCCGAATCCCATGCCGGGATAGGAGTAACGGCTGATTGGCACAGCCAGTAGCCGCCCAATGTGGTCGCTACAAACCGT
AACCATCCCAACCCAGCGGTGCGGTGCCGTGGGAATCCTCGCGCTTCAGCGCGGGGAGGATGTCAAATCGAGGGATGCGGAG
CGACGACCGCGAGCAGATGCGCTCAGCGCC

>gi|222451341|gb|CP001365.1|:495689-495884 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-taxID:416348
AGTATCTCCGAATCCCATGCCGGGATAGGAGTAACGGCTGATTGGCACAGCCAGTAGCCGCCCAATGTGGTCGCTACAAACCGC
AACCATCCCAACCCAGCGGTGCGGTGCCGTGGGAATCCTCGCGCTTCAGCGCGGGGAGGATGTCAACCGACGACGATCTCGAC
GTTTGTCTGGCTGATATCCCGTTACAGGC

>gi|222451341|gb|CP001365.1|:460228-460444 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-taxID:416348
GCCCCAGCCGTTGTTTACCGAATCTCTCGGTGAATAGGAGTAACGGCTGTTTGGCACAGCCATCGACGTTGTTTGTGACAAGG
AAACAGATGCGTTCGCACCTCAAGAACAGGTAACGTCGTAAACCAAAATATCCCAACCCAGCGGTGCGGTGCCGTGGGAATCC
TCGCGCTTCAGCGCGGGGAGGATGTCAAAAGCGATCGAATACTGCTCGGC

>gi|222451341|gb|CP001365.1|:1519779-1519974 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-
taxID:416348
AGTATCTCCGAATCCCATGCCGGGATAGGAGTAACGGCTGATTGGCACAGCCAGTAGCCGCCCAATGTGGTCGCTACAAACCGT
AACCATCCCAACCCAGCGGTGCGGTGCCGTGGGAATCCTCGCGCTTCAGCGCGGGGAGGATGTCAATCCGGGAGCGTCGGGC
GACTTGACCGCGAACCGTTTGTGAGCTGTGG

>gi|222451341|gb|CP001365.1|:1623269-1623484 Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence-

taxID:416348

AGTATCTCCGAGTTCACCGAATCTCTCGGTGAATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTGTTGTTGACAAGGA
AACTGGTGC GTTTCGCACCGCAAGAACAGGTAACGTCGTAAACCACAATATCCCAACTCAGCGGTGCAGTGCCGTGGGAATCCT
CGCGCTTCAGCGCGGGGAGGATGTCAAGCGTTCGCGAACGGCTCCAGC

>gi|222451341|gb|CP001365.1|:385595-385798 *Halorubrum lacusprofundi* ATCC 49239 chromosome 1, complete sequence-taxID:416348
TAACTCCGAGTCTGCACCGGATTTGGGTGGAGATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTATCTTCACGACCTCTC
TCGAGGATTCCAACCTCGGACAGAGGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCCCGC
GTCTTCAGCGCGGGGAGGATGTCAAACGTGCACCAAG

>gi|429134263|gb|CP003377.1|:1157284-1157492 *Natronobacterium gregoryi* SP2, complete genome-taxID:797304
AGAGATCTTAAGGTCTCTCGTCATCACGAAAGGCGCGAGCGCCTTTGAAACGACACAGCCATCGGTACACCTGTTGACCGCG
AGAGCGGGACGTTACACACGAGAGAACAGGTCTGTATCGTAGACCAGCAAATATCCCAACCTGCGGTACGGGAAGCCGCGTCTC
TCAGGGCGCGGAGGATGTACACCTGTGGCTGTGCCGCCATCGA

>gi|433671692|gb|CP003929.1|:423054-423253 *Natronococcus occultus* SP4, complete genome-taxID:694430
GAGAAAGATAACTCAGAGTCCCCGACAGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGGAGTCTTCTGTTGACCGCG
AGAGCGGACGGTTCACACGAGAGAACAGGGAGTCTGCAAACCCTAATATCCCAACGCTTGGGATTCTCCGCCTTCAGGCGG
AGGAGGATGTCAAGCTGTGTTGCGTTTCTGTGGT

>gi|109624723|emb|AM180088.1|:2392695-2392898 *Haloquadratum walsbyi* DSM 16790 complete genome-taxID:362976
TAACTCCGAGTCTGCACCTGATTTGGGTGGAGATAGGAGTACCGGCTGGTTAGCACAGCCATCAGTCTATCTTCATGACTTGTCC
CACGGATTCTACCGCGGGAACAAGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTTCGGTGCCGTGGGATTCTCCG
CCTTCAGGCGGATGGAGGATGTCAATATGGGGTGT

>gi|109624723|emb|AM180088.1|:2733171-2733368 *Haloquadratum walsbyi* DSM 16790 complete genome-taxID:362976
AGTAACTCCGAGTCTCTCGTCAGAGGATAGGAATACCGACTGGTTGGCACAGTCTTCAGTCTACCTGTATGACCCCTCTGCGGA
TTCCAACCGCAGGAGAGGTGGTGGAGTGTAGAGCTTACTATCCCCACGCTTGGGATTCTCCGCGCTCACGCAGAGGAGGAT
GTCGATTATGACACTGTGATTACACCTTTT

>gi|109624723|emb|AM180088.1|:2369814-2370019 *Haloquadratum walsbyi* DSM 16790 complete genome-taxID:362976
GGTACTCCCGAGTCTCGCCACGGATGGCGGTGATAGGGATAACGGCTGGTTGGCACAGCCAGCGGTTACCTGTTGACCGTA
GTGAGGGCTGTTTCGCCCCACGAGAACAGGTCTGGACCGCAATCTTAATATCCCAACGCGGTGCGGAATCTCGCCCTTTAG
GGCGGGGAGGATGTCAAAACACGTCTCCGACGAATATTA

>gi|109624723|emb|AM180088.1|:2948347-2948548 *Haloquadratum walsbyi* DSM 16790 complete genome-taxID:362976
GATAACTCTGAGTCTCACGGCGAGGATAGGAGTAACGGCTGTATGGCACAGCCATCGACATACTTGTTCGACCGCGAGAGCG
GGATGTTTCACACGAGAGAACAGGTCTGTCTGTAGACCGGCAAATATCCACCTGCGGTACGGGAAGCCCCATCGTTACGAC
GGGGAGGATGTACGCAGCACTCGAAGGCGAGTGAG

>gi|109624723|emb|AM180088.1|:1273474-1273679 *Haloquadratum walsbyi* DSM 16790 complete genome-taxID:362976
AGTAACTCCGAGTCTTCGGCCAGTTTGGACGAGGATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTACCTTCACGACTT
GTCCAGCGGATTCCAACCGCAGGGTCAAGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
TCCGTGTTTACGCGGAGGAGGATGTCAAACGTCTCAGTC

>gi|109624723|emb|AM180088.1|:1809395-1809610 *Haloquadratum walsbyi* DSM 16790 complete genome-taxID:362976
AGTATCTCTGAGTTCACCGAATCTGTCTCGGTGAATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTGACAAAAGA
GACTGGTGC GTTTCACCTGAAGAACGGGTTAGGTCGTAAACCACAATATCCCAACTCAGCGGTGCGGTGCCGTGGGAATCCT
CGCGCTTCAGCGCGGGGAGGATGTCAATGCGCACAGCGGTGCGGGGAAC

>gi|109624723|emb|AM180088.1|:17283-17498 *Haloquadratum walsbyi* DSM 16790 complete genome-taxID:362976
AGTATCTCTGAGTTCACCTGAATCTGTCTCGGTGAATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTGACAAGGA
GACTGGTGC GTTTCACCTGAAGAACGGGTTAGGTCGTAAACCACAATATCCCAACTCAGCGGTGCGGTGCCGTGGGAATCCT
CGCGCTTCAGCGCGGGGAGGATGTCAACCGAAGTGCTGTTGTTGATGTA

>gi|109624723|emb|AM180088.1|:2027035-2027250 *Haloquadratum walsbyi* DSM 16790 complete genome-taxID:362976
AGTATCTCTGAGTTCACCGAATCTGTCTCGGTGAATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTGACAAGGA
GACTGGTGC GTTTCACCTGAAGAACGGGTTAGGTCGTAAACCACAATATCCCAACTCAGCGGTGCGGTGCCGTGGGAATCCT

CGCGCTTCAGCGCGGGGAGGATGTCAATGACTAACAGGAGTGAGTATGC

>gi|109624723|emb|AM180088.1|:2106422-2106626 Haloquadratum walsbyi DSM 16790 complete genome-taxID:362976
AGTAACTCCCAGTTTGGCGTCCAACCTTGGACGGGGATGTGAGTGTGCGACTGGTTGGCGCAGTCCAGGGTCATCCTTCGTGGTTTA
TCAGATAGATTTTCATCGAGAAAATAAGTGGTAGAATGCAAACCCCTACTATCACAACCTGCGGTGCGAGAATAATCGCTCTTCAACG
CCGGGAGGATATCAATCTGAGAGGTCAATATGTGTG

>gi|109624723|emb|AM180088.1|:452331-452534 Haloquadratum walsbyi DSM 16790 complete genome-taxID:362976
TAACTCTGAGTCTGCACCCGATTTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCTGGAGTCTACTTTCATGACCTCTC
CCGAGGATTCCAACCTCGGATAGAGGTGGTGGACTGCAAACCGTAATATCCCAACCCAGCGGTGCGGTGCCGTGGGATTCTCTCC
GCGTTCACGCGGAGGAGGATGTCAATTCTGTAGATT

>gi|109624723|emb|AM180088.1|:c1291898-1291695 Haloquadratum walsbyi DSM 16790 complete genome-taxID:362976
TAACTCCGAGTCTGCACCCGATTTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCTGGAGTCTACCTTCATGACCTCTC
CCGAGGATTCCAACCTCGGATAGAGGTGGTGGACTGCAAACCGTAATATCCCAACCCAGCGGTGCGGTGCCGTGGGATTCTCTCC
GAGTTCACGCGGAGGAGGATGTCAATTGATAATGTC

>gi|109624723|emb|AM180088.1|:c1815524-1815321 Haloquadratum walsbyi DSM 16790 complete genome-taxID:362976
AGTATCTCCGAGTCTCGCTATGGACGACGGTGATAGGAGTAACGGCTAGTCGGCACAGCCATCGACGTACCTGTTTGACAAGGA
AATTGACTTGTTCAAGCCTCAACAACAGATAATGTCGTGAGATTCGGCGCTCTTCAGGCGCGAGAGGATGTCAAATATTTACTA
CTCATGACTTACTCTTCTAAGTGAAGCATGATTATT

>gi|109624723|emb|AM180088.1|:c2033313-2033109 Haloquadratum walsbyi DSM 16790 complete genome-taxID:362976
CTCTGAGTCTTCATCTGATTTGGGTGGGCATAGGAGTGCCGGTTGGTTGGCACAGTCCGGAGTCTCCCTTCAGGAGTTGTCCAG
CGGATTCCAACCGCAGGAAAAGTGGTAGACTGCAAACCCCTACCCTAATATCCCAACTGCGGTGCGGAATCCTCGCCGTTTCAGG
GCGGAGAGGATGTCAACTAGGTGAGTGTGATCAAA

>gi|109624723|emb|AM180088.1|:c1012067-1011870 Haloquadratum walsbyi DSM 16790 complete genome-taxID:362976
AGTACCTCCGAATCCCATATGCCGGAATAGGAGTAACGGCTGGTTGGCACAGCCAGGAGCCGTCATCGAGACGGGTACAAACC
TTGAACATCCCAATCCAGCGCGCGGTGCCGTGGGAATCTTTGCCCTTCAGGGCGGAGGGGATGTCAAATATAGGCGCGGACG
CTCATGAAGCAACTGTGGCTTATTAAGTTGCA

>gi|573482792|gb|CP007056.1|:c376821-376616 Halostagnicola larsenii XH-48 plasmid, complete sequence-taxID:797299
AGTAACTCCGAGTCTGCGTCCAGTTTGGACGAGGATAGGAGTACCGGCTGGTTGGCACAGCCCGAAGTCTACCTTCATGACCT
CTCCCGAGGATTCCGACCTCGGACAGCGGTGGGAGACTGCAAACCCCTAATATCCCAACTAGCGGTGCGGTACCGTGGGATTCT
CCGCGTCTTCAGGCGCGGGAGGATGTCAAAGATCGCGTGT

>gi|573484043|gb|CP007060.1|:617891-618098 Halobacterium sp. DL1, complete genome-taxID:751944
CAGACAAAAGATAACTCCGAATCCCTCTGTGGATAGGAGTAGCGGCTGTCTGGCACAGCCATCGACATACTTGTTTCGACCGCGA
GAGCGGGACGTTTTCACACGAGAGAACAAGCCGTGTCGTAGACCAGCAAATATCCACCTGCGGTACGGGAAGCCCCGGCGTT
CACGCGGGGAGGATGTACCTCGAATACGAACCTTGATATG

>gi|291370026|gb|CP001956.1|:1953791-1953997 Haloferax volcanii DS2, complete genome-taxID:309800
GGTACTCTCAAGTCTCGCCACGGATGGCGGTGATAGGGATAGCGGCAGGTTGGCACAGCCAGCGGTTACCTGTTTCGACCGTA
GTGAGGGCCGTTTCGCCCCACGAGAACAAGTTGCTAACCGCGAACCCTAATATCCCAACTGCGGTGCGGATTCCCGCGTTGAC
GAGACGCTCTGCGCCTCGTTTCGCACACCAGAACGCTCTGCG

>gi|640913992|gb|CP001868.2|:1299951-1300158 Haloferax mediterranei ATCC 33500, complete genome-taxID:523841
TCAGAAGATAACTCCAAGTCCTCATGGAGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGACATACTTGTTTCGACCGCG
AGAGCGGGACGTTTCGCACGAGAGAACAGGTGCTGTCGTAGACCGGCAAATATCCTACCTGCGGTACGGGAAGCCCCGTCGTT
TACGACGGGGAGGATGTACCGTTGCTTGGAGCGGCGGTTGC

>gi|291370026|gb|CP001956.1|:c1931002-1930801 Haloferax volcanii DS2, complete genome-taxID:309800
GATAACTCCGAGTCTCACGGTGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGGTACACCTGTTTCGACTCAGAGAGCG
GTTGCTTTGCACCGCGAGAACAGGTGCTGTCGTAGACTAGCAAATATCCCACTGCGGTACGGGAAGCCCCGTCGTTACGAC
GGGGAGGATGTACCTGCGGCCTCGACCCGCTCGGT

>gi|528524863|emb|HF571520.1|:c1931738-1931535 Halorhabdus tiamatea SARL4B complete genome-taxID:1033806
GTAACCTGAATCCCCATACCGAGGATAGGAGTAACGGCGGCTTGGCCCCGTCCTCAGAAATCTTTGATTTCTGATGTGCGAAC

GAGAACGGAGTTCTCGTTAACGCCAGGAGTCTGGTCATACCGACCGCCTGCAAACCTGAAACTCTCAACGTGGGAATCTTCG
CCCTTCTGGGCGGAGAGGAGGTCAACCTGTTCTGTC

>gi|631806635|gb|CP007554.1|:258952-259156 *Haloferax mediterranei* ATCC 33500 plasmid HMPLAS1, complete sequence-taxID:523841
ACGAGAGATAACTCAGAATCCCTCGACAGGGGATATGAGTAACGGTCGCTTGGCACGGCCAGCAGTCCACCTGTTCAACCAAA
CCTCGGGGCGCTTCGCACCGAGCGAACAGGCGGGTTGCGAATCTTAATATCCAATGCTCGGGAACCTCGTCGCCAGATAGCG
TCTGACGGGCAACCAGAACGCTTTGCGTTCTGGTGACG

>gi|148508047|gb|EF583986.1|:c10922-10721 Uncultured haloarchaeon clone fosmid eHwalsbyi 033 genomic sequence-taxID:160804
GATAACTCTGAGTCTCACGGCGAGGATAGGAGTAACGGCTGTATGGCACAGCCATCGACATACTTGTTCGACCGCGAGAGCG
GGATGTTTCACACGAGAGAACAGGTGCTGTCGTAGACCGGCAAATATCCCACCTGCGGTACGGGAAGCCCCATCGTTCACGAC
GGGGAGGATGTACCGCAGCACTCGAAGGCGGGTGAG

>gi|257168392|gb|CP001688.1|:1946819-1947019 *Halomicrobium mukohataei* DSM 12286, complete genome-taxID:485914
GAGAAAGATAACTCAGAGTCCCCGACGGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGAGTCTTCCTGTTTCGACCGCG
AGAGCGGACGGTTCACCGAGAGAACAGGGAGACTGTAGACCCTAATATCCCAACGCTCGGGAATCCTCGCCCTTCAGGGCG
GGGAGGATGTCAAGCGACGGCTGTGCGTCTGTCGGA

>gi|388246185|gb|CP001871.1|:c270828-270624 *Haloferax mediterranei* ATCC 33500 plasmid pHM500, complete sequence-taxID:523841
ACGAGAGATAACTCAGAATCCCTCGACAGGGGATATGAGTAACGGTCGCTTGGCACGGCCAGCAGTCCACCTGTTCAACCAAA
CCTCGGGGCGCTTCGCACCGAGCGAACAGGCGGGTTGCGAATCTTAATATCCAATGCTCGGGAACCTCGTCGCCAGATAGCG
TCTGACGGGCAACCAGAACGCTTTGCGTTCTGGTGACG

>gi|528524863|emb|HF571520.1|:1263410-1263612 *Halorhabdus tiamatea* SARL4B complete genome-taxID:1033806
CAGAAAGATAACTCAGAGTCCCCGACAGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGAGTCTTCCTGTTTCGACCGCG
AGAGCGGACGGTTCACACCGAGAGAACAGGGAGTCTGCAAACCCTAATATCCCAACGCTGAGCAAATCGGAGATTTGCGATGT
ACGCGAATCTTCGATTGCTTACTCGGGATTCTCTCC

>gi|822660691|gb|CP008874.1|:499893-500098 *Halanaeroarchaeum sulfuri-reducens* strain HSR2, complete genome-taxID:1604004
AGTAACCTCCGAGTCTTCGGCCAGTTTGGACGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCGAGTCTACCTTTACGACTT
GTCCACGGATTCTCACCGAGGGAACAAGTGGTAGACTGCAAACCTTAATATCCCAACGCTGAGCGAACCAGAGGTTTCGCGAC
GTACGCGAATCTTTGATTGCTTACTCGGGAATCTCTCGC

>gi|822660691|gb|CP008874.1|:929917-930122 *Halanaeroarchaeum sulfuri-reducens* strain HSR2, complete genome-taxID:1604004
GGTACTCCCGAGTCTCGCCACGGATGGCGGTGATAGGGATAACGGCTGGTTGGCACAGCCAGCGGTTACCTGTTTCGACCGTA
GTGAGGGTGTGTTTTGCCCCACGAGAACAGGTGAGAACCGCGAACCCTAATATCCCAACGCGGTTCGGGAATCCTCGCGCTTCA
GCGCGGGGAGGATGTCAACGTTCAAAGAACAGAAACGTTT

>gi|822660691|gb|CP008874.1|:404732-404933 *Halanaeroarchaeum sulfuri-reducens* strain HSR2, complete genome-taxID:1604004
AGTAACCTCCGAGTCTGCGTCCAGTTTGGACGAGGATAGGAGTACCGGCTGGTTGGCACAGCCTGGAGTCTACCTTCATGACCTC
TCCCGGGGATTCCAACCTCGGGCAGAGGTGGTAGATTGCAAACCGTAATATCCCAACCTCGGGATTCTCCGTCTTCAGGCGGA
GGAGGATGTCAAGGACGGGAACACATACCGGGAC

>gi|933677909|gb|CP011564.1|:499642-499847 *Halanaeroarchaeum sulfuri-reducens* strain M27-SA2, complete genome-taxID:1604004
AGTAACCTCCGAGTCTTCGGCCAGTTTGGACGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCGAGTCTACCTTTACGACTT
GTCCACGGATTCTCACCGAGGGAACAAGTGGTAGACTGCAAACCTTAATATCCCAACGCTGAGCGAACCAGAGGTTTCGCGAC
GTACGCGAATCTTTGATTGCTTACTCGGGAATCTCTCGC

>gi|933677909|gb|CP011564.1|:926613-926818 *Halanaeroarchaeum sulfuri-reducens* strain M27-SA2, complete genome-taxID:1604004
GGTACTCCCGAGTCTCGCCACGGATGGCGGTGATAGGGATAACGGCTGGTTGGCACAGCCAGCGGTTACCTGTTTCGACCGTA
GTGAGGGTGTGTTTTGCCCCACGAGAACAGGTGAGAACCGCGAACCCTAATATCCCAACGCGGTTCGGGAATCCTCGCGCTTCA
GCGCGGGGAGGATGTCAACGTTCAAAGAACAGAAACGTTT

>gi|933677909|gb|CP011564.1|:404525-404726 *Halanaeroarchaeum sulfuri-reducens* strain M27-SA2, complete genome-taxID:1604004
AGTAACCTCCGAGTCTGCGTCCAGTTTGGACGAGGATAGGAGTACCGGCTGGTTGGCACAGCCTGGAGTCTACCTTCATGACCTC
TCCCGGGGATTCCAACCTCGGGCAGAGGTGGTAGATTGCAAACCGTAATATCCCAACCTCGGGATTCTCCGTCTTCAGGCGGA
GGAGGATGTCAAGGACGGGAACACATACCGGGAC

>gi|76559182|emb|CR936258.1|:107695-107898 *Natronomonas pharaonis* DSM 2160 plasmid PL131 complete genome-taxID:348780

GCGAGCGACGGTGAAGTCCGACACAGGATAGGAGTAACGGCTGTTGGCTCAGCCTTCGGTACGCCTGTTGACAAAT
CCACGGGGCGAGTCGCCCCACAAGAACAGGTTGTCCCGTGAACCCCTAATATCCCAACGCGGGAATCCACGGCTTCAGCCGTG
TGGAAGAAGTCAAAATCAAGGAAATCAACGACAACCGC

>gi|312296031|gb|CP001693.1|c:113152-112953 Halogeometricum borinquense DSM 11551 plasmid pHBOR03, complete sequence-
taxID:469382

AGTATCTCCGAGTCCCGTTACGGATGGCGATGATAGGAGTAACAGCGGTTGGTATAGTCATCGACGCACCTGTTGATTTAGAGA
GTGATTGCTTCACACCACGAGAACAGGCTACGCTGTAAACCACAATATCCCCACATTGGGGAATTCCCGCCCTTCGGGGTGGGA
GGACGTAACTAAGCTGCATCAGTTATTTGT

>gi|76559182|emb|CR936258.1|c:127567-127367 Natronomonas pharaonis DSM 2160 plasmid PL131 complete genome-taxID:348780

GCGAGCGACGGTAACTCCGATCCGACACAGGATAGGAGTAACGGCTGTTTGGCACAGCCATCGGTCCGCCTGTTTCGAGAAAT
CAACGGGGCGAGTCGCCCCACAAGAACAGGTACGACCGTAAACCCCTAATATCCCAACGTGGGAATCCACGGCTTCAGCCGTG
TGGAGGATGTCAATGAAGAGGGCGGAGCAGTAGCC

>gi|822660691|gb|CP008874.1|c:320910-320705 Halanaeroarchaeum sulfureducens strain HSR2, complete genome-taxID:1604004

AGTAACTCCGAGTCTTCGGCCAGTTTGGACGGGGATAGGAGTACCGGCTGGTTGGCACAGCCTGGAGTCTACCTATACGACTTG
TCCAGCGGATTCCAACCGCAGGAACAAGTGGTGGACTGCAAACCCCTAATATCCCAACCTAAGCGGTGCGGTGCCGTGGGATTCC
CCGCGTGTTCACGCGCGGGGGGGATGTCAAGTAGACGTG

>gi|452081962|emb|HF582854.1|c:1879142-1879347 Natronomonas moolapensis 8.8.11 complete genome-taxID:268739

AGTAACTCCGAGTCTGCTCCCGATTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCAAAGTCTACCTTCATGGCCTC
TCCCAGGATTCCAACCTCGGGAAGAGGTGGTGGACTGCAAACCCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
GCCGCTTCAGGCGGAGGAGGATGTCAAGTACGGTCGAA

>gi|12057215|gb|AE004437.1|c:34775-34570 Halobacterium sp. NRC-1, complete genome-taxID:64091

ACGTATCTCCGAGTCCCGTCACGGATGGCGGTGATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTTGACAATG
AAACTGGTGGTTCGCACCTCAAGAACAGGCTACGTCGTAAACCATAATATCCCAACGCGGTGCGGAATCTCGCCCTTCAGGG
CGGGGAGGATGTCAATGCTGCACTCTCAGCTATCTGT

>gi|222454527|gb|CP001367.1|c:118168-117967 Halorubrum lacusprofundi ATCC 49239 plasmid pHLAC01, complete sequence-
taxID:416348

GATAACTCCGAGTCTTCACGGTGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGACATACTTGTTCGACCGCGAGAGCG
GGACGTTTCACACGAGAGAGAACAAGCCGTGTCGTAGACCAGCAAATATCCCACCTGCGGTACGGGAAGCCCCGGCGTTACGCC
GGGGAGGATGTCACTAGGAAATCTATCTTTTCACAA

>gi|222454527|gb|CP001367.1|c:117830-117627 Halorubrum lacusprofundi ATCC 49239 plasmid pHLAC01, complete sequence-
taxID:416348

GATAACTCCGAGTCTTCACGGCGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGGTACACTTGTTCGACCGCGAGAGCG
GGACGTTTCGCCCCGAGAGAACAGGTTGTATCGTAAACCAGTAAATATCCCACCTGCGGTACGGGAAGCCCCGGCGTTACGCC
GGGGAGGATGTCACTGCTGTAAAGTCTCTTATGCCGAA

>gi|433307828|gb|CP003373.1|c:260776-260569 Natrinema pellirubrum DSM 15624 plasmid pNATPE01, complete sequence-taxID:797303

CCAGAAGATAACTCTGAGTCCTCACGGTGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGGTACACCTGTTTCGACCGCG
AGAGCGGGACATTCCGTACGAGAGAACAGGTGCTGTCGTAGACCAGCAAATATCCCACCTGTGGTACGGGAAGCCTCGCCGTT
TACGGCGAGGAAGATGTACCCGATCTCCATGACGGTGAAG

>gi|433307828|gb|CP003373.1|c:280605-280403 Natrinema pellirubrum DSM 15624 plasmid pNATPE01, complete sequence-taxID:797303

CCGGCGAATAACTCAAATCTCCCAACCGGAGATAGTAAACGGCTGGTTGGCACAGCCCGGAGTCTACCTGTTTCGATACAAC
CGCAGGTGTGTTTACCCCCACGAACAGGTAGACTGCAAACCCCTAATATCCCACCGCTCGGGATTACCCCGTCTTCAGGCGG
GTGAGGATGTCAATCGAGGAGTCCGGTGAAAAAGA

>gi|452081962|emb|HF582854.1|c:97550-97345 Natronomonas moolapensis 8.8.11 complete genome-taxID:268739

AGTAACTCCGAGTCTGCTCCCGATTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCAAAGTCTACCTTCATGGCCTC
TCCCAGGATTCCAACCTCGGGAAGAGGTGGTGGACTGCAAACCCCTGATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
GCCGCTTCAGGCGGAGGAGGATGTCAATGCCCGTCGCG

>gi|452081962|emb|HF582854.1|c:2119423-2119218 Natronomonas moolapensis 8.8.11 complete genome-taxID:268739

AGTAACTCCGAGTCTGCTCCCGATTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCAAAGTCTACCTTCATGACCTC
TCCCGAGGATTCCAACCTCGGACAGAGGTGGTGGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
GCCGCCTTCAGGCGGAGGAGGATGTCAATTGAGTTCGTC

>gi|452081962|emb|HF582854.1|c10937-10732 *Natronomonas moolapensis* 8.8.11 complete genome-taxID:268739

AGTAACTCCGAGTCTGCTCCCGATTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCGAAGTCTACCTTCATGACCTC
TCCCGAGGATTCCAACCTCGGACAGAGGTGGTGGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
GCCGCCTTCAGGCGGAGGAGGATGTCAATGGTGCCCCGG

>gi|452081962|emb|HF582854.1|c1428173-1427968 *Natronomonas moolapensis* 8.8.11 complete genome-taxID:268739

AGTAACTCCGAGTCTGCTCCCGATTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCGAAGTCTACCTTCATGACCTC
TCCCGAGGATTCCAACCTCGGAAGAGGTGGTGGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
GCCGCCTTCAGGCGGAGGAGGATGTCAACAGTGAAGTCC

>gi|915987145|gb|KT322173.1|746-949 Uncultured haloarchaeon clone fosmid eHwalsbyiGI1-1 genomic sequence-taxID:160804

TAACTCCGAGTCTGCACCTGATTGGGTGGAGATAGGAGTACCGGCTGGTTGGCACAGCCAGCAGTCTATCTTCATGACTTGTCTC
CCACGGATTCTCACCGCGGGAACAAGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTTCGGTGCCGTGGGATTCTCTCC
GTCTTCAGGCGGATGGAGGATGTCAAATGAAGTGAA

>gi|55229005|gb|AY596294.1|114827-115030 *Haloarcula marismortui* ATCC 43049 plasmid pNG500, complete sequence-taxID:272569

TAACTCCGAGTCTACCGGCGAGTTGTCCGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTACCTGTACGACCTCTC
CTGCGGATTCCAACCGCAACGAGAGGTGGTAGACTTCAAACCCTAATATCCCAACCTCGGGATTCTCCCGGTTACGCGGAG
GAAGATGTCAAACCGCTGAACCTCGGAAGTACTGCTA

>gi|55229005|gb|AY596294.1|125457-125674 *Haloarcula marismortui* ATCC 43049 plasmid pNG500, complete sequence-taxID:272569

AAAACGCTACCCGTGACCCCTCCACTGGGGGGTAGCGGTAACGGCTGTTTGGCCAGCCTTGCGTCATTCAATTCAGCCGGACT
CGCGGATTCCAACCGCGAGCACCCGCGAGTGACGCGGAACCCTAATATCCCAACGCTTGCGGTGCGGTGCGGATTACCCCC
GACTTCAGGCGGGTGAGGAGGTCAAAATACGATCCGCTACTTTAATCTGGA

>gi|312291153|gb|CP001690.1|c574949-574744 *Halogeometricum borinquense* DSM 11551, complete genome-taxID:469382

AGTAACTCCGAGTCTCCGCCGGTTTGGACGGGAATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTACCTTCATGACTT
GTCCAGCGGATTCCAACCGCAGGAACAAGTGGTAGACTGCAAACCTTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
CCGCGTGTTCACGCGCGGGAGGATGTCAACCACGGTCATT

>gi|312291153|gb|CP001690.1|c26396-26196 *Halogeometricum borinquense* DSM 11551, complete genome-taxID:469382

ACGAAAGATAACTCAGAGTCCCCAACAGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGGAGTCTTCCTGTTGACCGCG
AGAGCGGGTCGTTCAACACGAGAGAACAGGGAGCCTGCAAACCGTAATATCCCAACGCTCGGGATTCTCCGGCTTCAGTCGG
AGGAGGATGTCAATGGTCGATGACTTCGGCGAACC

>gi|397680883|gb|CP003412.1|c1587117-1586916 *Natrinema* sp. J7-2, complete genome-taxID:406552

TAGGAAGATAACTCAGATCCTCTACAGAGGATATGAGTAACGGTCGTTTGGCACAGGCCAGTAGCCTACCTGTTCAACCAAAC
TTCGGGATCGTTCCACCCGAGGGAACAGATGGGTTGCGAACCTTAATATCCCAACGCTCGGGAATCCTCGCCCTTTAGATCGGG
GGAGGATGTCAAGTGTTGTTAACCGTACTAAT

>gi|312291153|gb|CP001690.1|1782688-1782888 *Halogeometricum borinquense* DSM 11551, complete genome-taxID:469382

ACGAAAGATAACTCAGAGTCCCCAACAGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGGAGTCTTCCTGTTGACCGCG
AGAGCGGGTCGTTCAACACGAGAGAACAGGGAGCCTGCAAACCGTAATATCCCAACGCTCGGGATTCTCCGGCTTCAGTCGG
AGGAGGATGTCAATTAAGCGATGCCCGGAATCCC

>gi|312291153|gb|CP001690.1|823278-823483 *Halogeometricum borinquense* DSM 11551, complete genome-taxID:469382

GGTACTCCCGAGTCTCGCCACGGATGGCGGTGATAGGGATAACGGCTGGTTGGCACAGCCAGCGGTTCGCCTGTTGACCGTA
GTGAGGGCCGTTTCGCCCCACGAGAACAGGTTGTGAACCACGAACCGTAATATCCCAACTCAGCGGTGCGGTGCCGTGGGATT
CCCGCGTTGACGAGACCAATGGTCTCGTCCGCCACCAGA

>gi|312291153|gb|CP001690.1|375074-375274 *Halogeometricum borinquense* DSM 11551, complete genome-taxID:469382

ACGAAAGATAACTCAGAGTCCCCGACAGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGGAGTCTTCCTGTTGACCGCG
AGAGCGGACGGTTACACCGAGAGAACAGGAGTCTGCAAACCGTAATATCCCAACGCTCGGGATTCTCCGGCTTCAGTCGG
AGGAGGATGTCAATTGCGGACGTTCTCGCCCCACC

>gi|312291153|gb|CP001690.1|:1802812-1803012 Halogeometricum borinquense DSM 11551, complete genome-taxID:469382
CAGAAGGATAACTCAGAGTCCCCGACAGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGAGTCTTCCTGTTGACCGCG
AGAGCGGACGGTTACACCGAGAGAACAGGAGTCTGCAAACCGTAATATCCCAACGCTCGGGATTCTCCGGCTGAAGCCG
GAGGAGGATGTCAACTGACGTGACTGGCGACTAACG

>gi|334353260|emb|FR746099.1|:1912618-1912833 Haloquadratum walsbyi C23 complete genome-taxID:768065
AGTATCTCTGAGTTCACCGAATCTGTCGGTGAATAGGAGTAACGGCTGGTTGGCACAGCCATCGACATTCTGTTGACAAGGA
GACTGGTGCCTTTGCACCTGAAGAACGGGTTAGGTCGTAAACCACAATATCCCAACTCAGCGGTGCGGTGCCGTGGGAATCCT
CGCGCTTCAGCGCGGGGAGGATGTCAAATATCTATCTGTAATAGAGGGG

>gi|334353260|emb|FR746099.1|:17304-17519 Haloquadratum walsbyi C23 complete genome-taxID:768065
AGTATCTCTGAGTTCAGTGAATCTGTCGGTGAATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTGACAAGGA
GACTGGTGCCTTTGCACCTGAAGAACGGGTTAGGTCGTAAACCACAATATCCCAACTCAGCGGTGCGGTGCCGTGGGAATCCT
CGCGCTTCAGCGCGGGGAGGATGTCAACCGAAGTGCTGTTCTCGATGTA

>gi|334353260|emb|FR746099.1|:1290206-1290411 Haloquadratum walsbyi C23 complete genome-taxID:768065
AGTAACCTCCGAGTCTTCGGCCAGTTTGGACGAGGATAGGAGTACCGGCTGGTTGGCACAGCCCGAGTCTACCTTCACGACTT
GTCCAGCGGATTCCAACCGCAGGGTCAAGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
TCCGTGTTTACGCGAAGGAGGATGTCAACTGAGCAGCGA

>gi|334353260|emb|FR746099.1|:2892290-2892494 Haloquadratum walsbyi C23 complete genome-taxID:768065
AGTAACCTCTGAGTCTGCACCCGATTTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCTGGAGTCTACCTTCATGACCTC
TCCCAGGATTCCAACCTCGGATAGAGGTGGTGGACTGCAAACCGTAATATCCCAACCCAAACGGTGCGGTGCCGTGGGATTCC
CTCCGCTGAACGCGGAGGAGGATGTCAATTGATGAGA

>gi|334353260|emb|FR746099.1|:243261-243464 Haloquadratum walsbyi C23 complete genome-taxID:768065
TAACCTCCGAGTCTGCACCTGATTTGGGTGGAGATAGGAGTACCGGCTGGTTGGCACAGCCAGCAGTCTATCTTCATGACTTGTC
CCACGGATTCTCACCGCGGGAACAAGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTTGCGGTGCCGTGGGATTCTCC
GTCTTCAGCGGATGGAGGATGTCAAATGAAGTGAA

>gi|334353260|emb|FR746099.1|:2196253-2196457 Haloquadratum walsbyi C23 complete genome-taxID:768065
AGTAACCTCCAGTTTTCGCTCCAACCTGGACGGGGATGTGAGTGTGACTGGTTGGCGCAGTCCAGGGTCACTCTCGTGGTTTA
TCAGATAGATTTTCATCGAGAAAATAAGTGGTAGAATGCAAACCCTACTATCACAACCTGCGGTGCGAGAATAATCGCTCTTCAACG
CCGGGAGGATATCAATCTGAGAGGTCAATATGTGTG

>gi|335339674|gb|CP002840.1|:c301285-301082 Halopiger xanaduensis SH-6 plasmid pHALXA01, complete genome-taxID:797210
GCGAGCAACGGTAACTCCGAGTCTGCGAAGGAGAGGAGTAACGGCTGGTTGGCGCAGCCAGCAGTCCGTCTGTTGATAAAT
ACACCGCGCTAGTAGCCGCGCGAGAACAGGTTGCTGACTGCAAACCATAATATCCCAACGCCCGTGAATCCTACGGCTTCAGCC
GTAGGAGGACGTCAAATTCGATCGTATCGTGATCTAA

>gi|335339674|gb|CP002840.1|:c313639-313439 Halopiger xanaduensis SH-6 plasmid pHALXA01, complete genome-taxID:797210
ACGAAAGATAACTCAGAGTCCCCGACCGGGGATATGAGTAACGGCTGGTTGGCCCAGCCCGGAGTCTTCCTGTTGACCGCG
AGAGCGGATGGTTACACCGAGAGAACAGGGAGACTGCAAACCGTAATATCCCAACGCTCGGGAATCCCACCCCTTCAGGGGT
GGGAGGAGGTCAAGCGGTTCCGATTTCCGGTGAA

>gi|573483295|gb|CP007057.1|:237537-237741 Halostagnicola larsenii XH-48 plasmid, complete sequence-taxID:797299
ACCCAGAACGCTTTGTGTTCTGGGGACGGCGATAGGGATAACAGCTGGTTGGAACAGCTTGCGATTACGCGGTTGACCGTAT
GAAGGTGTTTTGCTCCACGAGAACGGGTTGATAACCGCGAACCCTAATATCCCAACGCGGTGCGGATTCCCGCTCTTCAGGC
GCGGGAGGATGTCAATTAACGTTCTACCAAGTGCGA

>gi|573483295|gb|CP007057.1|:228044-228250 Halostagnicola larsenii XH-48 plasmid, complete sequence-taxID:797299
TCGACGAAAAGAACTCCGAGTTCGTCACCGGACAGGAGTAACGGCTGGTTGGCACAGCCAGCAGTTCACCTGTTCAATCGTA
GCGAGGGTCGTTTCTCCCGCGTGAACAGGTCATAGACTGCAAACCCTAATATCCCAACGGGGGTGCGGATTCCACGTCGGTTA
CGGCGTGAAGGATGTCAATCACCTCCTGTACTCTTCATT

>gi|148508018|gb|EF583985.1|:30394-30597 Uncultured haloarchaeon clone fosmid eHwalsbyi 023 genomic sequence-taxID:160804
AGGTACTCCAGAGTCTCGCCACGGGTGCGGTGAGAGGCATAACGGCTGGTTGGCACAGTTAGGGGTTACCCGTTGACCGGA
GTGAGGGCTGTTTCGCCCATGGGTTGCGAACCATCATATCTGACTCAGTGGTGCGGTGCTGTGAGATTCCCGCTCTTCAAG

CGTGGGAGGATGCCAATTCTCCAGAAAAGCTTTTGAT

>gi|344319636|gb|CP002988.1|:679829-680031 Halophilic archaeon DL31, complete genome-taxID:756883

CAGAAAGACAACGCAGAATCCTCCGACGGAGAATATGAATAACAGTTGTTTGGCACAGGCCAGCAGTTCGCTTGTTCATCAAA
GCTCGGGGCGTTTCGCACCGAGCGAACAGGCGAGTTGCGAACCTAATTCCCATCTCCGGTCGGGATTCCACCGCGTTCACGC
GGAGGAGGATGTCAACTGGTCCTCGGATACTGACTCG

>gi|76556520|emb|CR936257.1|:2256138-2256340 Natronomonas pharaonis DSM 2160 complete genome-taxID:348780

CGGCCTTCAGGCCGTGGATACGCGCCGTCACTTAATGAGTAACGGCTGGTTGGCACAGCCCGGAGTCTTCCTGTTGACCGCG
AGAGCGGACGGTTCACACCAAGAAAACAGGGAGTCTGCAAACCGTAATATCCCAACGCTGAGCAAATCGGAGATTGCGATGT
ACGCGAATCTTCGATTGCTTACTCGGGATTCTCC

>gi|344319636|gb|CP002988.1|:1955872-1955669 Halophilic archaeon DL31, complete genome-taxID:756883

AGTATCTCCGAGTCCCGTCACGGATGACGGTGATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTTGTGTTGACAAGGA
AACTGGTGCCTTCGCACCTCAAGAACAGGTCACGTCGTAAACCACAATATCCCAACCCAGCGGTGCGGTGCCGTGGGATTCCC
GCGTGTTCACGCGCGGGAGGATGTCAAGCTTCGTGGT

>gi|344319636|gb|CP002988.1|:1671543-1671340 Halophilic archaeon DL31, complete genome-taxID:756883

TAATCCGAGTCTGCGCCTGATTTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCATCAGTCTACCTTCATGACCTCTC
CCGAGGATTCCAACCTCGGGCAGAGGTGGTAGACTGCAAACCTAATATCCTAACTCAGCGGCGCGGTGCCGTGGGATTCCCG
CGTGTTCACGCGCGGGAGGATGTCAATTGGCGCCGAC

>gi|344319636|gb|CP002988.1|:278353-278154 Halophilic archaeon DL31, complete genome-taxID:756883

GATAACTCCGAGTCTCTACGGTGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGACACACCTGTTGACCGCGAGAGCG
GGACGTTTCGAACGAGAGAACAGGTCGTGTCTGACTGACCAACAAATATCCCAATCAACGGTGCGGTGCCGTGGGAAGCCCCG
TCATCTGTCGGGCTGCGCCCGACTGCCTGAGGGACG

>gi|344319636|gb|CP002988.1|:2297443-2297238 Halophilic archaeon DL31, complete genome-taxID:756883

AGTAACTCCGAGTCTCCGCCCATTAGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCATCAGTCTACCTTCATGACCTC
TCCCAGGATTCCAACCTCGGGCAGAGGTGGTAGACTGCAAACCTAATATCCCAATCAAGCGGTGCGGTGCCGTGGGATTCCC
GCGTGTTCACGCGCGGGAGGATGTCAATCTCGCGATTA

>gi|164521090|gb|EU080936.1|:136657-136453 Halobacterium sp. GN101 plasmid megaplasmid 2, complete sequence-taxID:88773

AAAGTTCTCCGAGTTCACCGAGCCTTTCCGGTGGATAGGAGTAACGGCTGGTTGGCACAGCCATCAACGTACTTGTTCGATTTCAG
AGAGCGGCTGCTTCGCACCGCGAGAACAAAGTCACGTCGTAAACCACAATATCCAACGGCAGTCGGGAATCCTCGCCCTTCAGC
GCGGGGAGGATGTCACTCCGGTATCGCTTGGCGATTCA

>gi|256553532|gb|GQ447672.1|:456-656 Uncultured prokaryote clone DeadSea1_IOBCFE001_09-C04-Rev genomic sequence-taxID:198431

AGTAACTCCGAGTCTGCGTCCAGTTTGGACGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTACCTGTATGACCTC
TCCTGCGGATTCCAACCGCAACGAGAGGTGGTAGACTGCAAACCTAATATCCCAACCTCGGGATCCTCCGCCTTCAGGCGGA
GGAGGATGTCAAGAACGCTCGTCATTGAGGCAG

>gi|167726115|emb|AM774415.1|:34778-34573 Halobacterium salinarum R1 complete genome-taxID:478009

ACGTATCTCCGAGTCCCGTCACGGATGGCGGTGATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTTGACAATG
AAACTGGTGCGTTTCGCACCTCAAGAACAGGCTACGTCGTAAACCATAATATCCCAACGCGGTGCGGAATCCTCGCCCTTCAGGG
CGGGGAGGATGTCAATGCTGCAGTCTCAGCTATCCTGT

>gi|433288832|gb|CP003050.1|:1069869-1069666 Halovivax ruber XH-70, complete genome-taxID:797302

ACCTCTGAGTCCGTCCGACGATACGTCCGCGGATAGGAGTAACGGCTGGTTGGCGCAGCCATCGACTCACCTGTTGACAAGT
CAACGGGTGGATTCCACCCACAAGAACAGGTTGTGTCGTAGACCTTAATATCCCAACTGCGGTGCGGAATCCTCGCCCTTCAGG
ACGGGAGGACGTCAACAGCCACGGTGTGGACACCCGT

>gi|433288832|gb|CP003050.1|:1084825-1084625 Halovivax ruber XH-70, complete genome-taxID:797302

CAGAAGGATAACTCAGAGTCCCCGACGGGAGATATGAGTAACGGCTGGTTGGCACAGCCCGGAGGCTTCCTGTTGACCGCG
AGAGCGGACGGTTCACACCGAGAGAATAGGGAGACTGTAAACCTAATATCCCAACGCTCGGGATTCTCCGGCTGAAGCCGG
AGGAGGATGTCAAACACGAGCGAAAAGTACTGCAT

>gi|572924095|gb|CP006885.2|:46365-46568 Haloarcula hispanica N601 chromosome 2, complete sequence-taxID:1417673

GCGAAAAACGGTAACTCCGAGTCTCGGAAGGATAGGAGTAACGGCTGCTTGGCTCAGCCATCGGTATACCTGTTGACAAAT
CCACGGGGCAAGTAGCCCCACAAGAACGGGTGCTTCCGTGAACCATAATATCCCAATCAAGCGGTGCGGTGCCGTGGGAATCC
CACGACTTCAGGCGTGTGGAGGAAGTCAAACCCGATCT

>gi|433304139|gb|CP003372.1|:c1777780-1777577 *Natrinema pellirubrum* DSM 15624, complete genome-taxID:797303

TAACTCCGAGTCTGCACCCAGTTTGGGTGGAGATAGGAGTACCGGCTGGTTGGCACAGCCCGAAGTCTACCTTCATAGCCTCTC
CTGCGGATTCCAACCGCAGACAGAGGTGGTGGACTGCAAACACTAATATCCCAACCGCGGTGCGGATTCTCCGCCTTCAGGC
GGAGGAGGATGTCAATACTATCTCGTTCGCCAACCGA

>gi|933677909|gb|CP011564.1|:c320703-320498 *Halanaeroarchaeum sulfireducens* strain M27-SA2, complete genome-taxID:1604004

AGTAACTCCGAGTCTTCGGCCAGTTTGGACGGGGATAGGAGTACCGGCTGGTTGGCACAGCCTGGAGTCTACCTATACGACTTG
TCCAGCGGATTCCAACCGCAGGAACAAGTGGTGGACTGCAAACCCTAATATCCCAACCTAAGCGGTGCGGTGCCGTGGGATT
CCGCGTGTTCACGCGCGGGGGGATGTCAAGTAGACGTG

>gi|528527607|emb|HF571521.1|:c25656-25457 *Halorhabdus tiamatea* SARL4B plasmid pHTIA complete sequence-taxID:1033806

AGTATCTCCGAGTTCACCGAATCTCTCGGTGAATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTTGACAAGGA
AACTGGCGCATTTCGCGCCTCAAGAACAGGTCACGTCGTAAACCACAATATCCCAACCGCGGAATCCTCGCCCTTCAGCGCGGG
GAGGATGTCAAGGAACGATGTTGCACACCAGAC

>gi|573483295|gb|CP007057.1|:c231197-230998 *Halostagnicola larsenii* XH-48 plasmid, complete sequence-taxID:797299

AGTATCTCCGAGTCTCTCCTGTGACGGGGGAGATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTTGTGTTGACACGGA
AACTGGTGCCTTCGCACCTCAAGAACACATAACGCCGTAAACCACAATATCCCACCGCGGAATCCTCGCCCTTCAGGGCGGG
GAGGATGTCAACTTGTCAACAATATCTACACCC

>gi|284016690|gb|CP001861.1|:593063-593264 *Haloterrigena turkmenica* DSM 5511 plasmid pHTUR01, complete sequence-taxID:543526

GATAACTCCGAGTCTCACGGTGAGGATAGGAGTAACGGCTGTGTGGCACAGCCATCGACATACCTGTTGACCGCGAGAGCG
GGACGTTTCGCACGAGAGAACAGGTCGTGTCGTAAACCGACAAATATCCCACCTGCGGTACGGGAAGCCCCGGCGTTTACGCC
GGGGAGGATGTACCAGCAGCTCATCCGACGGGCTC

>gi|312294779|gb|CP001692.1|:209121-209321 *Halogeometricum borinquense* DSM 11551 plasmid pHBOR02, complete sequence-taxID:469382

ACGAAAGATAACTCAGAGTCCCCCAACAGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGGAGTCTTCCTGTTGACCGCG
AGAGCGGGTCGTTCAACACGAGAGAACAGGGAGCCTGCAAACCGTAATATCCCAACGCTCGGGATTCTCCGGCTTCAGTCGG
AGGAGGATGTCAATACTGATGACTACTCGTCCGCC

>gi|343784702|gb|CP002922.1|:46311-46514 *Haloarcula hispanica* ATCC 33960 chromosome II, complete sequence-taxID:634497

GCGAAAAACGGTAACTCCGAGTCTCGGAAGGATAGGAGTAACGGCTGCTTGGCTCAGCCATCGGTATACCTGTTGACAAAT
CCACGGGGCAAGTAGCCCCACAAGAACGGGTGCTTCCGTGAACCATAATATCCCAATCAAGCGGTGCGGTGCCGTGGGAATCC
CACGACTTCAGGCGTGTGGAGGAAGTCAAACCCGATCT

>gi|433307828|gb|CP003373.1|:241623-241823 *Natrinema pellirubrum* DSM 15624 plasmid pNATPE01, complete sequence-taxID:797303

CCGAAGGATAACTCAGAGTCTCCGACAGGGGATATGAGTAACGGCTGGTTGGCACAGCCCGGAGTCTTCCTGTTGACCGCG
AGAGCGGGTCGTTCAACACGAGAGAACAGGGAGACTGTAAACCCTAATATCCCAACGCTCGGGATTCTCCGGCTTTAGTGATT
AGCAAATCTTATTGAGAAATCCACTGAGGTGATT

>gi|433307828|gb|CP003373.1|:72074-72279 *Natrinema pellirubrum* DSM 15624 plasmid pNATPE01, complete sequence-taxID:797303

AGTAACTCCGAGTCTTCGGCCAGTTTGGACGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTACCTGTATGACTTG
TCCAGCGGATTCCAACCGCAGGATCAAGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGCGCGGTGTCGTGGGATTCCC
GCGTCTTCAGGCGCGGGAGGATGTCAAGTATGACCGAG

>gi|433307828|gb|CP003373.1|:270967-271167 *Natrinema pellirubrum* DSM 15624 plasmid pNATPE01, complete sequence-taxID:797303

GGTACTCCCGAATCTCGCCTGTGATGGGGGAGATAGGGATAACGGCTGGATGGCACAGCCAGCGGTTTCGCCTGTTGACAAATC
CACGGGCAGAGTACGCCCACAAGAGCAGGTGGCCTGCGAACCATAATCTCCTAACGCTCAGGAAACCGCGCCGTTACAGGGC
CGGAGGATGTATCCACTGTGACGTTTCACACGT

>gi|289529249|gb|CP001932.1|:c2280572-2280373 *Natrialba magadii* ATCC 43099, complete genome-taxID:547559

AGTATCTCCGAGTCTCGCCTGTGACGGAGGAGATAGGAGTAACGGCTGGTTGGTACAGCCATCGACGTTCTTGTGTTGACAAGA
AAACTGGTGTGTTGCACCTCAAGAGCAGGTCACGTCGTAAACCACAATACCCCAACGTGGGAATCCTCGCGCTTCAGCGCGG

GGAGGATGTCAAGCCGTCGCTTCGGTCTCCTCGA

>gi|289529249|gb|CP001932.1|:c312695-312492 *Natrialba magadii* ATCC 43099, complete genome-taxID:547559

TAACTCCGAGTCTGCACCCAGTTTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCAGGAGTCTACCTTCATGACCTCTC
CCGAGGATTCCAACCTCGGACAGAGGTGGTAGACTCTAAACCATAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCTCTCC
GCGTTCACGCGGAGGAGGATGTCAATAATACGATCG

>gi|289529249|gb|CP001932.1|:c1353617-1353414 *Natrialba magadii* ATCC 43099, complete genome-taxID:547559

TAACTCCGAGTCTTCGGCCAGTTTGGGCGAGGATAGGAGTACCGGCTGGTTGGCACAGCCTGGAGTCTACCTTCATGACTTGTG
TCACGGATTCTCACCGAGAGACAAGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCTCTCC
GTGTTACGCGGAGGAGGATGTCAATCTGGGAGTAT

>gi|148508225|gb|EF583996.1|:1388-1593 Uncultured haloarchaeon clone fosmid eHwalsbyi 464 genomic sequence-taxID:160804

AGTAACTCCGAGTCTTCGGCCAGTTTGGACGAGGATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTACCTTCACGACTT
GTCCAGCGGATTCCAACCGCAGGGTCAAGTGGTAGACTGCAAACCCTAATATCCCAATCCAGCGGTGCGGTGCCGTGGGATTCC
TCCGTGTTTACGCGGAGGAGGATGTCAATCTGACGAGTG

>gi|429134263|gb|CP003377.1|:c55398-55190 *Natronobacterium gregoryi* SP2, complete genome-taxID:797304

TCACTCATTACGGTCTCTCGTCATCACGAAAGGCGCGAGCGCCTTTCGAACGACACAGCCATCGGTACACCTGTTGACCGCGA
AAGCGGGACGTTCAACACGAGAAGACAGGTCGTATCGTAGACCAGCAAATATCCCACCCTGCGGTACGGGAAGCCGCGTCCTT
CAGGGCGCGGAAGATGTCACCACCCGTAGATATTTACTGCGC

>gi|289533150|gb|CP001934.1|:14953-15156 *Natrialba magadii* ATCC 43099 plasmid pNMAG02, complete sequence-taxID:547559

AGTATCTCCGAGTCCCGCCACGGATGGGGCGATAGGAGTAACGGCTGGTTGGCACAGCCATCGACGTTCTGTTGACAAGG
AAACTGGCGTGTTGCGCCTCAAGAACAGGTCACGTCGTAAACCACAATATCCCAACGCGGTGCGGAATCTTCGCCCTTCAGG
GCGGAGAGGATGTCAACGGGTGACCCGCGGACCCGAAT

>gi|749699433|gb|CP010529.1|:c339323-339120 *Haloarcula* sp. CBA1115, complete genome-taxID:1592728

GCGAAAAACGGTAACTCCGAGTCTGCGAAGGATAGGAGTAACGGCTGCTTGGCTCAGCCATCGGTATACCTGTTGACAAAT
CCACGGGGCAAGTAGCCCCACAAGAACGGGTGCGTTCCGTGAACCATAATATCCCAATCAAGCGGTGCGGTGCCGTGGGAATCC
CACGACTTCAGTCGTGTGGAGGAAGTCAAACCCGATCT

>gi|334353260|emb|FR746099.1|:c1609515-1609312 *Haloquadratum walsbyi* C23 complete genome-taxID:768065

TAACTCTGAGTCTGCACCCGATTTGGGTGGTGATAGCAGTACCGGCTGGTTGGCACAGCCTGGAGTCTATCTTCATGACCTTTCC
CGAGGATTCCAACCTCGGACAGAGGTGGGAGACTACAAACCCTAATATCCCAACCCAGCGGTGCGGTGCCGTGGATTCTCTCC
AAGTTCACGCGGAGGAGGATGTCAATCTGGATAACG

>gi|334353260|emb|FR746099.1|:c2111580-2111376 *Haloquadratum walsbyi* C23 complete genome-taxID:768065

ACTCTGAGTCTTCACCTGATTTGGGTGGGCATAGGAGTGCCGGTTGGTTGGCACAGCACGGAGTCTCCCTTCAGGTTGTCCAGC
GGATTCCAACCGCAGGAACAAGTGGTAGACTGCAAACCCTACCCTAATATCCCAACTGCGGTGCGGAATCCTCGCCGTTTCAGG
GCGGAGAGGATGTCAACTAGGTCAGCTGAGTGATCAAA

>gi|334353260|emb|FR746099.1|:c1875438-1875235 *Haloquadratum walsbyi* C23 complete genome-taxID:768065

AGTATCCCCGAGTCTCGCCATGGATGGCGATGACAGGAGTAACGGCTAGTCGGCACAGCTATCGACGTACCTGTTTGACAAGGA
AACTGACTTGTTCAAGTTTCAACAATAGGTAATGTCGTGAGATTCCGGCGTCTTCAGGCGCGAGAGGATGTCAAATATTTACTAC
TCATGACTTATCTTCTAAGTGAAGCGTGATTATT

>gi|334353260|emb|FR746099.1|:c2209485-2209282 *Haloquadratum walsbyi* C23 complete genome-taxID:768065

TAACTCTGAGTCTGCACCCGATTTGGGTGGGGATAGGAGTACCGGCTGGTTGGCACAGCCCGGAGTCTACCTTTATGACCTTTC
CCGAGGATTCCAACCTCGGGCAGAGGTGGGAGACTGCAAACCCTAATATCCCAACCCAAGCGGTGTGGTGCCGTGGGATTCTCT
CCGCGTTACGCGGAGGAGGATGTCAATTTGATCGAG

>gi|284016690|gb|CP001861.1|:c166401-166201 *Haloterrigena turkmenica* DSM 5511 plasmid pHTUR01, complete sequence-taxID:543526

CAGAAAGATAACTCAGATCCTCCAACAGGGGATAGTAAACGGTCGTTTGGCACGGCCAGTAGCCTACCTGTTCAACCAAA
CCTCGGGGCATTTCTGATCCGAGCGAACAGGTGGATTGCAAACCGTAATATCCCAACGCTCGGGAATCCTCGCTCTTTAGGGCGG
GGAGGATGTCAACGATCGATCGCGGACCCTCCGT

>gi|749699433|gb|CP010529.1|:1588262-1588457 *Haloarcula* sp. CBA1115, complete genome-taxID:1592728

GGTAACTCCGAGTCCGTCACAGGATAGGAGTAACGGCTGTTTGGCACAGCCATCGGTGCGCCTGTTGATAAATCAACGGGCA

GATTACGCCACAAGAACAGGTTGTCCCGTAGACCACAATATCCCAACGGCAGTCGGGAATCTTCGCCCTTCAGGGCGGAGAG
GATGTCAAGCGGGTTCGTCCTGGTCCTCAA

ANEXO B- Lista completa de genes diferencialmente expressos identificados por RNA-seq da linhagem VNG_sot0042

New Annotation	Old Annotation	log2FoldChange	pvalue	Function / Domains
VNG_RS12025	VNG6442H	0,62	1,10E-02	ORF H1655
VNG_RS07325	VNG1889H	0,55	1,58E-02	
VNG_RS12000	new gene	1,40	3,04E-06	hypothetical protein
VNG_RS02735	new gene	0,89	3,70E-03	hypothetical protein
VNG_RS05075	new gene	0,57	4,51E-02	
VNG_RS00440	new protein	0,71	1,40E-03	hypot
VNG_RS00425	VNG0108G	0,69	9,17E-04	DEAD/DEAH box helicase
VNG_RS00430	VNG0110C	1,13	1,94E-05	hypothetical protein
VNG_RS00730	VNG0177G	0,80	1,78E-07	50S ribosomal protein L15e
VNG_RS00995	VNG0244H	0,54	2,21E-02	
VNG_RS01325	VNG0335a	0,56	1,52E-02	gptA1
VNG_RS01470	VNG0375G	0,59	2,76E-03	preprotein translocase subunit SecE
VNG_RS01615	VNG0415G	0,78	7,69E-04	adenylosuccinate lyase
VNG_RS01755	VNG0451G	0,67	1,15E-02	phosphate transport system regulatory protein PhoU
VNG_RS01775	VNG0457G	0,59	3,42E-03	phosphate ABC transporter substrate-binding protein
VNG_RS01790	VNG0461G	0,52	2,18E-02	aspartate--tRNA(Asp/Asn) ligase
VNG_RS02045	VNG0524G	0,53	1,36E-02	ABC transporter ATP-binding protein
VNG_RS02055	VNG0527C	0,52	2,48E-02	Fe-S cluster assembly protein SufD
VNG_RS02200	VNG0560C	0,58	2,00E-02	Na(+)/H(+) antiporter subunit D
VNG_RS02290	VNG0582C	0,89	9,22E-07	cytochrome b
VNG_RS02505	VNG0632G	0,67	3,11E-03	5-(carboxyamino)imidazole ribonucleotide synthase
VNG_RS02510	VNG0633G	0,53	2,59E-02	5-(carboxyamino)imidazole ribonucleotide mutase
VNG_RS02530	VNG0639G	0,63	1,17E-02	NADH dehydrogenase
VNG_RS02560	VNG0647G	0,72	5,30E-03	oxidoreductase
VNG_RS02565	VNG0648G	0,81	1,08E-03	oxidoreductase
VNG_RS02715	VNG0686C	0,68	1,24E-03	DUF159 family protein
VNG_RS03375	VNG0864G	0,72	2,53E-04	phosphoribosylformylglycinamide synthase
VNG_RS03770	VNG0967Gm	0,62	4,03E-02	chemotaxis protein CheD
VNG_RS03855	VNG0991H	0,77	5,03E-04	hypot
VNG_RS03860	VNG0992H	1,05	3,41E-06	hypothetical protein
VNG_RS04120	VNG1058H	0,61	1,15E-04	hypot
VNG_RS04130	VNG1060H	1,04	4,78E-04	hypothetical protein
VNG_RS04170	VNG1069C	0,70	4,68E-03	DUF368 domain-containing protein
VNG_RS04225	VNG1085H	0,60	3,91E-03	hypot
VNG_RS04465	VNG1143G	0,59	1,60E-03	30S ribosomal protein S2
VNG_RS04670	VNG1194H	0,61	2,03E-02	
VNG_RS04700	VNG1202C	0,67	1,10E-02	hydrolase
VNG_RS04790	VNG1224G	0,78	1,37E-04	cell division control protein Cdc6
VNG_RS04955	VNG1263C	0,64	1,27E-04	hypot

VNG_RS05270	VNG1351G	0,60	2,55E-02	transcriptional regulator
VNG_RS05360	VNG1375C	0,57	2,54E-02	histidine kinase
VNG_RS05585	VNG1433G	0,71	3,52E-04	30S ribosomal protein S17e
VNG_RS05590	VNG1435G	0,51	1,09E-02	aspartate-semialdehyde dehydrogenase
VNG_RS05645	VNG1452G	0,62	3,38E-02	translation initiation factor eIF-2B subunit delta
VNG_RS06060	VNG1557G	0,56	3,84E-02	precorrin-3B C(17)-methyltransferase
VNG_RS06210	VNG1601G	0,60	2,01E-02	glycine dehydrogenase subunit 2
VNG_RS06325	VNG1631G	1,08	1,16E-04	ABC transporter ATP-binding protein
VNG_RS06335	VNG1634G	0,94	1,68E-03	cobalt transporter CblN
VNG_RS06355	VNG1638H	0,69	3,58E-04	hypot
VNG_RS06360	VNG1640H	0,72	1,27E-02	
VNG_RS06420	VNG1657H	0,54	1,87E-02	
VNG_RS06670	VNG1719G	0,58	1,20E-02	preprotein translocase subunit SecY
VNG_RS06765	VNG1744H	2,00	2,27E-12	ZIP family metal transporter
VNG_RS06845	VNG1767G	0,71	1,56E-04	agmatinase
VNG_RS06875	VNG1773G	0,52	2,25E-02	4a-hydroxytetrahydrobiopterin dehydratase
VNG_RS07010	VNG1806H	0,63	5,63E-03	hypot
VNG_RS07015	VNG1807H	0,59	4,19E-03	hypot
VNG_RS07045	VNG1818G	0,76	9,29E-04	isopentenyl-diphosphate Delta-isomerase
VNG_RS07185	VNG1851G	0,54	4,18E-02	sugar kinase
VNG_RS07200	VNG1856G	0,57	2,24E-02	type I deoxyribonuclease HsdR
VNG_RS07290	VNG1880C	0,62	4,81E-03	amidohydrolase
VNG_RS07480	VNG1925H	0,63	1,80E-03	transcription regulator-like protein
VNG_RS07530	VNG1939G	0,56	1,38E-02	phosphoribosylaminoimidazolesuccinocarboxamide synthase
VNG_RS07550	VNG1945G	0,80	9,39E-04	phosphoribosylformylglycinamide synthase
VNG_RS07700	VNG1987G	0,69	6,15E-04	preprotein translocase subunit SecD
VNG_RS07815	VNG2017G	0,53	2,76E-02	lysine--tRNA ligase
VNG_RS08000	VNG2071G	0,59	1,96E-02	long-chain fatty acid--CoA ligase
VNG_RS08100	VNG2093G	0,51	5,39E-03	glutamine synthetase
VNG_RS08195	VNG2116C	0,63	4,52E-04	uracil permease
VNG_RS08290	VNG2138G	0,59	7,20E-04	V-type ATP synthase subunit B
VNG_RS08295	VNG2139G	0,73	2,96E-04	ATP synthase subunit A
VNG_RS08655	VNG2220G	0,65	8,54E-04	dihydrolipoamide dehydrogenase
VNG_RS09025	VNG2308G	0,58	4,40E-03	DNA-binding protein
VNG_RS09060	VNG2315H	0,65	2,98E-03	hypot
VNG_RS09160	VNG2338G	0,58	1,45E-03	DNA polymerase II large subunit, intein-containing
VNG_RS09180	VNG2342H	0,69	5,79E-03	hypot
VNG_RS09190	VNG2344G	0,71	4,30E-04	ABC transporter ATP-binding protein
VNG_RS09335	VNG2383G	0,64	1,03E-02	ribonucleotide-diphosphate reductase subunit beta
VNG_RS09385	VNG2397G	1,00	1,83E-05	ABC transporter
VNG_RS09870	VNG2521H	0,53	1,25E-02	universal stress protein UspA
VNG_RS09905	VNG2529G	0,75	4,87E-04	peptide ABC transporter permease
VNG_RS09910	VNG2531G	0,94	1,12E-04	peptide ABC transporter permease
VNG_RS09915	VNG2532H	0,65	1,49E-02	
VNG_RS10005	VNG2556H	0,68	1,64E-03	hypot
VNG_RS10060	VNG2573G	0,72	3,00E-03	halolysin
VNG_RS10215	VNG2611G	0,54	4,53E-02	GTPase HflX
VNG_RS11100	VNG6155H	0,52	4,33E-02	
VNG_RS11390	VNG6240G	0,79	7,27E-03	protein gvpD
VNG_RS11470	VNG6265G	1,15	2,11E-05	adhesin

VNG_RS05105	VNG1308G	-0,86	2,86E-06	succinate dehydrogenase
VNG_RS05145	VNG1318H	-0,83	4,54E-03	hypothetical protein
VNG_RS05215	VNG1337C	-0,53	3,45E-02	
VNG_RS05300	VNG1359G	-0,75	1,14E-03	endonuclease
VNG_RS05380	VNG1380H	-1,66	3,23E-12	hypothetical protein
VNG_RS05740	VNG1473H	-0,57	1,19E-02	
VNG_RS05765	VNG1479H	-0,51	1,18E-02	
VNG_RS05845	VNG1500H	-0,95	4,74E-04	hypothetical protein
VNG_RS05920	VNG1523G	-0,54	3,92E-03	methyl-accepting chemotaxis protein
VNG_RS05980	VNG1538H	-0,75	2,13E-05	
VNG_RS06480	VNG1672H	-0,63	3,99E-02	
VNG_RS06620	VNG1705G	-0,65	4,66E-03	50S ribosomal protein L5
VNG_RS06775	VNG1748C	-0,62	3,44E-02	short-chain dehydrogenase
VNG_RS06800	VNG1755G	-0,57	3,36E-02	phytoene dehydrogenase
VNG_RS07145	VNG1843C	-1,21	1,72E-13	MarR family transcriptional regulator
VNG_RS07360	VNG1898C	-0,82	1,37E-03	universal stress protein UspA
VNG_RS07425	VNG1912G	-1,04	7,33E-07	anthranilate phosphoribosyltransferase
VNG_RS07430	VNG1914G	-0,53	9,56E-03	peptidyl-prolyl cis-trans isomerase
VNG_RS07895	VNG2037C	-0,85	1,41E-05	histidine kinase
VNG_RS08090	VNG2089H	-0,56	1,45E-03	
VNG_RS08170	VNG2109H	-0,57	3,23E-03	
VNG_RS08430	VNG2166Cm	-0,60	4,71E-02	saccharopine dehydrogenase
VNG_RS08485	VNG2179H	-0,79	1,01E-02	
VNG_RS08560	VNG2196G	-0,68	3,64E-03	halocyanin-like protein
VNG_RS08565	VNG2197H	-0,67	1,26E-02	
VNG_RS08570	VNG2199H	-0,81	4,11E-03	hypothetical protein
VNG_RS08675	VNG2226G	-0,94	1,70E-06	thermosome subunit
VNG_RS08820	VNG2262H	-0,79	1,02E-02	
VNG_RS08925	VNG2286G	-0,59	4,28E-02	methylaspartate mutase
VNG_RS08960	VNG2294G	-0,83	3,70E-03	1-(5-phosphoribosyl)-5-((5-phosphoribosylamino)methylideneamino)imidazole-4-carboxamide isomerase
VNG_RS09145	VNG2334C	-0,67	7,66E-04	
VNG_RS09270	VNG2368G	-1,02	4,51E-09	DEAD/DEAH box helicase
VNG_RS09330 uvrC	VNG2381G	-0,64	1,96E-02	excinuclease ABC subunit C
VNG_RS09435	VNG2412H	-1,54	1,10E-10	hypothetical protein
VNG_RS09440	VNG2413H	-0,93	3,65E-05	
VNG_RS09555	VNG2443G	-0,52	6,22E-03	DNA starvation/stationary phase protection protein
VNG_RS09645	VNG2468C	-0,66	6,81E-03	
VNG_RS09760	VNG2501C	-0,66	7,90E-03	
VNG_RS09865	VNG2520C	-0,55	4,14E-02	universal stress protein UspA
VNG_RS10190	VNG2606G	-0,67	5,39E-04	bifunctional hydroxymethylpyrimidine kinase/phosphomethylpyrimidine kinase
VNG_RS11135	VNG6165H	-0,57	1,53E-02	
VNG_RS11335	VNG6225C	-0,62	1,73E-02	ATPase
VNG_RS11425	VNG6251H	-0,71	8,48E-05	
VNG_RS11435	VNG6255C	-1,18	1,89E-07	hypothetical protein
VNG_RS11440	VNG6256G	-0,57	1,73E-02	lipoate--protein ligase
VNG_RS11640	VNG6318G	-0,61	8,34E-03	transcriptional regulator
VNG_RS11850	VNG6372H	-0,60	8,82E-04	
VNG_RS12740	VNG7101	-0,97	1,69E-06	ORF H1224
VNG_RS12905	VNG7132	-0,94	9,90E-05	ORF H1522
VNG_RS00490	VNGt02	-0,67	9,49E-04	tRNA
VNG_RS00565	VNGt04	-0,59	3,62E-04	tRNA

VNG_RS11730	VNG6341H	0,59	2,35E-04	transposase
VNG_RS11755	VNG6348H	0,51	3,16E-02	
VNG_RS09670	VNG2474C	-0,57	4,99E-02	
VNG_RS09730	-	-0,98	1,22E-03	hypothetical protein
VNG_RS00300	new gene	-0,65	1,45E-04	
VNG_RS06980	new gene	-0,65	5,87E-03	
VNG_RS03895	new gene	-0,65	3,18E-02	
VNG_RS00035	VNG0008G	-0,63	6,93E-04	glucose-1-phosphate thymidyltransferase
VNG_RS00235	VNG0057H	-0,87	3,93E-06	hypothetical protein
VNG_RS00355	VNG0089G	-0,60	3,38E-02	protein-L-isoaspartate O-methyltransferase
VNG_RS00525	VNG0129G	-0,61	4,78E-03	protease
VNG_RS00610	VNG0149G	-0,83	6,16E-03	cadmium-transporting ATPase
VNG_RS00675	VNG0163G	-0,51	2,90E-02	DNA mismatch repair protein MutS
VNG_RS01030	VNG0252C	-0,70	5,70E-05	geranylgeranylgeranylheptaprenylgeranyl phosphate synthase
VNG_RS01055	VNG0259G	-0,58	7,39E-04	inorganic pyrophosphatase
VNG_RS02310	VNG0586C	-0,53	2,61E-03	halocyanin
VNG_RS02350	VNG0595H	-0,63	3,01E-02	
VNG_RS02455	VNG0620G	-0,61	7,79E-03	peptidase
VNG_RS02585	VNG0654C	-0,54	3,66E-02	N-acetyltransferase
VNG_RS02675	VNG0678G	-0,63	2,34E-02	3-ketoacyl-CoA thiolase
VNG_RS02720	VNG0688H	-1,02	6,32E-04	hypothetical protein
VNG_RS02800	VNG0715G	-0,73	2,51E-04	phosphomethylpyrimidine synthase
VNG_RS02810	VNG0717H	-0,52	1,19E-02	
VNG_RS02840	VNG0725H	-0,57	7,56E-03	transcriptional regulator
VNG_RS02880	VNG0734G	-0,57	6,09E-03	transcription initiation factor IIB 2
VNG_RS02960	VNG0753G	-0,70	1,45E-03	anion permease
VNG_RS02965	VNG0754C	-0,58	7,14E-03	
VNG_RS03115	VNG0795G	-0,70	1,98E-04	halocyanin
VNG_RS03230	VNG0824G	-0,60	4,95E-02	molybdenum cofactor guanylyltransferase
VNG_RS03260	VNG0831G	-0,58	4,31E-02	molybdopterin oxidoreductase
VNG_RS03385	VNG0867G	-0,69	1,61E-02	asparagine synthase
VNG_RS03550	VNG0913H	-0,73	2,53E-03	
VNG_RS03605	VNG0925C	-0,98	1,93E-10	hypothetical protein
VNG_RS03640	VNG0935G	-1,23	5,89E-05	NADH oxidase
VNG_RS03650	VNG0937G	-1,05	5,94E-05	glyceraldehyde-3-phosphate dehydrogenase
VNG_RS03730	VNG0959H	-0,69	1,03E-03	
VNG_RS03880	VNG0997G	-0,65	1,82E-03	acetyl-coenzyme A synthetase
VNG_RS04040	VNG1035C	-1,12	2,20E-06	GMC family oxidoreductase
VNG_RS04045	VNG1036H	-0,64	3,39E-02	
VNG_RS04370	VNG1120H	-0,62	1,04E-02	cyclase
VNG_RS04375	VNG1121G	-0,82	1,50E-03	aspartate aminotransferase
VNG_RS04400	VNG1131G	-0,52	3,42E-02	ATP-binding protein
VNG_RS04435	VNG1137G	-0,54	7,28E-04	50S ribosomal protein L18e
VNG_RS04615	VNG1181G	-0,68	9,60E-03	flagellin
VNG_RS04630	VNG1183H	-0,75	4,89E-03	
VNG_RS04625	VNG1184G	-0,99	5,47E-05	radical SAM/SPASM domain-containing protein
VNG_RS04640	VNG1187G	-1,18	1,11E-07	membrane protein
VNG_RS04645	VNG1188G	-1,05	4,21E-05	halocyanin
VNG_RS04655	VNG1190G	-0,64	9,82E-04	superoxide dismutase
VNG_RS04875	VNG1245C	-0,50	4,58E-03	shikimate kinase
VNG_RS04995	VNG1276C	-0,59	3,05E-02	rRNA metabolism protein

VNG_RS00805	VNGt06	-0,76	6,02E-06	tRNA
VNG_RS02070	VNGt08	-0,87	2,49E-04	tRNA
VNG_RS02185	VNGt09	-0,66	3,00E-04	tRNA
VNG_RS03100	VNGt11	-0,78	2,42E-07	tRNA
VNG_RS03210	VNGt13	-0,82	8,43E-07	tRNA
VNG_RS04755	VNGt19	-0,77	1,15E-02	tRNA
VNG_RS04785	VNGt20	-0,56	5,16E-03	tRNA
VNG_RS06400	VNGt28	-0,57	1,52E-03	tRNA
VNG_RS06545	VNGt29	-0,60	2,78E-02	tRNA
VNG_RS06715	VNGt30	-0,90	4,87E-06	tRNA
VNG_RS07440	VNGt32	-0,85	1,76E-04	tRNA
VNG_RS07665	VNGt33	-0,70	2,04E-05	tRNA
VNG_RS07860	VNGt34	-0,60	3,44E-02	tRNA
VNG_RS08420	VNGt37	-0,78	1,35E-06	tRNA
VNG_RS09280	VNGt39	-0,98	1,51E-07	tRNA
VNG_RS09795	VNGt41	-0,68	3,02E-03	tRNA

ANEXO C – Proteínas enriquecidas nas análises de SILAC

ANEXO D - GOMES-FILHO, J. V. et al. Sense overlapping transcripts in IS 1341 -type transposase genes are functional non-coding RNAs in archaea. *RNA Biology*, v. 12, n. 5, p. 490–500, 4 maio 2015.

Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea

José Vicente Gomes-Filho, Livia Soares Zaramela, Valéria Cristina da Silva Italiani, Nitin S Baliga, Ricardo Z N Vêncio & Tie Koide

To cite this article: José Vicente Gomes-Filho, Livia Soares Zaramela, Valéria Cristina da Silva Italiani, Nitin S Baliga, Ricardo Z N Vêncio & Tie Koide (2015) Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea, RNA Biology, 12:5, 490-500, DOI: [10.1080/15476286.2015.1019998](https://doi.org/10.1080/15476286.2015.1019998)

To link to this article: <http://dx.doi.org/10.1080/15476286.2015.1019998>

© 2015 The Author(s). Published with license by Taylor & Francis Group, LLC© José Vicente Gomes-Filho, Livia Soares Zaramela, Valéria Cristina da Silva Italiani, Nitin S Baliga, Ricardo Z N Vêncio, and Tie Koide. Accepted author version posted online: 25 Mar 2015.

View supplementary material [↗](#)

Submit your article to this journal [↗](#)

Article views: 369

View related articles [↗](#)

View Crossmark data [↗](#)

Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea

José Vicente Gomes-Filho¹, Livia Soares Zaramela¹, Valéria Cristina da Silva Italiani^{1,#}, Nitin S Baliga², Ricardo Z N Vêncio³, and Tie Koide^{1,*}

¹Department of Biochemistry and Immunology; Ribeirão Preto Medical School; University of São Paulo; Ribeirão Preto, Brazil; ²Institute for Systems Biology; Seattle, WA USA;

³Department of Computing and Mathematics; Faculdade de Filosofia Ciências e Letras de Ribeirão Preto; University of São Paulo; Ribeirão Preto, Brazil

[#]Present address: Universidade Paulista - Campi Ribeirão Preto; Ribeirão Preto, Brazil

Keywords: sense overlapping transcripts, IS1341, HgcC RNA family, IS1341-type, RNA-seq, non-coding RNA, overlapping transcripts, transposons, IS200/605, tetraloop

The existence of sense overlapping transcripts that share regulatory and coding information in the same genomic sequence shows an additional level of prokaryotic gene expression complexity. Here we report the discovery of ncRNAs associated with IS1341-type transposase (*tnpB*) genes, at the 3'-end of such elements, with examples in archaea and bacteria. Focusing on the model haloarchaeon *Halobacterium salinarum* NRC-1, we show the existence of sense overlapping transcripts (sotRNAs) for all its IS1341-type transposases. Publicly available transcriptome compendium show condition-dependent differential regulation between sotRNAs and their cognate genes. These sotRNAs allowed us to find a UUCA tetraloop motif that is present in other archaea (ncRNA family HgcC) and in a *H. salinarum* intergenic ncRNA derived from a palindrome associated transposable elements (PATE). Overexpression of one sotRNA and the PATE-derived RNA harboring the tetraloop motif improved *H. salinarum* growth, indicating that these ncRNAs are functional.

Introduction

Overlapping genomic signals have been discovered across all domains of life, showing that DNA sequences are multi-functional, harboring regulatory and coding information simultaneously. Transcriptome complexity includes not only interweaving promoter elements and intergenic transcription but also a plethora of transcripts that overlap annotated genes.^{1,2} Sense and antisense non-coding (ncRNA) transcripts have been identified, with a clear predominance of antisense transcripts characterization. Antisense ncRNAs are ubiquitous and affect diverse cellular processes, from transcription initiation to translation efficiency³; they have also been shown to regulate transposition events.⁴ Even with the widespread use of strand-specific sequencing technologies, the identification and characterization of sense overlapping transcripts have expanded more modestly, since the signals for these molecules are convoluted with mRNA signals and thus, can be difficult to detect.

Recently, we have identified sense ncRNAs overlapping the 5' end of coding sequences in archaea with focus in the haloarchaeon *Halobacterium salinarum* NRC-1.⁵ This extremophile thrives in environments with 4.5M NaCl concentration and has been used as a model organism for systems biology studies. As such, a massive amount of microarray and tiling array-based transcriptome data in diverse environmental conditions is currently available for this organism.^{6–11} In the present work, by revisiting the dataset from Koide *et al.*⁸, where transcriptome dynamics was assessed over a growth curve using high resolution tiling array data, we detected signals that could indicate the presence of sense overlapping transcripts inside IS1341-type transposase genes in *H. salinarum*.

IS1341-type transposases genes are also known in the literature as *tnpB* or *orfB*. They are part of the widespread IS200/605 family, the most ancient family in the archaeal domain of life.¹² Based on gene composition, IS200/605 family can be classified into 3 distinct groups: (a) IS200 group, with only the *tnpA* gene (also known as the IS200 gene); (b) IS605 group, where *tnpA*

© José Vicente Gomes-Filho, Livia Soares Zaramela, Valéria Cristina da Silva Italiani, Nitin S Baliga, Ricardo Z N Vêncio, and Tie Koide

*Correspondence to: Tie Koide; Email: tkoide@fmrp.usp.br

Submitted: 09/11/2014; Revised: 02/09/2015; Accepted: 02/09/2015

<http://dx.doi.org/10.1080/15476286.2015.1019998>

This is an Open Access article distributed under the terms of the Creative Commons Attribution-Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The moral rights of the named author(s) have been asserted.

and *tnpB* are found adjacent and (c) IS1341 group, where *tnpB* is found alone.^{13,14} Transposition of IS200/605 members requires the protein TnpA, which binds conserved palindromic sequences at the left end (LE) and right end (RE) of the insertion element.^{15,16} The mechanism of transposition is unique among IS elements since it requires single stranded DNA as substrates.^{15,17} The *tnpB* gene is not required for transposition^{14,18}, however, it is known to play a regulatory role by inhibiting excision and insertion of ISDra2 in *Deinococcus radiodurans*.¹⁹ Mechanisms of TnpB action are currently unknown.

Sense overlapping transcripts in IS1341-type transposase genes have been detected in other archaea as part of studies for the identification of ncRNAs. In *Sulfolobus solfataricus*, 2 ncRNAs overlapping transposases in the same strand were identified and in *Thermococcus kodakarensis*, 3 RNAs transcribed from transposase *loci* were detected.^{20,21} In *Pyrococcus abyssi*, ncRNAs are transcribed from intergenic repetitive *loci* and have sequence similarity to the 3' end of IS1341-type transposases.²² This apparent conservation of sense overlapping RNAs and similar sequences in intergenic ncRNAs prompted us to investigate not only the presence of these transcripts in *H. salinarum* NRC-1, but also to evaluate if they have a functional role. In the present work, we show that all IS1341-type transposases of *H. salinarum* possess sense overlapping transcripts, explore their expression profiles in diverse conditions, and show a conserved tetraloop motif that might be relevant for function of these ncRNAs.

Results and Discussion

H. salinarum has a large number of insertion sequences (114 according to ISfinder and ISbrowser^{23,24}, <https://www-is.biotoul.fr>), which are believed to be responsible for the unusual genome plasticity observed in this organism.²⁵ In *H. salinarum* NRC-1, reannotation based on R1 strain identified 10 IS1341-type transposases.²⁶ Five of these IS1341-type transposases are located adjacent to *tnpA*, 1 in divergent and 4 in convergent orientation (Table 1).

Analyzing publicly available tiling microarray data obtained in reference growth condition for *H. salinarum* NRC-1⁸, we discovered a conserved transcriptional signature in all 10 IS1341-type transposases genes. This signature features a change in the expression signal inside the insertion sequence near the 3' end, indicating the existence of possible sense overlapping transcripts (sotRNAs). Figure 1A-I illustrates this feature for VNG0042G. When the dynamics of these transcripts were analyzed over a growth curve, the putative sense overlapping transcripts showed an increase in relative expression levels, whereas the remaining of IS1341-type transposase gene showed decreasing levels (Fig. 1A-II). Tiling array data from TfbD transcription factor overexpression showed that sotRNAs presented decreasing expression levels, whereas IS1341-type genes presented increasing levels, stating an opposite regulation in varied conditions (Fig. 1A-III).⁸ Transcription signatures in the reference condition, together with the

dynamical information from the differential expression profiles provided evidence for the existence of sense overlapping transcripts at the 3' end of IS1341-type transposases genes in *Halobacterium salinarum* NRC-1 (Fig. 1A, B).

To validate their existence and define their boundaries with precision, we analyzed previously published small RNA-seq data (sRNA-seq).⁵ Figure 2A shows sRNA-seq based discovery of boundaries for VNG0042G associated sotRNA. An extended analysis (Fig. S1A) showed that all IS1341-type transposases in *Halobacterium salinarum* NRC-1 have an associated sotRNA. In addition, it was possible to map the 3' and 5' ends of all sotRNAs (Table 1) (Fig. S1A). We named these discovered features using "VNG_sot" as prefix and making reference to their cognate gene in the suffix. All sotRNAs mapped in this study start inside the coding sequence of IS1341-type transposase genes, at approximately 1100 nt from the start codon, with an average size of 218 nt and end, on average, 74 nt after gene's stop codon (Table 1).

Two targets were selected for further experimental validations: VNG_sot0042 and VNG_sot2652. C-RACE experiment for VNG_sot0042 confirmed its 5' and 3' ends (Fig. S2) and primer extension for VNG_sot2652 confirmed its 5' end (Fig. S3).

To further investigate the expression profiles of sotRNAs and IS1341-type transposases, we considered all tiling microarray data sets available in public databases (see Materials and Methods) and focused on probes designed to measure expression levels in the same region defined by sRNA-seq analysis. All sotRNAs were found to be expressed with diverse patterns of modulation, as would be expected for a typical functional transcript (Fig. S1B). Most of the sotRNAs were differentially expressed relative to the remaining signal of IS1341-type transposase cognate gene in different environmental conditions, which could possibly indicate some kind of regulation between these elements. Figure 2B shows the data for VNG0042G and its sotRNA. There is one case where the sotRNA (VNG_sot0286) is located antisense to the probable 3' UTR of VNG0287a (a putative RNA-binding protein). Analysis of their expression profiles over different conditions show that VNG_sot0286 and VNG0287a have anti-correlated expression profiles in some conditions, which could suggest a cis-antisense interaction (Fig. S4). None of the other 9 sotRNAs mapped in *H. salinarum* NRC-1 have this putative pattern for cis-antisense regulation.

The available data cannot conclusively differentiate between the 2 most probable sotRNA biogenesis hypotheses: primary transcription or processing products. When the upstream regions of the sotRNAs were analyzed, no classical BRE-TATA promoter sequence was found. Northern-blot experiments for 2 sotRNAs (VNG_sot0042 and VNG_sot2652, Fig. S5) might argue for processing, since the full-length transcript is not detected. Also, deep-sequencing data from primary transcript enriched libraries (dRNA-seq)⁵ only shows evidence of primary transcript enrichment for VNG_sot0026. However, it was previously reported that dRNA-seq data is prone to false negative results²⁷ which reinforces the finding that VNG_sot0026 is a primary transcript but does not necessarily disprove the others. In addition, some sotRNAs were also shown as primary transcripts in *T.*

Table 1. IS1341-type transposases (*tnpB* genes) analyzed in this work and their respective sense overlapping transcripts (sotRNAs)

Gene	IS200/605 family group	chromosome	<i>tnpB</i> start codon	<i>tnpB</i> stop codon	strand	sotRNA transcript start	sotRNA transcript end	sotRNA size [nt]	R1 strain equivalent genes
VNG0013C	IS1341	chr	12734	11478	reverse	11641	11428	214	OE1019R
VNG0026C	IS1341	chr	21789	20533	reverse	20701	20498	204	OE1040R
VNG0042G	IS1341	chr	35931	34651	reverse	34775	34570	206	OE1070R
VNG0044H	IS605 (VNG0043H – VNG0044H), convergent	chr	36715	37971	forward	37813	38042	230	OE1074F
VNG0286C	IS605 (VNG0285C – VNG0286C), convergent	chr	229443	230597	forward	230444	230629	186	OE1440F
VNG2652H	IS605 (VNG2653C – VNG2652H), divergent	chr	1989188	1987884	reverse	1988005	1987819	187	OE4727R
VNG6181H	IS605 (VNG6182H – VNG6181H), convergent	pNRC200	155603	154368	reverse	154508	154259	250	OE5062R
VNG6221H	IS605 (VNG6222H – VNG6221H), convergent	pNRC200	183054	181795	reverse	182008	181724	285	OE5102R
VNG6361G	IS1341	pNRC200	287504	288844	forward	288696	288915	220	OE6034F
VNG6406H	IS1341	pNRC200	316931	315786	reverse	315833	315631	203	OE6089R

Figure 1. Identification of sotRNAs in IS1341-type transposases using tiling array data. (A) (I) Tiling array signal in reference condition for VNG0042G and VNG_sot0042; (II) expression profiles over the growth curve⁸; (III) expression profiles during TfbD overexpression.⁸ Heatmaps are color-coded according to $\log_{10}$ expression ratios between each of the 13 time points relative to reference condition. (B) Tiling array signal in reference condition and expression profiles of the remaining 9 IS1341-type transposases (arrows in yellow for genes in forward strand and in orange for genes in reverse strand) and their sotRNAs (light blue arrows) over the growth curve and TfbD overexpression, as described in (A).

Figure 2. Analysis of sRNA-seq and expression profiles for VNG0042G and VNG_sot0042. **(A)** Strand-specific sRNA-seq data⁵ was used to map the 5' end of sotRNAs in IS1341-type transposases of *H. salinarum* NRC-1. sRNA-seq data is visualized as log₂ of total reads aligned in each genomic position for VNG0042G and VNG_sot0042. Enrichment of 5' ends of mapped reads are visualized as peaks right below small RNA-seq coverage. Light blue arrow: sotRNA. Dark orange arrows: genes annotated on reverse strand. **(B)** Expression profiles of VNG0042G (orange) and VNG_sot0042 (blue) in different environmental and genetic backgrounds. Each pair of orange/blue columns represent one published experimental condition. Data for the remaining genes and for each individual experiment is available in **Supplementary Figure 1**. Data for intergenic ncRNAs can be found in **Figure 4** and **Supplementary Figure 7**.

kodakarensis.²¹ Chromatin immunoprecipitation data (ChIP-chip) points to transcription factor binding sites at upstream vicinity of some sotRNAs along with statistically derived *H. salinarum* gene regulatory elements (Table S1).^{6,7,41} Finally, the distinct expression patterns of sotRNAs and their cognate genes across several experiments are unlikely to be generated by processing a single precursor molecule.

The presence in *H. salinarum* NRC-1 of a probable sotRNA in a nonfunctional IS1341-type transposase gene

(Fig. S1C), a truncated small 120 bp long pseudogene identical to OE5220R, suggests that sotRNA presence may somehow be the reason why these defective elements have not been lost from the genomes. Mining public available transcriptome data for other archaea and bacteria, we found that transcripts overlapping 3' ends of *tnpB* are more prevalent than previously appreciated. Besides the previously described *S. solfataricus* P2²⁰ and *T. kodakarensis* KOD1²¹ cases, we found examples in all archaea that contain IS1341 group

genes and RNA-seq data available at NCBI's SRA database: *Methanopyrus kandleri* AV19, *S. acidocaldarius* DSM639 and *P. furiosus* DSM3638. Moreover, we found bacterial examples in *Helicobacter pylori* 26695 and *Escherichia coli* K12 data (Figure S6, File S1). These observations suggest a conserved role for sotRNAs in prokaryotes. Computational secondary structure analysis reveals a very distinctive feature among almost all sotRNAs in *H. salinarum*: the presence of a stable hairpin motif near the end of the sequence. Using five different secondary structure prediction algorithms on each RNA sequence, it was possible to identify: (i) a clear tetraloop motif comprising an 8 bp long stem with a UUCA loop (VNG_sot0042, VNG_sot0044, VNG_sot6181 and VNG_sot6221); (ii) single nucleotide variants of this tetraloop (VNG_sot0286 with UUUA, VNG_sot0013 with CUCA and VNG_sot0026 with GUCA) (Fig. 3); and (iii) stem-loop structures created by other sequences (VNG_sot6406 with a 3 nt long loop in a 6 bp stem, VNG_sot2652 with a 6 nt long loop in a 10 bp stem, and VNG_sot6361 with a 3 nt long loop in a 4 bp stem) (File S2). While the 5 algorithms are considerably different in their approaches and underlying assumptions to predict secondary structures²⁸, they all made consistent predictions of tetraloops and collectively added to the significance of our discovery of this feature.

One of the key characteristics of IS200/605 family is the presence of conserved imperfect palindromic sequences that form a hairpin-like structure at DNA level at both left and right ends of the insertion elements, sometimes just upstream of CDS stop codons.^{15,16,19} Therefore, IS1341-type transposase genes should harbor the right end (RE) of the IS element, which may explain the sotRNA motifs found as a retained structure formed at DNA level, also stable at RNA level. We named the motif found, the RE-like tetraloop.

Similarly, in *P. abyssi*, sequence alignments of intergenic ncRNAs sRK48 and sRK52 with other thermococcal genomes showed similar conserved structures and the most conserved one (P1 loop)²² corresponds to the RE-like tetraloop,

Figure 3. Tetraloop motif found in most sotRNAs in *H. salinarum* NRC-1. Colors represent base pairing groups and color saturation corresponds to conservation, parenthesis correspond to pairing and dots to unpaired regions. Gray blocks below the sequences represent a histogram of the most conserved nucleotide. Raw data is available at **Supplementary File 2**.

Table 2. Intergenic RNAs harboring the RE-like tetraloop motif

ncRNA	VNG_R0052	VNG_R6334
Start	14397	268030
End	14299	268304
Length	99	275
Chr	chr	pNRC200
Strand	reverse	forward
RE-like Start	14359	268256
RE-like End	14340	268277

reinforcing the probable functionality of this region in archaea. A sequence similarity search for the RE-like tetraloops in *H. salinarum* NRC-1 genome showed that, besides all sotRNAs regions, 3 intergenic regions also have this secondary structure: 14335 to 14359 on main chromosome, 268252 to 268281 on pNRC200 plasmid, and 19532 to 19556 on main chromosome (File S3). From these 3, only the first 2 had sufficient sRNA-seq read coverage signal: VNG_R0052 and VNG_R6334 (Table 2). We determined the 5' and 3' end of these transcripts using RNA-seq data (Fig. 4A and Fig. S7A) and verified the expression profiles over diverse environmental and genetic perturbations. Intergenic ncRNAs showed differential expression, with trends similar to sotRNAs (Fig. 4B and Fig. S7B).

When we analyze the sequence of the intergenic ncRNA VNG_R0052, we can find not only the stem loop corresponding to the RE of the IS element, but also, fused to it at the 5' end, we can find part of the left end (LE) sequence of IS200/605 family (File S2). This phenomena has been previously observed in *Haloquadratum walsbyi* as a result of transposase deletion, generating elements that are denominated PATEs (palindrome associated transposable elements).²⁹ We show that a PATE element is indeed expressed at the RNA level and similarity search in the NCBI database shows that VNG_R0052 is not an exception: fused LE and RE are common among haloarchaea (File S4). We predict that, as in VNG_R0052 case, PATEs might also be functionally transcribed in other organisms. Moreover, since PATEs have counterparts in bacteria (REPINS, found for example in *Pseudomonas fluorescens*³⁰), we extend this prediction to bacteria.

Interestingly, the RE-like tetraloop is a previously unannotated structural motif in the HgcC ncRNA family (RFAM v12 accession code RF00062) (Fig. S8). HgcC family was first discovered in the genome of *Methanococcus jannaschii* by an *in silico* analysis designed for the identification of high guanine-cytosine content regions.³¹ Yet, functions of this ncRNA family are still unknown, thus the presence of the RE-like tetraloop inside RNAs from this family might help further studies focusing on elucidating their function.

To test if RNAs harboring this conserved structural motif may have some functional role, we overexpressed one sotRNA (VNG_sot0042) and one intergenic RNA (VNG_R0052), and analyzed the consequences on growth dynamics in standard batch culture. Remarkably, both strains harboring the

Figure 4. (A) Strand-specific sRNA-seq data⁵ was used to map 5' ends of intergenic RNAs in *H. salinarum* NRC-1. sRNA-seq data is visualized as log₂ of total reads aligned in each genomic position for VNG_R0052. Enrichment of 5' ends of mapped reads are visualized as peaks right below small RNA-seq coverage profile. (B) Expression profiles of VNG_R0052 in different environmental and genetic backgrounds.

RE-like tetraloops had improved growth relative to the control strain providing preliminary evidence that this conserved motif is somehow functional in *H. salinarum* NRC-1 (Fig. 5). In the future, the specific mechanism by which the tetraloop acts can potentially be investigated by analyzing the effect of systematically disrupting its stem and loop structures on this growth phenotype.

The exact role of IS1341-type transposases is still an open problem but recent reports indicate that they have an inhibitory effect on IS excision and insertion.¹⁹ Given that all transcripts analyzed in this work (sotRNAs and a PATE-derived RNA)

harbor similar motifs, we might speculate whether these molecules are related to this mechanism. A possibility is that this motif could work as a competitor for TnpA binding, and thus, making it less available for binding at single stranded DNA regions. PATE-derived RNAs could also present similar functionality. Thus, IS1341 group of transposases could act as reservoirs of stem loops, as previously hypothesized.³² Further experimental work would be required to test this hypothesis, nevertheless, our data point to the functionality of ncRNAs derived from IS1341-type transposases that harbor the conserved RE-like tetraloop motif in their sequences.

Figure 5. Growth curve of 4 different strains of *H. salinarum* NRC-1 under reference conditions. Strains overexpressing RNAs (VNG_R0052 and VNG_sot0042) that harbors the RE-like tetraloop present higher growth when compared to control strains (NRC-1, wild type and PMTF-0, which is the wild type strain with an empty overexpression vector).

Conclusions

In the present work, we report the discovery of sense overlapping transcripts in IS1341-type transposases in *H. salinarum* NRC-1 and show that they are generally differentially expressed relative to their cognate transposase gene in diverse environmental conditions. In addition, we detected the presence of a conserved tetraloop motif in several sRNAs and in intergenic ncRNAs that correspond to palindrome associated transposable elements (PATEs). We provide evidence that transcripts harboring this conserved tetraloop motif, which are very similar to right end (RE) of the insertion sequence, are probably functional since their overexpression improved *H. salinarum* NRC-1 growth.

Materials and Methods

H. salinarum NRC-1 transcriptome analysis

Publicly available RNA-seq data was used to locate ncRNA transcripts with high spatial resolution. Publicly available tiling microarray data was used to interrogate expression dynamics over time or combinatorial experimental conditions for the ncRNAs found in this work. The RNA-seq data is composed by deep-sequencing of replicated sRNA enriched libraries (<200 bp) obtained from standard *H. salinarum* NRC-1 growth conditions⁵ available at NCBI's SRA database (SRX433542). Tiling microarray data is a compendium of all publicly available datasets (Aug/2014) obtained by different platforms (NCBI's GEO database accession numbers: GPL17005, GPL13426, GPL7369, GPL7255 and GPL8468).

The data set used were: 13 points of a standard growth-curve (GSE12923); overexpression of the TfbD transcription factor (GSE15788)⁸; knockouts of transcription factor Tfb and overexpression of synthetically enhanced versions of it (GSE31308)⁹; knockout of ribonuclease VNG2099C (GSE45988)¹⁰; variation of extracellular salinity concentrations for 3 different salts (GSE53544).¹¹ Normalized expression ratios appropriate for each study were obtained directly from GEO records as published. Growth curve dynamics and TfbD overexpression tiling array data were visualized in detail at probe level using Gaggie Genome Browser.³³

Sequence similarity search in public databases

Sense overlapping transcripts for specific *tnpB* genes were found in different organism using standard BLASTn-SRA suite sequence similarity search. The RNA-seq datasets considered comprised archaeal and bacterial data: *S. acidocaldarius* DSM639 (SRA accession SRX327494), *Methanopyrus kandleri* AV19 (accession SRP019041), *H. pylori* 26695 (accessions SRX014018, SRX014030 to SRX014033 and SRX014050) and *E. coli* K12 (accession SRX254762). *H. salinarum* NRC-1 intergenic ncRNA containing RE-like tetraloop motifs were found through standard BLASTn searches using all 10 sRNAs sequences as queries.

RNA secondary structure computational analysis

RNA sequences of all 10 sRNAs and 2 intergenic ncRNAs were used for structural predictions using 5 different methods: RNAfold-MFE, which predicts the Minimum Free Energy (MFE) structure; RNAfold using thermodynamic partition function centroids algorithm³⁴; ContextFold, which uses a very large number of parameters in a machine learning approach³⁵; Vsfold5, which allow the prediction of pseudoknots by using sequential (5' to 3') folding and thermodynamically most-probable folding pathways³⁶ and Cylofold, which simulates the 3D folding process in a coarse-grain model.³⁷ All predictions are shown in extended dot-bracket format to include pseudoknots. All 5 methods were run using their respective web-server versions and default parameters.

Strains and culturing

H. salinarum NRC-1 was grown in enriched complex media (CM) consisting of 25% NaCl, 2% MgSO₄·7H₂O, 0.2% KCl, 0.3% Na-citrate and 1% peptone, at 37°C with constant light and agitation of 225 rpm. When required, the media was supplemented with 20 µg/mL of mevinolin (A.G. Scientific, San Diego, CA). *E. coli* strain DH5α was grown at 37°C with air shaking at 225 rpm in Luria-Bertani media and supplemented with 50 µg/mL of carbenicillin (Sigma) when necessary.³⁸

RNA preparation

Total RNA was prepared using the MirVana RNA extraction kit (Ambion) and treated with Turbo DNase I (Fermentas). Samples were analyzed by PCR to rule out DNA contamination.

Primer extension

10 µg of RNA was hybridized with 1pmol of 5' fluorophore (VIC or NED) labeled oligonucleotide and 1 µl of 10 mM dNTPs in a total volume of 10µl (Temperature cycle for annealing: 95°C 5', 62°C 15', 59°C 15', 53°C 15', 50°C 15'). After annealing, a Synthesis Mix 2 µl 10x RT Buffer (200 mM Tris-HCl (pH 8.4), 500 mM KCl); 4 µl MgCl₂(25 mM); 2 ul 0.1 M DTT; 1 µl RNase Out; 1 µl SuperScript III (Invitrogen) was added and the reaction was incubated at 50°C for 50'. The cDNA was purified and analyzed in an ABI 3500XL (Applied Biosystems). Data processing was made using Gene Mapper V5.0 (Life Technologies).

Northern blot

For Northern blot analyzes, 50 µg of total RNA treated with RNase-free DNaseI (Fermentas) was separated on polyacrylamide gel (8% acrylamide:bisacrylamide [29:1], 8M urea, 1xTris–borate–EDTA buffer). RNAs were transferred to Hybond-N+ membranes (GE Healthcare) and hybridized with ³²P-labeled oligonucleotides (5'-GACGTAGCCTGTTCTT-GAGGTGCGAACGCACCAGTTTCAT-3' for VNG0042G and VNG_sot0042; 5'-CCCTGTTCCGCGAGGGAACGG-GATGAGGTTGGGCGGCTTA-3' for VNG2652H and VNG_sot2652) using Rapid-hyb buffer (GE Healthcare). Signals were detected by autoradiography using a M35A X-Omat Processor (Kodak).

Circularization and RACE (Rapid Amplification of cDNA Ends)

5' and 3' ends of RNAs were determined by RNA circularization.³⁹ Briefly, RNA free of DNA contamination was treated with Tobacco Acid Phosphatase (Epicentre) following manufacturer's instructions and treated with T4 RNA ligase (Thermo Scientific) for intramolecular ligation of 5' and 3' ends. cDNA was synthesized using SuperScript III First Strand Synthesis System (Life Technologies) using random hexamers. Divergent PCR (CircVNG0042-F — 5'-ATGGTTTACGACGTAGCC-TGTTTC —3', CircVNG0042-R — 5'-GTTCTGTTTGA-CAATGAAACTG -3') using circular cDNA was made using Kapa Taq ReadyMix (Kapa Biosystems). Amplification products were purified from agarose gels and cloned into pGEM T-easy Vector (PROMEGA) for sequencing. RACE assays were performed as previously described⁴⁰ using 5'/3' RACE Kit 2nd Generation (Roche) for VNG_R0052 using the following primer sequences: RSP1 — 5' - GAAGGGCGAGGCTTCTCTC — 3', RSP2 — 5' - CGAGGCTTCTCTCTCGATTTC — 3', RSP3 — 5' - CTGTTACGGAGAATCGAGAAG — 3', RSP4 — 5' GAATCGAGAAGAAAGCCTCG — 3'.

General cloning procedures

VNG_R0052 and VNG_sot0042 were cloned into pMTF over-expression vector for *H. salinarum* NRC-1.⁴¹ DNA fragments were amplified by PCR using 50 ng of *H. salinarum* NRC-1 genomic DNA, 10pmol of specific oligos, 6.25 µl of 2x KAPA2G Fast ReadyMix (with loading dye) (Product Number: KK5102), 0.25 µl

of MgCl₂ (25 mM), 3.875 µl of DMSO and DEPC treated water to a final volume of 12.5µl. The oligos used for VNG_R0052 were 5'-GACATATGCGGAGAATCGAGAAGA AAGCC-3' and 5'-GGGGATCCTCAGTCTAACT-TCTGGGGTGG-3'; and for VNG_sot0042 were 5'-GCCGGGCATATGACGTATCTCC-GAGTCCCGTCA-3' and 5'-GCGGGGGGATCCACAGGA-TAGCTGAGACTGCAGCAT-3'. The selected oligos added a BamHI site located at the 5' region and a NdeI site located at 3' region. PCR products were purified from 1% agarose gel by illustra GFX PCR DNA and Gel Extraction Kit (GE Healthcare). All fragments were cloned as BamHI/NdeI fragments into pMTF vector and transformed in chemically competent *E. coli* DH5α. Plasmidial DNA was extracted using Wizard Plus SV Miniprep Kits (Promega) and sequenced using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). The confirmed plasmids were used to transform *H. salinarum* NRC-1. Transformant selection was obtained by plating the cells into solid CM supplemented with mevinolin (20 µg/ml).

Growth curves

For phenotypical characterization, isolated colonies of each *H. salinarum* strain were grown in liquid CM, supplemented with 20µg/ml of mevinolin when needed, until they reached OD600 = 0.5. Each culture was diluted to OD600 = 0.05 and grown under standard conditions for 8 days, all growth curves were made in 2 biological and technical replicates.⁴²

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Silvia Epifanio for technical support. We thank reviewers for constructive criticisms that improved our work.

Funding

This work was supported by Projeto Jovem Pesquisador em Centros Emergentes da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, <http://fapesp.br/en/>) [09/09532–0 to TK]; Edital Universal do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [473660/2013–0 to TK, 470120/2009–6 to TK, 476724/2013–9 to RZNV]; 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) [1640/2009 to TK]; Núcleo de Pesquisa em Ciência Genômica (NAP-CG) da Universidade de São Paulo; and fellowships FAPESP [11/07487–7 to LSZ, 2013/21522–5 to JVGf, 10/10195–5 to VCSI].

Supplemental Material

Supplemental data for this article can be accessed on the publisher's website.

References

- Kapranov P, Willingham AT, Gingeras TR. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* 2007; 8:413-23; PMID:17486121; <http://dx.doi.org/10.1038/nrg2083>
- Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* 2010; 11:9-16; PMID:19935729; <http://dx.doi.org/10.1038/nrg2695>
- Pelechano V, Steinmetz LM. Gene regulation by antisense transcription. *Nat Rev Genet* [Internet] 2013 [cited 2014 Jul 14]; 14:880-93. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24217315>; <http://dx.doi.org/10.1038/nrg3594>
- Brantl S. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr Opin Microbiol* 2007; 10:102-9; PMID:17387036; <http://dx.doi.org/10.1016/j.mib.2007.03.012>
- Zaramela LS, Vencio RZN, ten-Caten F, Baliga NS, Koide T. Transcription start site associated RNAs (TSSaRNAs) are ubiquitous in all domains of life. *PLoS One* 2014; 9:e107680; PMID:25238539; <http://dx.doi.org/10.1371/journal.pone.0107680>
- Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC, et al. A predictive model for transcriptional control of physiology in a free living cell. *Cell* [Internet] 2007 [cited 2011 Jul 22]; 131:1354-65. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18160043>; PMID:18160043; <http://dx.doi.org/10.1016/j.cell.2007.10.053>
- Brooks AN, Reiss DJ, Allard A, Wu W-J, Salvanha DM, Plaisier CL, Chandrasekaran S, Pan M, Kaur A, Baliga NS. A system-level model for the microbial regulatory genome. *Mol Syst Biol* 2014; 10:740; PMID:25028489; <http://dx.doi.org/10.15252/msb.20145160>
- Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo F-Y, et al. Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol* [Internet] 2009 [cited 2011 Aug 12]; 7:5285. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2710873&tool=pmcentrez&rendertype=abstract>; PMID:19536208; <http://dx.doi.org/10.1038/msb.2009.42>
- Turkarslan S, Reiss DJ, Gibbins G, Su WL, Pan M, Bare JC, Plaisier CL, Baliga NS. Niche adaptation by expansion and reprogramming of general transcription factors. *Mol Syst Biol* 2011; 7:554; PMID:22108796; <http://dx.doi.org/10.1038/msb.2011.87>
- Wurtmann EJ, Ratushny AV, Pan M, Beer KD, Aitchison JD, Baliga NS. An evolutionarily conserved RNase-based mechanism for repression of transcriptional positive autoregulation. *Mol Microbiol* 2014; 92:369-82; PMID:24612392; <http://dx.doi.org/10.1111/mmi.12564>
- Beer KD, Wurtmann EJ, Pintel NN, Baliga NS. Model organisms retain an "ecological memory" of complex ecologically relevant environmental variation. *Appl Environ Microbiol* [Internet] 2014 [cited 2014 Jul 31]; 80:1821-31. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24413600>; PMID:24413600; <http://dx.doi.org/10.1128/AEM.03280-13>
- Filee J, Siguier P, Chandler M, Filée J, Siguier P, Chandler M. Insertion sequence diversity in archaea. *Microbiol Mol Biol Rev* [Internet] 2007 [cited 2013 Mar 31]; 71:121-57. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1847376&tool=pmcentrez&rendertype=abstract>; PMID:17347521; <http://dx.doi.org/10.1128/MMBR.00031-06>
- Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: the genomic impact and diversity. *FEMS Microbiol Rev* [Internet] 2014 [cited 2014 Jul 15]; 38(5):865-91. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24499397>
- Ton-Hoang B, Guynet C, Ronning DR, Cointin-Marty B, Dyda F, Chandler M. Transposition of ISHP608, member of an unusual family of bacterial insertion sequences. *EMBO J* 2005; 24:3325-38; PMID:16163392; <http://dx.doi.org/10.1038/sj.emboj.7600787>
- Barabas O, Ronning DR, Guynet C, Hickman AB, Ton-Hoang B, Chandler M, Dyda F. Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. *Cell* 2008; 132:208-20; PMID:18243097; <http://dx.doi.org/10.1016/j.cell.2007.12.029>
- Hickman AB, James JA, Barabas O, Pasternak C, Ton-Hoang B, Chandler M, Sommer S, Dyda F. DNA recognition and the precleavage state during single-stranded DNA transposition in *D. radiodurans*. *EMBO J* 2010; 29:3840-52; PMID:20890269; <http://dx.doi.org/10.1038/emboj.2010.241>
- Pasternak C, Ton-Hoang B, Coste G, Bailone A, Chandler M, Sommer S. Irradiation-Induced *Deinococcus radiodurans* Genome Fragmentation Triggers Transposition of a Single Resident Insertion Sequence. *PLoS Genet* [Internet] 2010; 6:e1000799. Available from: <http://dx.doi.org/10.1371/journal.pgen.1000799>; PMID:20090938; <http://dx.doi.org/10.1371/journal.pgen.1000799>
- Kersulyte D, Kalia A, Zhang M, Lee H-K, Subramaniam D, Kiuduliene L, Chalkauskas H, Berg DE. Sequence Organization and Insertion Specificity of the Novel Chimeric ISHP609 Transposable Element of *Helicobacter pylori*. *J Bacteriol* [Internet] 2004; 186:7521-8. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC524915/>; PMID:15516563; <http://dx.doi.org/10.1128/JB.186.22.7521-7528.2004>
- Pasternak C, Dulerme R, Ton-Hoang B, Debuchy R, Siguier P, Coste G, Chandler M, Sommer S. ISDra2 transposition in *Deinococcus radiodurans* is downregulated by TnpB. *Mol Microbiol* 2013; 88:443-55; PMID:23461641; <http://dx.doi.org/10.1111/mmi.12194>
- Zago MA, Dennis PP, Omer AD. The expanding world of small RNAs in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol Microbiol* [Internet] 2005 [cited 2013 May 27]; 55:1812-28. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15752202>; PMID:15752202; <http://dx.doi.org/10.1111/j.1365-2958.2005.04505.x>
- Jager D, Forstner KU, Sharma CM, Santangelo TJ, Reeve JN. Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis*. *BMC Genomics* 2014; 15:684; PMID:25127548; <http://dx.doi.org/10.1186/1471-2164-15-684>
- Phok KK, Moisan A, Rinaldi D, Brucato N, Carpusis AJ, Gaspin C, Clouet-d'Orval BB. Identification of CRISPR and riboswitch related RNAs among novel noncoding RNAs of the euryarchaeon *Pyrococcus abyssi*. *BMC Genomics* [Internet] 2011 [cited 2011 Aug 8]; 12:312. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3124441&tool=pmcentrez&rendertype=abstract>; PMID:21668986; <http://dx.doi.org/10.1186/1471-2164-12-312>
- Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* 2006; 34:D32-6; PMID:16381877; <http://dx.doi.org/10.1093/nar/gkj014>
- Chandler M, Kichenadja P, Siguier P, Pe J, Pérochon J, Chandler M. ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. *Nucleic Acids Res* [Internet] 2010 [cited 2014 Jul 29]; 38:D62-8. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808865&tool=pmcentrez&rendertype=abstract>; PMID:19906702; <http://dx.doi.org/10.1093/nar/gkp947>
- Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Shrogha J, et al. Genome sequence of *Halobacterium* species NRC-1. *Proc Natl Acad Sci U S A* [Internet] 2000; 97:12176-81. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3111458&tool=pmcentrez&rendertype=abstract>; PMID:11016950; <http://dx.doi.org/10.1073/pnas.190337797>
- Pfeiffer F, Schuster SC, Broicher A, Falb M, Palm P, Rodewald K, Ruepp A, Soppa J, Tittor J, Oesterhelt D. Evolution in the laboratory: the genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1. *Genomics* 2008; 91:335-46; PMID:18313895; <http://dx.doi.org/10.1016/j.ygeno.2008.01.001>
- Shao W, Price MN, Deutschbauer AM, Romine MF, Arkin AP. Conservation of transcription start sites within genes across a bacterial genus. *MBio* 2014; 5:e01398-14; PMID:24987095; <http://dx.doi.org/10.1128/mBio.01398-14>
- Puton T, Kozlowski LP, Rother KM, Bujnicki JM. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res* 2013; 41:4307-23; PMID:23435231; <http://dx.doi.org/10.1093/nar/gkt101>
- Dyall-Smith ML, Pfeiffer F, Klee K, Palm P, Gross K, Schuster SC, Rampp M, Oesterhelt D. *Halobacterium walsbyi*: limited diversity in a global pond. *PLoS One* [Internet] 2011 [cited 2013 Apr 30]; 6:e20968. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3119063&tool=pmcentrez&rendertype=abstract>; PMID:21701686; <http://dx.doi.org/10.1371/journal.pone.0020968>
- Bertels F, Rainey PB. Curiosities of REPINs and RAYTs. *Mob Genet Elements* [Internet] 2011; 1:262-8. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3337134/>; PMID:22545236; <http://dx.doi.org/10.4161/mge.18610>
- Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci U S A* [Internet] 2002; 99:7542-7. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=124278&tool=pmcentrez&rendertype=abstract>; PMID:12032319; <http://dx.doi.org/10.1073/pnas.112063799>
- Delhas N. Stem loop sequences specific to transposable element IS605 are found linked to lipoprotein genes in *Borrelia* plasmids. *PLoS One* [Internet] 2009 [cited 2014 Jul 28]; 4:e7941. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2775950&tool=pmcentrez&rendertype=abstract>; PMID:19936201; <http://dx.doi.org/10.1371/journal.pone.0007941>
- Bare JC, Koide T, Reiss DJ, Tenenbaum D, Baliga NS. Integration and visualization of systems biology data in context of the genome. *BMC Bioinformatics* 2010; 11:382; PMID:20642854; <http://dx.doi.org/10.1186/1471-2105-11-382>
- Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Website. *Nucleic Acids Res* [Internet] 2008; 36:W70-4. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447809/>; PMID:18424795; <http://dx.doi.org/10.1093/nar/gkn188>
- Zakov S, Goldberg Y, Elhadad M, Ziv-Ukelson M. Rich parameterization improves RNA structure prediction. *J Comput Biol* 2011; 18:1525-42; PMID:22035327; <http://dx.doi.org/10.1089/cmb.2011.0184>
- Dawson WK, Fujiwara K, Kawai G. Prediction of RNA Pseudoknots Using Heuristic Modeling with Mapping and Sequential Folding. *PLoS One* [Internet] 2007; 2:e905. Available from: <http://dx.plos.org/10.1371/journal.pone.0000905>; PMID:17878940; <http://dx.doi.org/10.1371/journal.pone.0000905>
- Bindewald E, Kluth T, Shapiro BA. CyloFold: secondary structure prediction including pseudoknots. *Nucleic Acids Res* 2010; 38:W368-72; PMID:20501603; <http://dx.doi.org/10.1093/nar/gkq432>
- Gerhardt P. Methods for General and Molecular Bacteriology [Internet]. American Society for Microbiology; 1994. Available from: <http://books.google.co.in/books?id=FGYXAAIAAJ>
- Abdelrahman YM, Rose LA, Belland RJ. Developmental expression of non-coding RNAs in Chlamydia trachomatis during normal and persistent growth. *Nucleic Acids Res* 2011; 39:1843-54; PMID:21051342; <http://dx.doi.org/10.1093/nar/gkq1065>

40. Schaefer BC. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal Biochem* 1995; 227:255-73; PMID:7573945; <http://dx.doi.org/10.1006/abio.1995.1279>
41. Facciotti MT, Reiss DJ, Pan M, Kaur A, Vuthoori M, Bonneau R, Shannon P, Srivastava A, Donohoe SM, Hood LE, et al. General transcription factor specified global gene regulation in archaea. *Proc Natl Acad Sci U S A* [Internet] 2007; 104:4630-5. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1838652&tool=pmcentrez&rendertype=abstract>; PMID:17360575; <http://dx.doi.org/10.1073/pnas.0611663104>
42. Robb FT, DasSarma S, Fleischmann EM. *Archaea: a laboratory manual*. Halophiles [Internet]. Cold Spring Harbor Laboratory; 1995. Available from: <http://books.google.com.br/books?id=sYvjSAAACAJ>