UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

FELIPE TEN CATEN

A transcrição pervasiva na archaea *Halobacterium salinarum* NRC-1
e a identificação de novos transcritos

Ribeirão Preto
2016

FELIPE TEN CATEN

A transcrição pervasiva na archaea *Halobacterium salinarum* NRC-1
e a identificação de novos transcritos

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:
Bioinformática

Orientadora: Prof^a Dr^a Tie Koide
Co-orientador: Prof. Dr. Ricardo Z. N.
Vêncio

Ribeirão Preto
2016

TEN-CATEN, F. **A transcrição pervasiva na archaea *Halobacterium salinarum* NRC-1 e a identificação de novos transcritos.** 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 esse trabalho aos meus pais Pedro Marino e Vera,
que provocaram em mim a curiosidade pela natureza*

Agradecimentos

Agradeço a minha família por todo o apoio durante essa jornada, que serviu de base para a realização desse trabalho. Ao meu pai Pedro Marino e a minha mãe Vera, que sempre se dedicaram para que olhássemos além. Aos meus irmãos Elaine, Alexandre e Roberto, que sempre tiveram carinho com o irmão mais novo e são minha referência para trilhar novos caminhos.

Agradeço ao meu filho Pedro, que desde muito pequeno teve que assimilar os conceitos de distância e saudade. Agradeço também a Daiane, que com muito empenho e dedicação enfrenta de forma exemplar os desafios da maternidade.

Agradeço a minha namorada Ana Carolina, que está sempre ao meu lado, incentivando, aconselhando, ajudando e compartilhando felicidades e tristezas. Essa jornada se tornou mais alegre e inspiradora ao seu lado. Te amo!

Agradeço a Prof^a Tie e ao Prof. Ricardo Vêncio, pela confiança e ensinamentos passados ao longo do doutorado. Vocês contribuíram enormemente para minha formação e me deram segurança para enfrentar novos desafios.

Agradeço ao Prof. Elgion Loreto, que me deu a primeira oportunidade em um laboratório e revelou a beleza de observar o mundo pelos olhos da Ciência.

Agradeço a Bruna Corrêa, a Patrícia Pauli e a Marjorie Pontelli, pelo companheirismo e por tornar a saudade de casa um pouco menor. A RePam sempre será lembrada com boas recordações.

Agradeço a Dona Lourdes e a Tia Edi, por me acolherem com tanto carinho e serem minha família em São Paulo.

Agradeço ao Ricardo Silva, Leonardo Rippel, ao Diego Martinez Salvanha e ao Ricardo Cacheta Waldemarin, por me receberem quando cheguei aqui a alguns anos atrás. Vocês são grandes amigos e ajudaram muito na minha adaptação.

Agradeço aos amigos e colegas Livia Soares Zaramela, Ricardo Silva, José Vicente Gomes Filho, Marjorie Cornejo Pontelli, Marcos Abraão, Bruna Corrêa, Diego Martinez Salvanha, Guilherme Firman, Gabriela Guardia, Victoria Maruyama, Andrea Aguiar, André Medina, Rafael Silva Rocha e Gilvan Pessoa Furtado por todos esses anos de convivência, trabalho, cafés, conversas e festas. Vocês são pessoas muito especiais para mim.

Agradeço aos integrantes do LaBiSisMi NextGen, Alan, João Paulo, Junier, Jadson e Victor, pela amizade e boas conversas nos últimos meses. Espero que tenham uma boa jornada pelo laboratório e que ela seja tão inspiradora quanto foi para mim.

Agradeço a Sílvia Helena Epifânio, pela grande amizade e por todo o suporte técnico. Os resultados obtidos durante esse trabalho também são fruto da sua dedicação e empenho. Você é um exemplo de pessoa e de profissional.

Agradeço ao Lucas Lorenzon, Diego Ângelo Duarte, Fábio Natividade, André Brandão, Felipe Valer, Lucas Tavares, Gustavo Tibério, Felipe Canto, Luiz Carlos Vieira, Tayllon

Garcia, Mário Cervoni, João Paulo Mesquita e Henrique Oliveira Frank pela amizade e pelos churrascos. Além de me deixarem bem seguro de que não sirvo para jogar futebol.

Agradeço aos Mineiros, Gaúchos e associados Ana Carolina Santana, Diego Ângelo Duarte, Amanda Antoniêto, Fábio Natividade, Flávia Mendonça, André Brandão, Felipe Canto, Geisa Santos, Luiz Carlos Vieira, Karol Nogueira, Lucas Tavares, Felipe Valer, Aline Tolardo, Tayllon Garcia e Rélber Gonçalves por toda a amizade e companheirismo ao longo desses anos.

Agradeço a Juliana Doblas Massaro, a Sandra Rodrigues e a Flávia Tremeschin, por toda a ajuda e suporte com o sequenciamento.

Agradeço a Patrícia Martorelli e a Cristiane de Fátima Braulino por seu trabalho junto a secretaria de pós-graduação e por diminuírem as dificuldades que o termo “interunidades” pode trazer.

Agradeço a todos os integrantes do laboratório do Prof. Roberto do Nascimento Silva por disponibilizarem equipamentos e ajuda para a realização dos experimentos conduzidos durante essa trabalho.

Agradeço a todos os integrantes do laboratório da Prof^a. Angela Kaysel Cruz que também foram muito prestativos e solícitos, auxiliando em diferentes etapas da realização desse trabalho.

Agradeço a Alexandra Elbakyan, a mente brilhante por trás da ferramenta Sci-Hub. Além de “remover todas as barreiras no caminho da ciência”, seu trabalho contribui de maneira imensurável para o progresso do conhecimento, questionando os modelos arcaicos de publicação científica.

Agradeço a todas as pessoas ao redor do mundo que dedicaram alguns minutos de seu tempo postando questionamentos e respostas nos diferentes fóruns disponíveis na Internet. Vocês representam o que há de melhor nesse ambiente: o compartilhamento de conhecimento em qualquer lugar e a qualquer momento. Vocês mitigaram dúvidas e dificuldades, apontando caminhos para obtenção de resultados aqui apresentados. Vocês também fazem parte desse trabalho.

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

Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP
Processo N° 2011/14455-4; Processo N° 2015/12012-9

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 - FAEPA

"Nature is a tinkerer, not an inventor"
François Jacob, 1977

Resumo

TEN-CATEN, F. **A transcrição pervasiva na archaea *Halobacterium salinarum* NRC-1 e a identificação de novos transcritos.** 2016, 168 f. Tese (Doutorado) – Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, 2016

A caracterização em larga escala do transcrito de diferentes organismos revelou um cenário complexo da expressão gênica, levando a identificação de inúmeros transcritos produzidos ao longo dos genomas de eucariotos e procariotos. Esse fenômeno recebeu o nome de transcrição pervasiva e tem sido fonte de estudos na busca de novos RNAs com importâncias regulatórias e também transcritos envolvidos na tradução de proteínas ainda não caracterizadas. A abundância de dados de transcrito e proteômica, além de informações completas a respeito do genoma, fazem do extremófilo halófilo *Halobacterium salinarum*, um organismo modelo ideal para os estudos da transcrição pervasiva. Esse micro-organismo pertence ao grupo Archaea, o último dos três domínios da vida a ser descrito e com características compartilhadas entre bactérias e eucariotos. Através do uso da técnica de *differential* RNA-seq (dRNA-seq), a qual permite a distinção entre transcritos primários e processados, identificamos 179 TSSaRNAs em *H. salinarum*, esses pequenos RNAs estão associados ao início de transcrição e ainda não haviam sido descritos em archaea. A aplicação do dRNA-seq em amostras de RNA extraídas ao longo da curva de crescimento permitiu a identificação de 4540 TSS no genoma de *H. salinarum* NRC-1. Parte desses inícios de transcrição está localizada *upstream* a genes conhecidos, permitindo a identificação de inícios de transcrição em 1545 genes. 59,2% desses inícios de transcrição estão localizados até 10 pb. de distância do códon de início de tradução, confirmando a ausência de regiões UTRs em grande parte dos genes. A análise de expressão, em diferentes condições, das regiões relacionadas a inícios de transcrição *antisense* a genes revelou que a maioria dessas regiões apresenta um perfil de expressão correlacionado com os genes na fita oposta, indicando um possível papel regulatório desses transcritos. De forma similar, a análise da expressão de inícios de transcrição intergênicos permitiu a identificação de 132 regiões diferencialmente expressas e que não estão relacionadas a nenhum outro elemento no genoma de *H. salinarum* NRC-1. A análise comparativa com dados de proteômica revela que algumas dessas regiões podem estar envolvidas com a produção de pequenas proteínas. Além disso, a identificação de 1365 inícios de transcrição internos a genes sugere que a produção de transcritos intragênicos (intraRNAs) seja um fenômeno amplamente distribuído no genoma desse halófilo. Experimentos de *Northern blot* confirmaram a produção de um transcrito correspondente a porção final do gene VNG_RS05220, e experimentos de *Western blot* revelaram que a tradução desses intraRNAs é responsável pela produção de pequenas proteínas correspondentes a domínios proteicos individuais, com importante papel funcional em condições específicas de crescimento. A análise de inícios de transcrição *upstream* a regiões codificantes de domínios similares em bactérias e outras archaea sugere que a produção de intraRNAs codificantes é um fenômeno amplamente distribuído em procariotos e pode ser responsável pelo aumento da diversidade do proteoma através da geração de isoformas de proteínas a partir de um único gene. Por fim, a análise de dados de RNA-seq, em conjunto com a busca por assinaturas conhecidas de término de transcrição em archaea, permitiu a identificação da posição final de 58 genes. Os dados obtidos a partir dos experimentos e análises realizados ajudam a traçar um panorama mais completo do transcrito de *H. salinarum* NRC-1 e revelam a presença de novos transcritos que podem ser amplamente distribuídos em procariotos e apresentar importantes papéis funcionais.

Palavras-chave: Bioinformática, Transcrição gênica, Transcrição pervasiva, Archaea, Transcritoma, RNA-seq, Inícios de transcrição.

Abstract

TEN-CATEN, F. **Pervasive transcription in the archaeon *Halobacterium salinarum* NRC-1 and the identification of new transcripts.** 2016, 168 f. Tese (Doutorado) – Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, 2016

The large-scale transcriptome characterization of different organisms revealed a highly complex scenario of gene expression, leading to the identification of numerous transcripts in the genomes of eukaryotes and prokaryotes. This phenomenon has been named pervasive transcription and has been an important source for the search of new RNAs with regulatory functions or involved in the translation of unknown proteins. The abundance of transcriptomic and proteomic data, as well as complete information regarding the genome, allowed the halophilic extremophile *Halobacterium salinarum* to be an ideal model organism for studies of pervasive transcription. This microorganism belongs to the Archaea group, the last one of the three domains of life to be described, which presents shared characteristics with bacteria and eukaryotes. The use of differential RNA-seq (dRNA-seq) approach, which allows the distinction between primary and processed transcripts, allowed the identification of 179 TSSaRNAs, small RNAs associated with the transcription initiation in *H. salinarum*. The application of dRNA-seq in RNA samples collected along the growth curve allowed the identification of 4540 transcription start sites (TSS) in *H. salinarum* NRC-1. Some of these transcription initiation are located upstream to known genes, enabling the identification of TSSs for 1545 genes. 59.2% of these positions are located up to 10 bp away from the translation initiation codon, confirming that most of genes are leaderless. The expression analysis of regions related to antisense TSS under different conditions revealed that most of these regions have a correlated expression profile with genes in the opposite strand, indicating a possible regulatory role. Similarly, analysis of the expression of intergenic TSS allowed the identification of 132 differentially expressed regions that are not related to any other element in *H. salinarum* NRC-1 genome. Integration with proteomic data reveals that some of these regions may be involved in the production of small proteins. The identification of 1365 TSS located within genes suggests that the production of intragenic RNAs (intraRNAs) is a widely distributed phenomenon in *H. salinarum* NRC-1 genome. Northern blot experiments confirmed the production of a transcript corresponding to the final portion of VNG_RS05220 gene and Western blot experiments also revealed that the translation of intraRNAs is responsible for producing small proteins corresponding to individual protein domains with important functional role in specific growth conditions. Analysis of TSS upstream to the coding regions of similar protein domains in bacteria and other archaea suggests that the production of coding intraRNAs is a widely distributed phenomenon in prokaryotes and may be responsible for the increased proteome diversity through the generation of protein isoforms from a unique gene. Finally, the RNA-seq data analysis, combined with a search for known signatures for transcription termination in archaea, allowed the identification of the final position of 58 genes. The present work help to give a more complete picture of *H. salinarum* transcriptional landscape and reveals the presence of new transcripts that can be widely distributed in prokaryotes, with important functional roles.

Keywords: Bioinformatics, Transcription, Pervasive transcription, Archaea, Transcriptome, RNA-seq, Transcription start sites

Sumário

1 Introdução.....	12
1.1 Transcrição pervasiva.....	12
1.2 O domínio <i>Archaea</i>	20
1.3 <i>Halobacterium salinarum</i> NRC-1.....	24
1.4 RNA-seq, identificação de novos RNAs e suas extremidades.....	30
2 Objetivos.....	38
2.1 Geral.....	38
2.2 Específicos.....	38
3 Materiais e Métodos.....	39
3.1 Oligonucleotídeos.....	39
3.2 Linhagem e condição de cultivo.....	39
3.3 Extração dos RNAs.....	40
3.4 Depleção dos RNAs processados.....	41
3.5 Preparação das bibliotecas de dRNA-seq.....	41
3.6 Sequenciamento e processamento dos dados.....	42
3.7 Dados públicos analisados.....	43
3.8 Programas e <i>scripts</i> utilizados nas análises.....	45
3.9 TAG cromossomal.....	45
3.10 Tricina-SDS-PAGE e <i>Western blot</i>	46
3.11 Superexpressão de intraRNAs.....	48
3.12 Avaliação da superexpressão de intraRNAs.....	50
4 Resultados e Discussão.....	51
4.1 Identificação de inícios de transcrição (TSS).....	51
4.1.1 RNAs associados ao início de transcrição.....	51
4.1.2 TSSs durante a curva de crescimento de <i>H. salinarum</i> NRC-1.....	53
4.1.3 Identificação automatizada de TSS em dados de dRNA-seq.....	56
4.1.4 Anotação dos TSSs.....	58
4.1.5 Inícios de transcrição primários (pTSS).....	63
4.1.6 Genes com pTSS associados.....	66
4.1.7 TSS antisense (aTSS).....	70
4.1.8 TSS órfãos (oTSS).....	72
4.1.9 TSS internos (iTSS).....	77
4.1.10 iTSS estruturados.....	80
4.1.11 Análise de enriquecimento funcional.....	84
4.1.12 Fatores de transcrição associados a iTSS.....	84
4.2 Transcritos intragênicos (intraRNAs).....	86
4.2.1 iTSS dão origem a intraRNAs.....	86
4.2.2 intraRNAs diferencialmente expressos.....	89
4.2.3 intraRNAs estão associados a domínios proteicos e são conservados em procariotos.....	91
4.2.4 RNAs intragênicos produzem proteínas alternativas.....	99
4.2.5 Superexpressão de domínios através de intraRNAs.....	110
4.3 Identificação do término de transcrição (TTS).....	114
5 Conclusão.....	119
Referências.....	123
Anexos.....	135

1 Introdução

1.1 Transcrição pervasiva

Com a publicação dos primeiros trabalhos envolvendo o sequenciamento de genomas eucariotos ficou claro que boa parte das sequências identificadas não era responsável pela produção de proteínas e uma grande região dos genomas estudados não tinha uma função bem estabelecida (Venter & et al, 2001; Consortium., 2001; Waterston et al, 2002). O desenvolvimento das tecnologias de quantificação da expressão gênica em larga escala permitiu verificar que mesmo os genomas procariotos, conhecidos por sua compactação e baixo número de regiões intergênicas, também apresentavam regiões responsáveis pela produção de transcritos não codificantes sem uma funcionalidade clara (Selinger *et al.*, 2000). Esses resultados levaram a proposição do fenômeno da transcrição pervasiva, segundo o qual a transcrição não se limita apenas as fronteiras gênicas; a maior parte dos genomas é expresso em alguma condição ou momento do desenvolvimento. Esse fenômeno tem expandido o universo de transcritos identificados e revelado um cenário muito mais complexo da expressão gênica, demonstrando a importância de regiões pouco estudadas do genoma e revelando novas funções para regiões já conhecidas (Kapranov et al, 2007; Wade & Grainger, 2014).

Diferentes trabalhos contestam a existência da transcrição pervasiva, argumentando que os transcritos gerados por esse fenômeno são geralmente associados a baixos níveis de expressão (Johnson *et al.*, 2005), apresentam grande variação na validação experimental ou podem ser associados a genes já conhecidos (van Bakel *et al.*, 2010). Em alguns casos, esses transcritos não canônicos apresentam uma baixa taxa de conservação, o que poderia indicar uma ausência de funcionalidade e reforçar seu caráter de geração espúria (Wang *et al.*, 2004; Brosius, 2005). Os resultados demonstrando a existência da transcrição pervasiva são muitas

vezes interpretados como ruído gerado pelas técnicas utilizadas para quantificar a expressão gênica, como por exemplo a contaminação de amostras por DNA genômico ou mesmo a baixa especificidade de sondas utilizadas nos experimentos de *microarray* (Johnson et al. 2005; Robinson, 2010). Apesar disso, ao longo dos anos esse fenômeno tem sido detectado em diferentes espécies e em alguns casos, como *Escherichia coli*, tem sido confirmado através do uso de diferentes técnicas, como *tiling array*, RNA-seq, técnicas de mapeamento de início de transcrição, detecção de sítios de ligação de fatores de transcrição e transcrição *in vitro*, eliminando assim a possibilidade de associação com artefatos experimentais (Wade & Grainger, 2014). Outro ponto em favor da existência desses transcritos é a identificação de alterações nos seus níveis de expressão e sua relação com fatores regulatórios dos organismos onde são detectados (Singh et al., 2014; Lasa et al., 2011), indicando que esses elementos são alvos de regulação da expressão gênica e devem ser importantes para o funcionamento do organismo. Dessa forma, a discussão em torno da transcrição pervasiva tem se aprofundando no sentido de compreender sua funcionalidade e mecanismos regulatórios (Wade & Grainger, 2014).

O uso de técnicas em larga escala para investigação da expressão gênica possibilitou a identificação da transcrição pervasiva tanto em eucariotos quanto procariotos, tratando-se, dessa forma, de um fenômeno amplamente distribuído. Análises do transcrito humano mostram que 75% do genoma está envolvido com a produção de transcritos (Djebali et al., 2012), já em *Saccharomyces cerevisiae* esse número pode chegar a 85% do genoma (David et al., 2006). Em bactérias, diferentes casos de transcrição pervasiva também foram descritos: em *Escherichia coli* foram detectados em torno de 1000 transcritos antisense a genes codificantes (Dornenburg et al., 2010), já em *Synechocystis* sp. PCC6803 64% dos inícios de transcrição identificados estão envolvidos com a produção de RNAs antisense ou não-codificantes (ncRNAs) (Mitschke et al., 2011). No domínio archaea, a análise combinada de

sítios de ligação a fatores de transcrição e dados de expressão gênica em larga escala permitiu a identificação da expressão de subunidades de operons e RNAs intergênicos em *Halobacterium sp.* NRC-1 (Koide *et al.*, 2009). Além disso, resultados do nosso grupo de pesquisa apontam que 85% das sequências codificantes (CDS – *coding sequences*) desse micro-organismo apresentam transcrição antisense em, pelo menos, metade da sua extensão (Vencio *et al.*, em preparação).

A transcrição pervasiva atua como fonte de inúmeras espécies de RNAs, e apesar de diferentes transcritos serem produzidos ao longo de toda a extensão dos genomas analisados, muitos deles parecem se agrupar em torno de regiões codificantes (Kampa *et al.*, 2004; Fejes-Toth *et al.*, 2009; Cheng *et al.*, 2005). Tal observação levou Kapranov *et al.*, 2007 a propor um modelo de organização genômica para eucariotos onde as regiões codificantes atuam como geradoras de novas espécies de RNAs nas fitas sense e antisense, intercalado por pequenos RNAs nas regiões intergênicas (Figura 1). Dessa forma, podemos encontrar diversos tipos de transcritos: pequenos ou longos RNAs associados às regiões promotoras (PASR/PALR) e também *upstream* a essa região (PROMPT) (Kapranov *et al.*, 2007; Preker *et al.*, 2008); pequenos RNAs nucleolares (snoRNAs) (Bachellerie *et al.*, 2002); micro RNAs (miRNAs) (Landgraf *et al.*, 2007); pequenos RNAs associados ao início (tiRNAs) ou término de transcrição (TASR) (Taft *et al.*, 2009; Seila *et al.*, 2008; Kapranov *et al.*, 2007); RNAs derivados da região 3' UTR (uaRNA) (Mercer *et al.*, 2011) e RNAs não codificantes longos (lncRNAs) antisense a regiões codificantes, em regiões intergênicas, dentro de íntrons e em regiões responsáveis por aumentar a afinidade da maquinaria de transcrição pelo promotor (*enhancer*) (Carninci *et al.*, 2006; Engström *et al.*, 2006; Nakaya *et al.*, 2007; Kim *et al.*, 2010; Ponting *et al.*, 2009).

Figura 1: Transcrição pervasiva em eucariotos. Modelo da transcrição pervasiva baseado no proposta de Kapranov et al., 2007 onde as regiões codificantes atuam como centros geradores de novas espécies de RNAs. O transcrito primário é representado pelo 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 codificantes 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.

Cenário semelhante tem sido encontrado em genomas procariotos, indicando uma grande quantidade de informações sobrepostas no genoma e adicionando mais uma camada de complexidade a genomas antes considerados simples (Güell et al., 2011; Stazic & Voß, 2016) (Figura 2). Além de pequenos RNAs em regiões intergênicas (Mitschke et al., 2011; Bilusic et al., 2014; Koide et al., 2009) e snoRNAs antisense a tRNAs e rRNAs em Archaea (Omer, 2000), podemos encontrar também RNAs associados ao início de transcrição (TSSaRNAs) (Yus et al., 2012; Zaramela et al., 2014), RNAs antisense a genes (asRNA) e transcritos produzidos a partir de regiões 5' e 3' UTR (Kawano, 2005; Chao et al., 2012; Bilusic et al., 2014). Em bactérias, foram identificados também longos RNAs (lasRNA) que podem se estender por mais de um gene e apresentam uma dupla função: atuar como RNA

regulatório de genes antisense e ainda ter a informação codificante na fita sense (Wurtzel *et al.*, 2012; Sesto *et al.*, 2012). Mais recentemente, diferentes trabalhos têm identificado RNAs transcritos a partir de posições dentro de ORFs, em sobreposição ao mRNA. Pequenos RNAs sense (seRNAs) foram identificados em *Synechocystis* sp. PCC6803 e *Neisseria gonorrhoeae* (Mitschke *et al.*, 2011; Wachter & Hill, 2015) e podem apresentar função de mimetizar o mRNA, interagindo com asRNA produzidos na fita oposta (Mitschke *et al.*, 2011). Transcritos similares foram identificados também próximos a extremidade 3' de transposases em diferentes procariotos (Gomes-Filho *et al.*, 2015). Em *E. coli* 25 RNAs intragênicos (intraRNAs) foram identificados através do sequenciamento de transcritos interagindo com a chaperona de RNA *Hfq* (Bilusic *et al.*, 2014).

Figura 2: Transcrição pervasiva em procariotos. Similar ao que ocorre em eucariotos existem diversos transcritos sobrepostos a genes (setas azuis) no genoma de procariotos. A região em azul claro nos genes representa as porções 5' e 3' UTRs. TSSaRNA, transcritos associados ao início de transcrição; intraRNA, RNAs intragênicos; asRNA, transcritos antisense; seRNA, transcritos sense; lasRNA – excludon, RNAs antisense longos, os quais apresentam uma porção codificante e uma porção não codificante capaz de regular o mRNA na fita oposta. A direção das setas indica a orientação dos genes e transcritos.

A análise da transcrição pervasiva também tem levado a identificação de inúmero transcritos codificantes, principalmente em regiões intergênicas. Devido a limitações nos procedimentos de anotação automatizada de genomas e nos experimentos de espectrometria

de massa, proteínas menores de 100 aminoácidos foram durante muito tempo desconsideradas das análises, visando diminuir o número de falsos positivos identificados (Su *et al.*, 2013). Recentemente, protocolos específicos têm permitido a identificação dessas pequenas proteínas em diferentes genomas, e revelado importantes papéis funcionais. Em genomas mais complexos, como de *Drosophila*, estima-se que existam mais de 4000 pequenas ORFs (smORFs) (Ladoukakis *et al.*, 2011), enquanto em *Saccharomyces cerevisiae*, análises de expressão gênica e busca por homologias revelaram a existência 299 smORFs, das quais 247 foram validadas experimentalmente (Kastenmayer *et al.*, 2006). Na bactéria *Mycoplasma pneumoniae* 67 smORFs foram identificados, sendo que aproximadamente metade se mostrou essencial para o desenvolvimento do organismo, demonstrando também a importância desses elementos em genomas reduzidos (Lluch-Senar *et al.*, 2015). Análises de sequenciamento de sítios de interação de ribossomos (*ribosome profiling*) tem levado a identificação de regiões ativamente traduzidas a partir de pequenas ORFs também dentro de RNAs não codificantes longos (lncRNAs) (Nam *et al.*, 2016). Além disso, sítios de início de tradução tem sido identificados em pequenos RNAs não codificantes (Andrews & Rothnagel, 2014; Lee *et al.*, 2012). Assim, muitos transcritos, antes considerados somente não-codificantes podem dar origem a proteínas, existindo ainda exemplos de transcritos bifuncionais, fazendo com que a distinção entre transcritos codificantes e não-codificantes seja cada vez mais complexa e condicional.

Até o momento acredita-se que os mecanismos mais importantes para geração da transcrição pervasiva em procariotos estejam envolvidos com alterações em processos usuais de transcrição, como a localização de promotores ativos em regiões não canônicas e o não reconhecimento de sítios de término de transcrição (TTS) pela RNA polimerase, estendendo os transcritos muito além da porção gênica (Wade & Grainger, 2014). Em *E. coli*, a maioria dos sítios de interação do fator σ^{54} está localizado dentro de genes (Bonocora *et al.*, 2015);

essa observação tem se repetido para outros fatores de transcrição em diferentes organismos (Zhu *et al.*, 2006; Tanay, 2006; Shimada *et al.*, 2008; Koide *et al.*, 2009), sugerindo que a existência de regiões promotoras internamente a regiões gênicas pode ser um fenômeno amplamente distribuído. A existência desses sítios de interação a fatores de transcrição poderia estar relacionada com a pressão evolutiva presente nas regiões codificantes, impedindo grandes variações nas sequências e atuando na manutenção das regiões promotoras intragênicas (Wade & Grainger, 2014).

Em relação ao controle da transcrição pervasiva em bactérias, são conhecidos mecanismos que atuam no início de transcrição, alongação, terminação e degradação, sendo que tais mecanismos devem atuar em cooperação para controlar a produção de transcritos espúrios. Quanto ao controle de início de transcrição, o mecanismo melhor caracterizado é a regulação via H-NS, uma proteína similar as histonas de eucariotos e responsável pela manutenção da estrutura do nucleóide no interior da célula procariótica. Em *E. coli* essa proteína pode impedir o início de transcrição a partir de posições internas do operon *ehXCABD* na linhagem O157, e também a nível genômico na linhagem K12 (Singh *et al.*, 2014). Se o mecanismo de controle de início de transcrição via H-NS é evadido e a fase de alongação é iniciada, o principal mecanismo de controle está relacionado com os fatores de terminação Rho e NusG. O mecanismo mais aceito é de que a proteína NusG é capaz de interagir com a RNA polimerase, tornando-a mais suscetível a desestabilização pelo fator Rho (Peters *et al.*, 2012). Além disso, os sítios de ligação da proteína H-NS estão colocalizados com as regiões de terminação pelo fator Rho, indicando que a interação da proteína H-NS pode dificultar o progresso da RNA polimerase e permitir a desestabilização do complexo transcricional pelo fator Rho (Peters *et al.*, 2012). A etapa final de controle da transcrição pervasiva está relacionada com a atuação de RNases na degradação de transcritos não canônicos. Em *Bacillus subtilis* a depleção da RNase Y, RNase J1 e RNase III promove um

aumento na abundância de ncRNAs antisense (Durand *et al.*, 2012) sendo que o mesmo ocorre para a depleção da RNase III em *Staphylococcus aureus* (Lasa *et al.*, 2011). Em eucariotos, enzimas com os domínios da RNase III, como Dicer e Drosha, tem papel fundamental na maturação de microRNAs, ressaltando o papel desses domínios no processamento de transcritos não codificantes (Lee *et al.*, 2003).

Ainda não estão muito claras as possíveis funções associadas aos RNAs produzidos através da transcrição pervasiva, mas de forma geral acredita-se que tais transcritos devam atuar principalmente em vias regulatórias da expressão gênica através da interação com mRNAs, levando a degradação dos mesmos ou impedindo o acesso dos ribossomos (Wade & Grainger, 2014). A existência de diferentes mecanismos de controle sugere que a transcrição pervasiva deve apresentar um custo energético para a célula, reduzindo sua eficiência. Apesar disso, existe uma grande diversidade de transcritos não canônicos no interior da célula. Uma das hipóteses para explicar a manutenção da geração dessa grande quantidade de transcritos seria a pressão seletiva em regiões gênicas, a qual atua na estabilização de sequência funcionais e poderia levar a fixação de eventuais promotores dentro de genes. Além disso, a produção intensa de diferentes transcritos poderia fornecer um conjunto de RNAs para uma resposta rápida frente a novas condições de crescimento, uma vez que RNAs tem um menor custo energético envolvido e apresentam uma maior capacidade de interação com diferentes moléculas, aumentando a plasticidade dos genomas (Heinen *et al.*, 2009; Wade & Grainger, 2014).

A transcrição pervasiva parece ainda ter um efeito importante na evolução de genomas. Análises filogenéticas de *E. coli* demonstraram que regiões transcricionalmente inativas, muitas vezes colocalizadas com sítios de ligação a proteína H-NS, acumulam mutações mais facilmente que regiões altamente transcritas (Martincorena *et al.*, 2012). Esse fenômeno pode ser explicado pela capacidade da RNA polimerase de detectar regiões do

DNA danificados e recrutar uma maquinaria de reparo associada a transcrição (Savery, 2007). Dessa forma, a transcrição pervasiva poderia aumentar a taxa de transcrição de uma determinada região e diminuir a taxa de mutações.

1.2 O domínio *Archaea*

Enquanto analisavam as características da subunidade menor do RNA ribossomal de organismos procariotos, Carl Woese e colaboradores identificaram uma grande diferença na composição dessas moléculas em algumas espécies do grupo. Análises mais detalhadas demonstraram a existência de um grupo monofilético distinto dentro do domínio Bactéria, levando-os a sugerir a existência de um novo domínio da vida, o domínio Archaeobacteria, posteriormente renomeado para Archaea (Woese et al., 1990; Woese & Fox, 1977) (Figura 3). Análises a nível molecular indicam que as Archaea apresentam um mosaico de características compartilhadas entre os demais domínios (Tabela 1). Exemplo disso é a maquinaria de processamento de informação (replicação, transcrição, tradução), que apresenta proteínas com alto grau de similaridade com Eucariotos; já a organização genômica compacta, com poucas regiões intergênicas e muitos genes dispostos em operons é uma propriedade compartilhada entre procariotos em geral (Grabowski & Kelman, 2003; Rivera et al, 1998; Jain et al, 1999).

Além das similaridades com os demais domínios, as Archaea apresentam também características únicas, como por exemplo a composição da bicamada lipídica da membrana, caracterizada pela presença de uma cadeia lateral de isoprenóides unidos por uma ligação éter a um grupo glicerol-1-fosfato. Em bactérias e eucariotos a cadeia lateral de hidrocarbonetos derivados de ácidos graxos é unida por uma ligação éster a um grupo glicerol-3-fosfato (Jain *et al.*, 2014) (Tabela 1). Essa diferença tem importantes implicações na estabilidade da membrana e tem sido relacionada a sua capacidade adaptativa a diferentes ambientes (Koga,

2012). Além disso, esse é o único domínio que abriga organismos capazes de empregar a metanogênese em seu metabolismo.

Tabela 1: Características apresentadas por Bactérias, Archaea e Eucariotos. Adaptado de Cavicchioli, 2011

Característica	Bactéria	Archaea	Eucarioto
Ligação dos carbonos de lipídios	Éster	Éter	Éster
Fosfatos dos lipídios de 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

Muitos dos representantes desse grupo habitam ambientes inóspitos que remetem às condições primitivas do planeta Terra (éon Arqueano), levando a proposição de que esse domínio pode estar ligado a origem da vida e também ao desenvolvimento de formas de vida mais derivadas, como os Eucariotos (Yutin et al., 2008; Gribaldo et al., 2010). De acordo com os ambientes onde são encontrados, esses microorganismos recebem diferentes classificações. Archaea metanogênicas são organismos anaeróbicos que habitam ambientes com baixa disponibilidade de oxigênio, como pântanos e o interior do intestino de alguns animais. Esses organismos obtêm energia através da redução do dióxido de carbono, gerando metano durante o processo. Halófilos são organismos capazes de viver em ambientes de extrema salinidade. Já os alcalófilos e acidófilos são capazes de suportar ambientes nas faixas mais extremas de pH, sendo que geralmente organismos acidófilos são também termófilos, habitando ambientes com temperaturas por volta dos 100° C, e muitos organismos alcalófilos são halófilos. Existem também as archaea psicrofílicas, capazes de habitar ambientes de baixa temperatura, como o interior de lagos congelados. Além de habitar nichos incomuns, muitas archaea têm

sido encontradas em ambientes com condições mais amenas de desenvolvimento, como por exemplo em águas oceânicas, possuindo um papel importante nos ciclos biogeoquímicos do planeta (Cavicchioli, 2011).

O domínio Archaea apresenta dois filos principais: Euryarchaeota, o mais diverso e composto por archaea metanogênicas, halófilas, termófilas e psicrofilas, e o filo Crenarchaeota, o qual apresenta principalmente organismos hipertermófilos, como *Sulfolobus solfataricus*, que apresenta condições ideais de crescimento a 80° C (Figura 3). Dentre as características que diferenciam esses dois filos podemos destacar a presença de histonas com alta similaridade a homólogos eucarióticos em Euryarchaeota, enquanto Crenarchaeota utilizam um conjunto diferenciado de proteínas de ligação ao DNA (White & Bell, 2002). Os demais filos presentes no domínio Archaea são os filos Nanoarchaeota, Korarchaeota e Thaumarchaeota, os quais apresentam poucos representantes ou espécies não cultiváveis detectadas em amostras de DNA ambiental. A posição filogenética desses filos ainda é incerta (Allers & Mevarech, 2005).

A capacidade das Archaea de habitar ambientes inóspitos têm feito delas organismos de grande interesse biotecnológico. Exemplo disso é a enzima *Pfu*, uma DNA Polimerase derivada da archaea *Pyrococcus furiosus* capaz de suportar altas temperaturas e amplamente empregada em reações de PCR. Outro grupo de enzimas derivadas de archaea e com grande potencial biotecnológico são as celulasas de haloarchaea empregadas na produção de bicomcombustíveis. Como os ambientes de conversão de biomassa são inibitórias para a maioria das celulasas comumente utilizadas, as enzimas de haloarchaea, que são bem-adaptadas a ambiente de alta concentração iônica, podem apresentar uma alternativa viável (Zhang *et al.*, 2011). Os organismos metanogênicos, por sua vez, apresentam um importante papel na geração de biocomcombustíveis. Além disso, esse grupo de Archaea está relacionado com aumento do aquecimento global uma vez que é um dos principais constituintes da microbiota

intestinal de espécies criadas intensivamente para alimentação humana, como bovinos, ocasionando um aumento na liberação de metano, um importante gás do efeito estufa (Patra, 2012). Por habitarem ambientes que remetem a Terra pré-biótica, as Archaea também tem sido alvo de estudo da Astrobiologia, a qual busca avaliar sua capacidade de colonizar outros planetas e também auxiliar na explicação da origem da vida em nosso planeta (Cavicchioli, 2002; Koonin & Martin, 2005).

Figura 3: Árvore filogenética baseada na comparação do RNA ribossomal. A proposição de Carl Woese de analisar a relação entre as espécies através da subunidade menor do rRNA permitiu a distinção dos procariotos em dois grupos, Bactérias e as Archaea. Em destaque estão os quatro grupos do domínio Archaea, sendo que o mais diverso deles é o grupo Euryarchaeota, abarcando organismos metanogênicos e halófilos, além de alguns psicrófilos e termófilos. O grupo Crenarchaeota apresenta os organismos hipertermófilos, com as espécies capazes de sobreviver as condições mais extremas de temperatura. Os grupos Nanoarchaeota e Korarchaeota apresentam poucas espécies e o posicionamento dos seus filis na árvore filogenética ainda é incerta (linhas tracejadas). Adaptado de Allers & Mevarech, 2005.

1.3 *Halobacterium salinarum* NRC-1

Dentre as Archaea uma espécie que se estabeleceu como organismo modelo é a halófila obrigatória *Halobacterium salinarum* (Figura 4). Essa archaea extremófila é de fácil cultivo em laboratório, com condições ideais de crescimento a 37° C e 4-5 M de NaCl, e apresenta um grande potencial para estudos de biologia molecular, com uma ampla gama de protocolos disponíveis (DasSarma *et al.*, 1995). *H. salinarum* foi primeiramente identificada no início do século XX, mas os estudos com esse organismo ganharam impulso devido principalmente à descoberta da bacteriorodopsina. Essa proteína de membrana é homologa as rodopsinas, proteínas envolvidas com a visão, e é capaz de captar a energia luminosa do ambiente para gerar um gradiente de prótons na membrana. A entrada desses prótons através de ATP sintases localizadas na membrana é responsável pela produção de energia em baixas concentrações de oxigênio. Essa archaea é normalmente encontrada em salinas, as quais assumem colorações avermelhadas devido à presença da bacteriorodopsina e pigmentos carotenóides presentes nesse micro-organismo (Figura 4A). Apresenta formato de bastonete, com flagelos e uma camada de glicoproteínas que circundam a membrana, denominada *surface layer* (S-layer). Além disso é capaz de produzir vesículas de gás responsáveis pela sua flutuação até a superfície do ambiente aquático, permitindo o acesso a melhores condições de oxigenação e luminosidade (Figura 4B). É um organismo anaeróbico facultativo e em laboratório é cultivado em meio complexo (peptona) apresentando um tempo de geração entre 8 e 12 horas, sendo capaz de utilizar a fermentação de certos aminoácidos, como arginina e aspartato, para obtenção de energia. *H. salinarum* é um halófilo obrigatório, apresentando concentrações internas de íons potássio em torno de 4 M, apesar de a concentração externa girar em torno de 0,1 M. Essa grande quantidade interna de potássio é responsável pelo equilíbrio osmótico frente a alta concentração externa de sódio (Christian & Waltho, 1962).

Figura 4: *Haloarchaea Halobacterium salinarum* NRC-1. A) Vista aérea de lagos hipersalinos com desenvolvimento de archaea halofílicas. A coloração avermelhada se deve a proteína de membrana bacteriorodopsinae pigmentos carotenóides envolvidos na fototrofia. B) Microscopia de corte longitudinal de *H. salinarum* NRC-1. Os pontos brancos são vesículas de gás produzidas com objetivo de promover a flutuação do organismo até a porção superior do meio aquoso, favorecendo a captação de oxigênio e luz.

A linhagem selvagem de *H. salinarum*, NRC-1, possui o genoma completamente sequenciado, apresentando um cromossomo principal circular de aproximadamente 2,01 Mb e dois plasmídeos de 0,37 e 0,19 Mb, denominadas pNRC200 e pNRC100, respectivamente (Ng *et al.*, 2000). Os plasmídeos são por diversas vezes denominados minicromossomos pois possuem genes de grande importância, como fatores de transcrição e genes de vesículas de gás. Uma característica importante do seu genoma é seu alto conteúdo GC, que pode chegar até 68% no cromossomo principal. No total já foram anotados 2.629 genes, sendo 2042 no cromossomo, 318 no plasmídeo pNRC200 e 180 no plasmídeo pNRC100. Destes, 676 (25,7%) são classificados como hipotéticos. Em relação ao seu proteoma, *Halobacterium* apresenta proteínas bem-adaptadas ao ambiente de alta salinidade, apresentando em geral, proteínas ácidas, com um pI médio de 5,1 (Ng *et al.*, 2000). Além da linhagem NRC-1, a linhagem R1 de *H. salinarum* é também bastante estudada, sendo que ambas compartilham alta similaridade de genoma e provavelmente se originaram a partir de um mesmo isolado na natureza (Pfeiffer *et al.*, 2008). O cromossomo principal das duas linhagens apresenta apenas 12 alterações, como mutações pontuais e alterações na fase de leitura em alguns genes. A

principal diferença se deve a uma inserção de 10 kb no cromossomo da linhagem NRC-1, a qual é flanqueada por uma duplicação de 8bp, indicando que essa inserção se originou a partir de um evento de transposição. A linhagem R1 apresenta 4 megapasmídeos, sendo que apenas 4,5 kb dos plasmídeos da linhagem NRC-1 não estão presentes nos plasmídeos da linhagem R1, enquanto que os plasmídeos da linhagem R1 apresentam uma porção adicional de 210 kb. Assim como em NRC-1, os plasmídeos da linhagem R1 apresentam grandes duplicações (Pfeiffer et al, 2008). A anotação mais recente e curada manualmente da linhagem R1 (Pfeiffer, 2015) tem sido uma importante ferramenta para os estudos em escala genômica da linhagem NRC-1 realizados no nosso grupo de pesquisa.

Halobacterium sp., assim como as demais archaea, apresenta uma maquinaria de início de transcrição similar ao encontrado em Eucariotos, com a presença de um sítio de interação com o fator de transcrição B (BRE) nas posições -32 a -38 em relação a posição de início de transcrição, e um TATA box iniciando nas posições -26/-27, além disso apresenta um elemento de iniciação em torno da posição de início de transcrição (Soppa, 1999). Outra característica importante em haloarchaea é a preponderância de transcritos sem regiões 5' UTRs, os quais são denominados *leaderless* (Brenneis et al., 2007). Na linhagem R1 de *H. salinarum* 70% dos transcritos avaliados experimentalmente são *leaderless* (Brenneis et al, 2007), enquanto na haloarchaea *Haloferax volcanii*, informações em larga escala determinaram que 72% dos transcritos não apresentam região 5' UTR (Babski et al., 2016). Apesar da proporção de transcritos *leaderless* ser similar nas duas espécies, o tamanho médio das regiões 5' UTRs em *H. volcanii* é de 19 nucleotídeos, enquanto em *H. salinarum* R1 essa região apresenta em média 33 nucleotídeos (Brenneis et al., 2007).

Uma característica muito importante de *H. salinarum* está relacionada com o comportamento bifásico de sua curva de crescimento, a qual apresenta duas fases de aumento na densidade óptica (DO) intercaladas por um período com poucas alterações (Figura 5). A

primeira fase está relacionada com o crescimento exponencial no meio de cultura, enquanto a segunda fase está relacionada com a morte das células e liberação das vesículas de gás no meio de cultura, demonstrada pela diminuição de unidades formadoras de colônia nesse período (Figura 5). A análise do perfil de expressão gênica ao longo da curva de crescimento de *H. salinarum* NRC-1 permitiu a caracterização da resposta transcricional nos diferentes estágios do crescimento (Lange *et al.*, 2007; Facciotti *et al.*, 2010). Durante a primeira fase do crescimento, os principais transcritos encontrados estão relacionados à maquinaria de transcrição, tradução, fosforilação oxidativa e proteínas envolvidas com o transporte de ferro. Por outro lado, durante a fase estacionária os genes expressos estão relacionados principalmente a transportadores ABC e genes envolvidos em vias alternativas de obtenção de energia, como fototrofia, fermentação de arginina e respiração utilizando os aceptores de elétrons DMSO/TMAO, além de vesículas de gás e proteínas de choque térmico (Facciotti *et al.*, 2010).

Figura 5: Curva de crescimento e viabilidade celular de *H. salinarum* NRC-1. As duas fases de aumento na densidade óptica durante o crescimento são intercaladas por uma fase estacionária com poucas alterações na DO (círculo tracejado). A diminuição no número de unidades formadoras de colônia concomitantemente ao segundo aumento de DO indica que esse comportamento pode estar relacionada com a morte das células e liberação de vesículas de gás no meio de cultura, alterando a dispersão do feixe luminoso utilizado para aferimento da concentração de células. Adaptado de Facciotti *et al.*, 2010.

Por se de fácil cultivo e manipulação em laboratório, *H. salinarum* se estabeleceu como um organismo modelo dentro do domínio Archaea. Dessa forma, existe um grande número de trabalhos que buscam avaliar esse microorganismo nas mais variadas condições, usando perturbações genéticas e ambientais e análises em escala genômica para compreender sua fisiologia, principalmente da linhagem NRC-1. Exemplo disso são as análises de crescimento e sobrevivência frente a variações em diferentes elementos, como oxigênio, luminosidade, metais de transição (manganês, ferro, cobalto, níquel, cobre e zinco) e também radiações UV e gama (Baliga et al., 2004; Kottemann et al., 2005; Kaur et al., 2006). Além disso existe também uma grande quantidade de informações sobre suas respostas transcricionais frente a deleção de fatores de transcrição (Turkarslan *et al.*, 2014). Outros dados disponíveis para esse microorganismo estão relacionados a identificação de sítios de ligação fatores de transcrição (Facciotti *et al.*, 2007) e dados de proteômica obtidos em uma grande variedade de experimentos (Van *et al.*, 2008).

Essa grande quantidade de informações torna esse microorganismo atrativo para os estudos de Biologia Sistêmica que buscam integrar as informações obtidas a partir de diferentes plataformas e condições para avaliar a fisiologia do organismo de forma global (Ge *et al.*, 2003). A análise de sítios de interação de fatores de transcrição em conjunto com o sinal de expressão gênica ao longo da curva de crescimento, levou a identificação de promotores condicionais dentro de operons de *H. salinarum* NRC-1. Esses promotores são responsáveis pela expressão de subunidades operônicas transcritas de forma modular em condições específicas do crescimento (Koide *et al.*, 2009). A partir da integração de diferentes dados Bonneau et al, 2007 foram capazes de identificar grupos de genes que respondiam de forma similar de acordo com a condição analisada, possibilitando classificar 1900 genes em grupos co-regulados e elaborar modelos que descrevessem a dinâmica de resposta desses grupos de genes. Com a construção desse modelo foi possível prever as modificações transcricionais de

aproximadamente 80% dos genes em novas condições. Uma consideração importante do trabalho diz respeito as situações onde não foi possível prever o comportamento da resposta transcricional através do modelo, indicando que esses casos devem estar relacionados com a carência de informações a respeito de alguns elementos, como modificações epigenéticas, modificações pós-traducionais de proteínas, respostas baseadas em metabólitos e também informações sobre ncRNAs, ainda escassas nesse organismo (Bonneau *et al.*, 2007).

Mais recentemente, trabalhos do nosso grupo de pesquisa tem avançado no sentido de identificar novas espécies de RNAs presentes em *H. salinarum* NRC-1 e conservados em outros procariotos. Análises de dados de expressão gênica permitiram identificar um perfil diferenciado de expressão próximo a extremidade 3' de transposases da família IS1341 associado a um grupo de transcritos denominados de sotRNAs (*sense overlap* RNAs). A superexpressão de um desses elementos, e de um ncRNA intergênico com o mesmo motivo estrutural, promoveu um aumento no crescimento de *H. salinarum* NRC-1 demonstrando seu papel funcional. Esses transcritos foram identificados também nas archaea *S. solfataricus* (Zago *et al.*, 2005), *T. kodakarensis* (Jäger *et al.*, 2014), *Methanopyrus kandleri* AV19, *Sulfolobus acidocaldarius* DSM639 e *Pyrococcus furiosus* DSM3638, além das bactérias *Helicobacter pylori* 26695 e *Escherichia coli* K12, sugerindo sua ampla distribuição entre os procariotos (Gomes-Filho *et al.*, 2015). De forma similar, análise de dados de expressão gênica em 10 diferentes transcritomas de archaea demonstram um perfil diferenciado de expressão na extremidade 5' de regiões codificantes, atribuídos a RNAs associados ao início de transcrição (TSS *associated* RNAs – TSSaRNAs). Esses transcritos foram experimentalmente validados em *H. salinarum* NRC-1, e sua produção está possivelmente associada a pausas da RNA polimerase no processo de alongação da transcrição (Zaramela *et al.*, 2014).

Além de transcritos associados a regiões codificantes, a análise do perfil de expressão gênica de *H. salinarum* NRC-1 ao longo da curva de crescimento permitiu a identificação de 61 RNAs intergênicos diferencialmente expressos. Esses pequenos RNAs não apresentaram uma região codificante definida e não possuem correspondência com os dados disponíveis de proteômica, sendo, dessa forma, classificados como ncRNAs (Koide *et al.*, 2009). Pequenos RNAs já foram identificados em diferentes espécies de archaea, como em *H. volcanii*, onde foram identificados 145 pequenos RNAs intergênicos e 45 pequenos RNAs antisense a regiões codificantes (Heyer *et al.*, 2012) e *Pyrococcus abyssi*, outra espécie do filo Euryarchaeota, onde foram identificados 107 sRNAs intergênicos e 215 sRNAs antisense a regiões codificadoras (Toffano-Nioche *et al.*, 2013). No filo Crenarchaeota, *S. solfataricus* apresenta um total de 43 sRNAs intergênicos e 185 sRNAs antisense (Wurtzel *et al.*, 2010).

Analisadas em conjunto, as informações sobre *Halobacterium salinarum* apontam para a existência de importantes espécies de RNAs que podem contribuir para o entendimento da regulação da expressão gênica nesse organismo. Além disso, a existência de uma grande base de dados, obtidos em diferentes plataformas e em diferentes condições, torna esse organismo muito atrativo para o estudo da transcrição pervasiva e seus efeitos.

1.4 RNA-seq, identificação de novos RNAs e suas extremidades

As tecnologias de sequenciamento em larga escala se tornaram uma das principais ferramentas para identificação e caracterização de genomas e a sua expressão no nível transcricional. Tais tecnologias visam, principalmente, o aumento da capacidade de aquisição de dados e a diminuição nos custos envolvidos nessa aquisição. Com isso, tornou-se mais

difundido o desenvolvimento de pesquisas capazes de extrair informações globais a respeito do funcionamento dos sistemas biológicos.

Dentre essas novas tecnologias, o sequenciamento de RNAs (RNA-seq) tem ganhado grande destaque dentro do campo de estudo da transcritômica (Mortazavi *et al.*, 2008). A técnica de RNA-seq baseia-se nas metodologias de sequenciamento de ácidos nucleicos desenvolvidas nas últimas décadas e que representaram uma quebra do paradigma sobre sequenciamento, permitindo a geração de um volume massivo de dados. O maior responsável pelo aumento de performance na nova geração de sequenciadores é a miniaturização das reações e a alta capacidade de paralelização do sequenciamento (Metzker, 2010).

Por apresentar um maior nível de resolução e também por possuir uma maior sensibilidade na detecção dos níveis de expressão, o RNA-seq permite o discernimento entre pequenas variações dos níveis de expressão de regiões muito próximas, até mesmos adjacentes (Wang *et al.*, 2009). Além disso, a técnica possibilita a detecção mais acurada de genes expressos em níveis muito baixos ou muito elevados (maior *dynamic range*) (Mortazavi *et al.*, 2008), e é amplamente empregada em estudos que visam comparar os níveis de expressão gênica global entre organismos selvagens e mutantes (Roschetto *et al.*, 2012), organismos expostos a diferentes condições ambientais (Villar *et al.*, 2011) ou também tecidos afetados por determinadas doenças (Sinicropi *et al.*, 2012).

Além de fornecer resultados a respeito dos níveis de expressão gênica, os dados de RNA-seq trazem outra informação relevante para o entendimento do transcritoma: a estrutura do sinal de expressão gênica. Uma vez que as sequências (*reads*) tenham sido geradas e mapeadas ao genoma referência, podemos contabilizar a cobertura em cada posição do genoma e gerar um sinal de expressão para todas as posições mapeadas. Com essa abordagem, o RNA-seq vem auxiliando na detecção da expressão em regiões onde não existem genes anotados. Por meio da análise dessas regiões, os pesquisadores têm conseguido

determinar novos elementos com importância regulatória como ncRNAs e regiões UTRs, e regiões codificantes ainda não anotadas (Alexander *et al.*, 2010).

Contudo, analisar a composição do sinal de expressão gênica gerado envolve uma série de dificuldades. Como os RNAs sequenciados constituem uma amostra de todos os transcritos presentes na célula no momento da extração, o sinal de expressão gênica é composto por RNAs iniciando o processo de transcrição, transcritos que estavam sendo processados para desempenhar corretamente sua função e também aqueles envolvidos em vias de degradação (Miller *et al.*, 2011). Dessa forma, o RNA-seq pode ser comparado com uma espécie de fotografia estática de um processo altamente dinâmico, sendo que a miscelânea de espécies de RNA em diversos estágios do seu ciclo de vida pode introduzir grande variabilidade na composição do sinal analisado.

Além das variações produzidas pelas etapas naturais do processo de transcrição, diversas fontes de ruído já foram identificados na técnica de RNA-seq, sendo que podemos dividi-las em dois tipos principais: (i) o ruído gerado durante a etapa de preparação das bibliotecas e sequenciamento e (ii) o ruído gerado na etapa de análise dos dados. Durante a etapa de preparação das bibliotecas e sequenciamento, uma das características que introduzem viés nos dados é o conteúdo GC das sequências. Alguns trabalhos demonstram um menor nível de sinal de sequenciamento em regiões ricas em conteúdo GC (Sendler *et al.*, 2011). Outro ponto importante é o viés introduzido através da ligação de adaptadores aos fragmentos que serão sequenciados, já que as enzimas utilizadas para ligação dos adaptadores ao RNA podem apresentar sítios preferenciais de ligação (Jayaprakash *et al.*, 2011). Além disso, as etapas de transcrição reversa dos fragmentos de RNA e amplificação através de PCR podem atuar na geração de ruído nas bibliotecas de RNA-seq (Mamanova & Turner, 2011).

Dessa forma, a soma de todos os fatores técnicos e biológicos, leva a constituição de um sinal de transcrição altamente variado em dados de RNA-seq, e o novo desafio imposto é

conseguir analisar esse sinal de forma automatizada e global. Com isso se faz necessário o desenvolvimento de metodologias que consigam lidar com essa variabilidade para processar e segmentar o sinal e assim obter informações importantes como, por exemplo, as fronteiras das regiões codificantes já identificadas, regiões onde existe sobreposição de sinais de transcritos diferentes, e também novos elementos regulatórios como ncRNAs (Wang et al., 2009).

Alguns trabalhos têm sido publicados na tentativa de focar os esforços de sequenciamento em grupos específicos de RNAs, o que pode levar a uma diminuição da complexidade do sinal e facilitar as etapas de processamento. Um exemplo é o estudo publicado por Sharma et al., 2010 onde os pesquisadores foram capazes de mapear globalmente os inícios de transcrição de *Helicobacter pylori*, um importante patógeno humano, através da comparação de uma biblioteca enriquecida com RNAs primários com uma biblioteca controle. O enriquecimento da amostra com RNAs primários é realizado com a enzima *Terminator 5' Phosphate-Dependent Exonuclease* (TEX). Essa enzima é capaz de degradar todos os RNAs monofosfatados na extremidade 5', ou seja, RNAs que passaram por alguma etapa de processamento nessa região. Sequenciando uma biblioteca de RNAs primários e uma biblioteca controle, os pesquisadores puderam determinar os inícios de transcrição identificando as posições que apresentavam um maior número de contagens na biblioteca enriquecida pra RNAs primários quando comparado a biblioteca controle (Figura 6).

Figura 6: 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 posições de início de transcrição (TSS). Adaptado de Sharma et al , 2010.

A identificação em larga escala de inícios de transcrição é uma importante ferramenta para obter uma melhor caracterização da transcrição pervasiva. A análise de dados provenientes do dRNA-seq tem permitido a identificação de TSSs por toda a extensão dos genomas procariotos analisados, fornecendo informações sobre inícios de transcrição de genes já conhecidos, identificação de TSS dentro de regiões gênicas, antisense a regiões codificantes e também em regiões intergênicas sem anotação conhecidas (Sharma & Vogel, 2014). Através dessas informações, diferentes elementos podem ser melhor caracterizados nos organismos estudados. Os inícios de transcrição localizados *upstream* a regiões anotadas, normalmente denominados de TSSs primários (pTSS), fornecem informações importantes não somente sobre o início de transcrição dos genes, mas também sobre a presença de regiões 5' UTRs. Dados de dRNA-seq de *H. pylori* indicam que a maioria dos transcritos dessa bactéria apresentam uma região 5' UTR entre 20 a 40 nucleotídeos, enquanto na haloarchaea *H. volcanii* 72% dos genes analisados são classificados como *leaderless* (Babski et al., 2016). A identificação dos inícios de transcrição primários auxiliam também na caracterização da região

promotora, além de auxiliar na identificação de genes que apresentam mais de um promotor, gerando transcritos com diferentes regiões 5' UTRs e que podem ser diferencialmente regulados (Sharma & Vogel, 2014).

A identificação de inícios de transcrição antisense a regiões codificantes (aTSS) e também aqueles localizados em regiões intergênicas sem relação com regiões gênicas (órfãos, oTSS) podem estar relacionados com a produção de pequenos RNAs não codificantes (sRNAs) com potencial regulatório. Em *H. pylori* dados de dRNA-seq permitiram a identificação de mais de 60 sRNAs, enquanto em *Salmonella typhimurium* em torno de 200 sRNAs primários foram identificados através do dRNA-seq (Kröger *et al.*, 2013). Os inícios de transcrição órfãos também podem fornecer informações importantes acerca das smORFs, pequenas ORFs como potencial codificante e usualmente desconsideradas dos processos de anotação. Em *H. pylori* e *Bacillus subtilis* inícios de transcrição órfãos auxiliaram na identificação de sistemas de toxina-antitoxina do tipo I com a presença de sRNA na fita oposta (Sharma *et al.*, 2010; Imov *et al.*, 2010).

Por fim, os inícios de transcrição internos a genes (iTSS) constituem uma classe muito abundante de inícios de transcrição nos experimentos de dRNA-seq realizados em diferentes organismos (Thomason *et al.*, 2015; Mitschke *et al.*, 2011). Em geral esses TSSs são associados a transcrição de genes *downstream* ou a erros de anotação da extremidade 5' de regiões codificantes (Stazic & Voß, 2016) além de sinalizarem o início de transcrição de pequenos RNAs não codificantes sobrepostos a extremidade 3' dos genes (Gomes-Filho *et al.*, 2015; Chao *et al.*, 2012; Bilusic *et al.*, 2014). Na cianobactéria *Synechocystis sp.* PCC6803, a combinação de dados de dRNA-seq com experimentos de *Northern Blot* demonstrou o acúmulo de pequenos RNAs produzidos internamente e na mesma orientação do gene *ntcA*, codificante de uma proteína regulatória com efeito positivo na via de captação de nitrogênio (Mitschke *et al.*, 2011). O cruzamento das informações de iTSSs com experimentos de

microarray demonstrou a existência de pequenos RNAs *sense* internos a outros genes dessa espécie (Mitschke *et al.*, 2011). Resultados similares foram encontrados também em genes da bactéria patogênica *Neisseria gonorrhoeae* (Wachter & Hill, 2015). Inícios de transcrição localizados dentro de operons também podem estar relacionados com a expressão modular dessas regiões (Güell *et al.*, 2011). Em *H. salinarum* NRC-1, a identificação de sítios de ligação de fatores de transcrições dentro de genes dispostos em operons, em conjunto com a análise de dados de expressão gênica ao longo da curva de crescimento, levou a identificação de expressão de subunidades dos operons, os quais eram diferencilmente regulados de acordo com a condição de crescimento (Koide *et al.*, 2009). A análise integrada de inícios de transcrição dentro de genes em oito diferentes espécies do gênero *Shewanella* levou a identificação de 307 iTSS no genoma de *Shewanella oneidensis* MR-1, dos quais, 30% são conservados em outras espécies do gênero, indicando que esses inícios de transcrição devem apresentar funcionalidade nesse grupo de organismos (Shao *et al.*, 2014).

Além da identificação das posições associadas ao início de transcrição outra informação relevante para a caracterização dos RNAs componentes do transcrito é a identificação dos termos de transcrição. Como apresentado anteriormente, os dados de RNA-seq representam o sequenciamento de um compêndio de transcritos em diferentes estágios de expressão, desde RNAs sendo transcritos até aqueles envolvidos em vias de degradação. Essa composição do sinal dificulta a identificação em larga escala de uma posição específica de término de transcrição, a qual é usualmente é representada por um decaimento gradativo no sinal de cobertura próximo a extremidade 3' dos genes. Em Bactérias existem dois principais mecanismos de terminação da transcrição: o mecanismo dependente a proteína *Rho* e outro mecanismo relacionado com a presença de estruturas em forma de *hairpin* e o reconhecimento de sequências específicas presente no DNA, sendo denominado de terminador intrínseco da transcrição (Santangelo & Artsimovitch, 2011). Em

Archaea ainda não são conhecidos em muitos detalhes os mecanismos associados ao término de transcrição. Em *Methanooccus thermolithotrophicus*, uma sequência oligo-dT se mostrou suficiente a terminação intrínseca da transcrição (Santangelo & Reeve, 2006). Além disso, a busca por terminadores em *T. kadakaraensis* confirmou que sequências poliT atuam como terminadores intrínsecos de transcrição, e também demonstrou que esse processo é completamente independente da presença de estruturas (Santangelo *et al.*, 2009). Apesar da presença de sequências poliT ser o mecanismo mais provável de terminação em Archaea ainda é incerto se essa é a sequência responsável pela terminação em genomas com alto conteúdo GC, como em *H. salinarum*. Análises integradas com informações sobre sítios de término de transcrição e dados de RNA-seq podem ser úteis para a confirmação de que sequências poliT podem realmente atuar como terminadores de transcrição nessa Archaea, auxiliando na identificação correta dos termos de transcrição.

2 Objetivos

2.1 Geral

- Analisar dados de expressão gênica da archaea extremofílica *Halobacterium salinarum* NRC-1 gerados através da técnica de RNA-seq, utilizando abordagens computacionais e experimentais, visando à identificação de novos transcritos e suas extremidades em escala genômica.

2.2 Específicos

- Preparar e analisar os dados de RNA-seq de *H. salinarum* NRC-1 utilizando metodologias de processamento e segmentação de sinais no intuito de identificar posições relacionadas ao início ou término de transcrição;
- Utilizar a metodologia de dRNA-seq em amostras de RNA extraídas de *H. salinarum* NRC-1 visando à detecção dos inícios de transcrição, e consequentemente identificar novos elementos de importância biológica no transcrito dessa archaea;
- Analisar dados públicos de expressão gênica de *H. salinarum* NRC-1, e outros organismos, no intuito de integrar as informações previamente obtidas com dados provenientes de outras plataformas ou obtidos em diferentes condições de crescimento, visando a compreender o comportamento de transcritos de interesse em diferentes condições;
- Validar experimentalmente a existência de novos transcritos e avaliar seu efeito no fenótipo de crescimento de *H. salinarum* NRC-1.

3 Materiais e Métodos

3.1 Oligonucleotídeos

A Tabela 2 apresenta o nome, sequência e aplicação de cada um dos oligonucleotídeos utilizados em diferentes etapas do desenvolvimento do trabalho. Os oligonucleotídeos foram sintetizados pelas empresas Exxtend ou IDT.

Tabela 2: Oligonucleotídeos utilizados no trabalho

Nome	Sequência	Aplicação
Forward_intraVNG0045C	ATCGCCTAGGGTTCTTGACGAGGTCGATG	Superexpressão
Reverse_intraVNG0045C	ATCTTCTAGACTACAGCACGCCGTTCCGA	Superexpressão
Forward_intraVNG1339C	ATTACCTAGGCCCGTGAGGTCGAGGAGT	Superexpressão
Reverse_intraVNG1339C	ATCTTCTAGAGCCGGGTCAGTCCTCGT	Superexpressão
2068_fwd_ecorI	CGGCGAATTCATGCGAGTCATCATCGTC	Superexpressão
2068_rev_hindIII	ATGAAAGCTTGCGCGACTACCCGTCGTC	Superexpressão
6363_fwd_ecorI	TATTGAATTCGGGGAGGTCGCCTACGA	Superexpressão
6363_rev_hindIII	TATCAAGCTTCGTCGTCCGCACCCAGTCA	Superexpressão
M13 Reverse	AGCGGATAACAATTTACACAG	Validação de construção
M13 Forward	CGCCAGGGTTTTCCAGTCACGAC	Validação de construção
+19pNRC200Fow	AACCACCGTGACCCGGAAGC	Confirmação da degradação do DNA genômico
+19pNRC200Rev	GCCGATCCGATGGGAGTGACT	Confirmação da degradação do DNA genômico
16S Fow	GTGAGGGACGAAAGCTAGGG	Confirmação da degradação do DNA genômico
16S Rev	AAGGGTACTGCTGGCAATTA	Confirmação da degradação do DNA genômico

3.2 Linhagem e condição de cultivo

Os RNAs foram extraídos de *H. salinarum* NRC-1 crescendo em meio CM (250g/l NaCl, 20g/l MgSO₄, 2 g/l KCl, 3 g/l Citrato de Sódio, 10 g/l Peptona Bacteriológica Oxoid) na condição padrão como previamente descrito (DasSarma *et al.*, 1995). Quando a cultura atingiu a densidade óptica (DO) em torno de 0,5 (metade da primeira fase log de

crescimento), alíquotas de 2 ml foram coletadas em tubos *ependorfs* e centrifugados a 14.000 rpm por 1 minuto. Após a centrifugação o sobrenadante foi desprezado e os *pellets* foram armazenados a - 80° C até que fosse realizada a extração dos RNAs. Para a coleta de amostras ao longo da curva de crescimento a linhagem NRC-1 foi cultivada em meio CM em volume de 200 ml à 37° C sob agitação (125 rpm) e luminosidade constantes. Os *pellets* foram coletados como previamente apresentado em três momentos distintos do crescimento, a 17 horas, 37 horas e 86 horas (Figura 7). Todos os experimentos foram realizados com réplica biológica.

3.3 Extração dos RNAs

Para a extração dos RNAs utilizamos o kit *mirVana miRNA Isolation* (Ambion). As extrações foram feitas em réplicas a partir de um único *pellet*, e foi empregado o protocolo para extração de RNA total. Ao final da extração as amostras foram quantificadas em espectrofotômetro.

A fim de eliminar qualquer contaminação com DNA genômico ou plasmidial, as amostras foram tratadas com Turbo™ DNase Ambion® na proporção de 1U para cada 3µg de RNA na presença de 2U de inibidor de RNase RiboLock™ Thermo Scientific para cada 1µg de RNA. A reação foi incubada à 37° C por 30 minutos e logo após foi realizada purificação com Fenol/Clorofórmio e precipitação com etanol. Para confirmar a degradação dos possíveis contaminantes de DNA foram realizadas reações de PCR utilizando 10 µM de dois pares de primers +19pNRC200Fow/Rev e 16SFow/Rev (Tabela 2), 0,5 mM de MgCl₂, 1X de solução KAPA2G Fast ReadyMix, 5% de DMSO (Kapa Biosystems). As condições da PCR foram de 95°C por 15 minutos, seguidos de 25 ciclos de 95° C por 15 segundos, 58° C por 15 segundos e 72°C por 30 segundos. Como controle positivo foi utilizado DNA genômico de *H.*

salinarum NRC-1. O produto da PCR foi submetido à eletroforese em gel de agarose 3% em tampão de corrida TAE, utilizando o marcador de peso molecular *GeneRuler 50 pb*

3.4 Depleção dos 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).

3.5 Preparação das bibliotecas de dRNA-seq

As amostras utilizadas na preparação das bibliotecas de dRNA-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. Optamos pela utilização do kit para pequenos RNAs pois seu

protocolo não possui a etapa de fragmentação dos RNAs, o que invalidaria o enriquecimento para RNAs primários. 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 segundo 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 abaixo do marcador de 145 pb e acima 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.

3.6 Sequenciamento e processamento dos dados

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. Os dados foram processados para visualização na ferramenta Gaggle Genome Browser – GGB (Bare *et al.*, 2010). Para determinar de forma automatizada os inícios de transcrição de *H. salinarum* NRC-1, os dados de dRNA-seq foram analisados com a ferramenta TSSAR (Amman *et al.*, 2014). Brevemente a ferramenta 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. 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. A ferramenta está disponível em <http://nibiru.tbi.univie.ac.at/TSSAR/>.

3.7 Dados públicos analisados

A identificação de TSS em outros procariotos foi realizada através da análise de dados públicos de dRNA-seq disponibilizados em trabalhos anteriores conforme Tabela 3. Nos casos onde os dados brutos não foram disponibilizados utilizamos a lista de TSSs fornecida como material suplementar dos trabalhos analisados. Os dados foram processados de maneira

similar ao descrito na seção anterior (3.6). Para identificação dos intraRNAs diferencialmente expressos realizamos a análise de dados de *tiling microarray* disponível para *H. salinarum* NRC-1 em oito condições distintas (Tabela 4).

A curadoria de inícios de transcrição internos foi realizada com anotações de *H. salinarum* NRC-1 disponíveis nos bancos de dados *Integrated Microbial Genomes* (<https://img.jgi.doe.gov/>), *UniProt* (<http://www.uniprot.org/>), *RefSeq* versão 2011 e a anotação disponibilizada pelo *Institute for Systems Biology*, responsável pelo sequenciamento do genoma da linhagem NRC-1, além de anotações realizados por meio das ferramentas *EasyGenes* e *BlastP* utilizando a duas anotações de proteínas da linhagem R1.

Tabela 3: Organismos utilizados para análise de conservação de inícios de transcrição internos a regiões codificantes.

Organismos	Acesso NCBI	Referência
Campylobacter jejuni RM1221	GSE38883	Dugar <i>et al</i> , 2013
Haloferax volcanii	GSE82206	Babski <i>et al</i> , 2016
Helicobacter pilory	SRA010186	Sharma <i>et al</i> , 2010
Neisseria lactamica	GSE44582	Zhang <i>et al</i> , 2013
Shewanella oneidensis MR-1	GSE58337	Shao <i>et al</i> , 2014
Thermococcus kodakarensis KOD1	GSE56262	Jäger <i>et al</i> , 2014

Tabela 4: Dados de tiling microarray de H. salinarum NRC-1 utilizados nesse trabalho.

Condição	Acesso NCBI	Referência
Superexpressão TfbD	GSE15788	Koide <i>et al</i> , 2009
Alteração nos níveis de oxigênio	-	dados não publicados
Curva de crescimento	GSE12923	Koide <i>et al</i> , 2009
Alteração nas condições de salinidade	GSE53544	Beer <i>et al</i> , 2014
Δ VNG2099 (RNase)	GSE45988	Wurtmann <i>et al</i> , 2014
Estresse oxidativo	GSE61975	Plaisier <i>et al</i> , 2014
Deleção de fatores de transcrição e produção de fator de transcrição sintético	GSE31308	Turkarlan <i>et al</i> , 2011
Δ lsm	-	dados não publicados

3.8 Programas e *scripts* utilizados nas análises

Todos os *scripts* implementados em nosso laboratório foram escritos utilizando a linguagem de programação R (R Core Team, 2016). Para identificação do enriquecimento funcional foi utilizado o pacote GOstats (Falcon & Gentleman, 2007), disponível na plataforma Bioconductor (<http://bioconductor.org>). A anotação de domínios pFAM no genoma de *H. salinarum* NRC-1 foi realizada através do software Sifter-T (Almeida-E-Silva & Vêncio, 2015) em colaboração com o MSc. Danilo Silva sob supervisão do Prof. Dr. Ricardo Vêncio, para os demais organismos utilizamos a anotação de domínios pFAM disponível no banco de dados Uniprot (Bateman *et al.*, 2015) disponível em <http://www.uniprot.org/>. A geração dos logos de sequência da região promotora dos TSSs foi realizada através da aplicação *web* Weblogo (Crooks *et al.*, 2004) disponível em <http://weblogo.berkeley.edu/>. Para visualização dos dados de RNA-seq, e também dos resultados dos algoritmos de segmentação e subsegmentação, foi utilizado o *software* Gaggie Genome Browser (Bare *et al.* 2010).

3.9 TAG cromossomal

Para validação das proteínas traduzidas a partir dos transcritos internos (intraRNAs) foram produzidas linhagens de *H. salinarum* NRC-1 transformadas para adição de um TAG cromossomal, para expressão do epítipo FLAG – DYKDDDDK, na mesma fase de leitura do gene cognato e precedendo o códon de parada de tradução (Munro & Pelham, 1984). As transformações foram realizadas pela Dr^a Livia Zaramela Soares de acordo com protocolos previamente descritos (Silva-Rocha *et al.*, 2015; Wilbanks *et al.*, 2012). Dessa forma foi possível detectar tanto a versão original da proteína alvo quanto sua versão alternativa através de *Western Blot* com anticorpo específico para o TAG utilizado.

3.10 Tricina-SDS-PAGE e *Western blot*

Após a confirmação da presença da sequência correspondente ao epítipo FLAG as linhagens foram inoculadas em meio de cultura CM e crescidas de acordo com a condição padrão descrita anteriormente até atingir a DO de 0,5. Para obtenção dos *pellets* 4 ml de meio de cultura foram centrifugados a 10000 rpm por 1 minuto a temperatura ambiente, o sobrenadante foi descartado. Logo após a obtenção dos *pellets* foi realizada a extração total de proteínas através de lise celular com o uso de 100 µL de Tampão de Lise Triplo (Tris-HCl pH 8 10mM, NaCl 50mM, NaF 50mM, Triton X-100 1%, SDS 0,1% e Deoxilato de Sódio 6nM) e 1,5µl de coquetel inibidor de protease (Sigma Aldrich P8340). As amostras foram sonicadas por 30 segundo com intervalos de 5 segundos por 6 vezes e incubadas em gelo por 20 minutos. Após essa etapa foi realizada a centrifugação a 13500 rpm a 4° C durante 20 minutos. O sobrenadante, contendo as proteínas, foi transferido para novo tubo e a concentração de proteínas foi aferida pelo método de Bradford (Green & Sambrook D.W., 2012). Durante todas as etapas as amostras foram mantidas em gelo para evitar a degradação das proteínas. Após a extração e quantificação as amostras foram armazenadas em freezer a -80° C para uso posterior.

As proteínas presentes nas amostras foram separadas por meio da técnica de eletroforese Tricina-SDS-PAGE, na qual é utilizada um tampão de corrida baseado em Tricina, ao invés da usual Glicina, com o objetivo de obter uma melhor separação e detecção de proteínas de baixo peso molecular (1 – 100 kDa). Os géis foram elaborados conforme especificações da Tabela 5.

Tabela 5: Reagentes e volumes empregados na preparação dos géis

	Gel de empilhamento (4%)	Gel de resolução (15%)
H ₂ O	2,05 ml	1,45 ml
3M Tris-HCl/SDS	0,775 ml	2,5 ml
40% Acrilamida/Bis-Acrilamida 19:1	0,3 ml	2,8 ml
Glicerol	-	0,75 ml
30% Persulfato de Amônia (APS)	7 µL	7 µL
TEMED	7 µL	7 µL

Para realização da eletroforese 20 a 80 µg de proteínas foram misturadas ao tampão de amostra (glicerol 10% v/v; 50 mM Tris-HCl pH 6.8, SDS 2% p/v, Azul de bromofenol 0,003% p/v, β-Mercaptoetanol 1% v/v) e desnaturadas por 10 minutos a 95° C. As amostras foram aplicadas no gel e a corrida foi realizada a uma voltagem inicial de 40 V para promover a entrada das proteínas no gel de empilhamento e após 100 V por aproximadamente 2,5 horas. Como marcador de peso molecular foi utilizado o *PageRuler Prestained Protein Ladder* (Thermo Fisher Scientific). Foram utilizados dois tampões de corrida distintos, o tampão do cátodo (100 mM Tris, 100 mM Tricina, 0,1 % SDS, pH 8,3) e o tampão do ânodo (200 mM Tris pH 8,8). Após a corrida as proteínas foram transferidas para membrana de nitrocelulose 0,2 µm (Amersham Protran Premium – GE Healthcare Life Sciences) com tampão de transferência sem SDS (Tris 24,7 mM, Glicina 192 mM, Etanol 18,46% v/v) a 80 V por 1 h. A eficiência da transferência foi avaliada corando as membranas com o corante Ponceau S por 10 min. Após a confirmação da transferência as membranas foram lavadas com água destilada para remoção do corante.

Para realização do *Western Blot* as membranas foram bloqueadas em leite 5% em TBS-Tween (0,1%) por 1 hora a temperatura ambiente. A marcação foi realizada com anticorpo monoclonal anti-FLAG conjugado a HRP (Sigma A8592) diluído na concentração de 1:1000 em TBS-T por 1 h a temperatura ambiente e após foram realizadas 5 lavagens de 5

minutos com TBS-T. A revelação das membranas foi realizada utilizando solução de ECL preparada em nosso laboratório (Solução 1: Tris-HcL pH 8.5 100mM, Luminol 2,5 mM, Ácido Cumárico 396 µM; Solução 2: Tris-HcL pH 8.5 100mM, Peróxido de hidrogênio 0,018%). As duas soluções foram misturadas em quantidades iguais em volume suficiente para cobrir a membrana. As membranas foram incubadas na solução reveladora por 1 minuto e reveladas no equipamento *ChemiDoc MP system* com a opção de detecção de quimioluminescência e aquisição de 20 imagens com incremento e 30 segundos de exposição.

3.11 Superexpressão de intraRNAs

Para superexpressão dos intraRNAs utilizamos o vetor de superexpressão pHSal-E, o qual é baseado no promotor constitutivo forte do gene *fer2* de *H. salinarum* NRC-1 e apresenta origens de replicação de *E. coli* e *H. salinarum* e os genes de resistência aos antibióticos carbenicilina e mevinolina (Silva-Rocha *et al.*, 2015). Oligonucleotídeos específicos para os intraRNAs alvos foram desenhados com a adição de sítios de restrição para as seguintes enzimas AVRII/XbaI ou EcoRI/HindIII (Tabela 2). Os intraRNAs alvos foram amplificados por meio de reações de PCR contendo 1 mM de MgSO₄, 300 µM de dNTP, 1X de Pfx Amplification Buffer, 1X de PCR Enhancer Solution, 1 U de Platinum *Pfx* DNA Polimerase (Thermo Fisher Scientific) e 50 ng de DNA genômico em uma reação com volume final de 50 µl. As condições da PCR foram de 95°C por 5 minutos, seguidos de 35 ciclos de 95° C por 15 segundos, 60° C por 30 segundos e 68°C por 1 minuto, com uma etapa final de extensão de 68° C por 5 minutos. Após a amplificação os produtos de PCR foram purificados com o kit *Wizard SV Gel and PCR Clean-Up System* (Promega) e foi realizada uma digestão dupla com as enzimas de restrição específicas para cada produto com 10 U das enzimas HindIII, EcoRI e XbaI, 2,5 U da enzima AvrII, 1X do Buffer 2.1 para o par Hind III/EcoRI e 1X do Buffer CutSmart para o par XbaI/AvrII (New England BioLabs) em um

volume final de 50µL a 37° C por 1 hora. Os produtos da digestão foram purificados como descrito anteriormente. De forma similar realizamos a digestão do vetor pHSal-E com 50U de cada enzima de restrição e 2,8 µg do vetor. Após a digestão isolamos a porção de interesse do vetor através de eletroforese e purificação do fragmento em gel de agarose por meio do kit *Wizard SV Gel and PCR Clean-Up System* (Promega). O fragmento de interesse do vetor digerido foi desfosforilado utilizando 3 U da enzima SAP (New England BioLabs) e 1X de Buffer a 37° C por 1 hora. A ligação dos produtos de PCR e vetor digeridos foi realizado com 5 U da enzima T4 DNA Ligase (Thermo Fisher Scientific) e 1X de Buffer em uma reação de volume final de 10µl a 4° C durante 16 horas. O produto da ligação foi clonado em linhagem competente de *Escherichia coli*, de acordo com o protocolo padrão, a qual foi plaqueada em meio Luria-Bertani (LB) sólido suplementado com o antibiótico marcador de resistência carbenicilina (Sigma) e incubadas a 37° C por 24 horas. Para identificação das colônias com os vetores ligados de forma correta aos insertos de interesse foi realizado PCR de colônia contendo 0,5 mM de MgCl₂, 1X de solução KAPA2G Fast ReadyMix, 5% de DMSO (Kapa Biosystems) e um oligonucleotídeo específico para o inserto e outro específico para o vetor (M13) (Tabela 2). As condições da PCR foram de 95°C por 15 minutos, seguidos de 25 ciclos de 95° C por 15 segundos, 58° C por 15 segundos e 72°C por 30 segundos. Após a identificação das colônias de *E. coli* contendo a construção de interesse as mesmas foram inoculadas em meio LB líquido suplementado com carbenicilina (Sigma) a 37° C pelo período de 24 horas. A extração plasmidial foi realizada utilizando o kit *PureYield Plasmid Miniprep System* (Promega). A linhagem NRC-1 de *H. salinarum* foi então transformada com a adição dos vetores de interesse através de protocolos já estabelecidos (DasSarma *et al.*, 1995). A linhagem transformada foi plaqueada em meio CM sólido contendo o antibiótico mevinolina (A.G. Scientific) e incubada a 37° C pelo período de 21 dias. Após esse período as colônias presentes na placa foram inoculadas em meio CM líquido suplementado com

mevinolina e crescidas na condição padrão anteriormente descrita até atingir a DO 0,5, quando foram coletadas amostras e realizados estoques em glicerol para armazenar em freezer -80° C até uso posterior.

3.12 Avaliação da superexpressão de intraRNAs

Para avaliar o efeito da superexpressão dos intraRNAs foram realizadas curvas de crescimento em meio CM, ou meio deficiente em potássio (250 g/l NaCl, 27 g/l MgSO₄, 0,14 g/l KCl, 3 g/l Citrato de Sódio, 10 g/l Peptona), suplementado com 20 µg/mL de mevinolina em banho-maria a 37°, agitação a 125 rpm e iluminação constante, em volume de 50 ml. Aos meio com restrição de potássio foi adicionado 100mM de KCl após a entrada na fase estacionária (Wurtmann *et al.*, 2014). A avaliação do crescimento foi realizada através da medida da densidade óptica a 600 nm ao longo do tempo.

4 Resultados e Discussão

4.1 Identificação de inícios de transcrição (TSS)

A identificação de inícios de transcrição no genoma permite o discernimento entre transcritos primários e produtos de processamento, auxiliando na identificação de novos transcritos, no estudo de suas regiões regulatórias, adicionando assim elementos para o estudo do fenômeno de transcrição pervasiva. A integração dessas informações com dados de expressão gênica, proteômica, sítios de ligação de fatores de transcrição e análise de conservação em espécies relacionadas, pode fornecer indícios importantes sobre o papel funcional de novos transcritos, além de auxiliar na melhor caracterização de regiões genômicas já conhecidas.

Para obter dados experimentais relacionados aos inícios de transcrição em *H. salinarum* NRC-1 optamos pela realização do *differential* RNA-seq. Essa técnica permite a identificação de forma global dos inícios de transcrição através da comparação de uma biblioteca enriquecida para RNAs primários (TEX+) e uma biblioteca controle não enriquecida (TEX-). Essas informações permitiram (i) refinar a identificação de RNAs associados ao início de transcrição (TSSaRNAs) e (ii) mapear TSSs em diferentes etapas da curva de crescimento em *H. salinarum*, ampliando o espectro de transcritos identificados nesta archaea, muitos com potenciais papéis funcionais.

4.1.1 RNAs associados ao início de transcrição

Os RNAs associados ao início de transcrição (**TSS** associated RNAs; TSSaRNAs) são uma classe de ncRNAs já descrita em eucariotos e bactérias, cujo papel funcional pode estar envolvido na regulação do início da transcrição (Seila et al., 2008, Taft et al., 2009 e

Henriques et al., 2013). Em bactérias por exemplo, esses ncRNAs podem estar envolvidos em vias regulatórias que impedem o início de transcrição até que o RNA polimerase esteja montada (Yus *et al.*, 2012). Em eucariotos, a biogênese desses ncRNAs está associada à clivagem endonucleolítica dos RNAs nascentes, já em bactérias a produção desses ncRNAs está associada com paradas da RNA polimerase durante o processo de elongação da transcrição, liberando o transcrito nascente (Yus *et al.*, 2012; Seila *et al.*, 2008).

Em archaeas, essa classe de RNAs foi identificada pelo nosso grupo de pesquisa, no artigo científico intitulado *Transcription start site associated RNAs (TSSaRNAs) are ubiquitous in all domains of life* (Anexo K), publicado em setembro de 2015. O uso da técnica de dRNA-seq na identificação desses RNAs permitiu a diferenciação entre verdadeiros TSSaRNAs e pequenos RNAs gerados através do processamento de mRNAs.

A identificação inicial dos TSSaRNAs foi realizada através da análise de dados de sequenciamento de RNA-seq convencional para pequenos RNAs, buscando identificar os *reads* completos (menores de 151 nt – *full length*) mais abundantes associados ao início dos genes de *H. salinarum* NRC-1. De acordo com essa metodologia, foi possível identificar 652 RNAs associados ao início de transcrição. Para identificação dos TSSaRNAs primários, realizamos o dRNA-seq a partir de RNAs extraídos de *H. salinarum* NRC-1 na condição referência de crescimento. O sequenciamento foi realizado com a presença de réplica biológica, no entanto como uma das réplicas apresentou um grande desbalanço do número de *reads* obtidos, ela foi desconsiderada das análises (Tabela 6 - REF 1 2014). Prosseguindo a análise somente com os *reads* pertencentes a segunda biblioteca foram identificados 435.339 *reads* correspondentes aos TSSaRNAs. 179 TSSaRNAs apresentaram um enriquecimento de 95% no número de *reads* na biblioteca TEX + em relação a biblioteca TEX- e foram considerados como primários. A biogênese dos TSSaRNAs está associada como a parada da

RNA polimerase durante o processo de transcrição e foi modelada computacionalmente levando em consideração o tempo de retenção em uma determinada posição e o tempo entre eventos de início de transcrição. O uso de diferentes sítios de parada dentro de um gene, em conjunto com diferentes tempos de retenção, é condizente com o acúmulo de *reads* do RNA-seq com tamanhos distintos mas com início em uma mesma posição do gene. Essa observação foi validada experimentalmente por meio de *Northern blot*. Para o gene VNG0101G (VNG_RS00395) foi possível identificar diferentes bandas com tamanhos próximos aos tamanhos apresentados pelo acúmulo de *reads*, indicando a existência de múltiplos sítios de parada da RNA polimerase dentro do gene. Já o gene VNG1213C (VNG_RS04745) apresentou uma única banda no experimento de *Northern blot*, correlacionada com o acúmulo de *reads* de um único tamanho (Figura 3 – Anexo K). A biogênese desses transcritos através da retenção da RNA polimerase durante o processo de transcrição também está relacionada com a expressão diferencial dos TSSaRNAs em relação ao seu gene cognato, a qual foi identificada em 10 TSSaRNAs de *H. alinarum* NRC-1. Esse comportamento também pode ser reproduzido explorando os dois parâmetros do modelo computacional e sugere a existência de uma regulação nos tempos de retenção da RNA polimerase em diferentes condições de crescimento, ocasionando uma expressão diferencial dos TSSaRNAs. Por fim, a análise adicional em dados públicos de *microarray* identificaram a existência de TSSaRNAs em outros 10 genomas de archaea, demonstrando que esses transcritos são amplamente distribuídos nos três domínios da vida.

4.1.2 TSSs durante a curva de crescimento de *H. salinarum* NRC-1

Por se tratar de uma técnica de identificação de inícios de transcrição baseada na análise do sinal de expressão gênica, a identificação dos TSSs através do dRNA-seq se

restringe somente aos genes expressos na condição analisada. Para expandir o número de posições identificadas e anotar inícios de transcrição de genes expressos em condições específicas, o dRNA-seq pode ser realizado a partir de amostras de RNA obtidas em diferentes condições de crescimento ou a partir do crescimento frente a diferentes agentes estressantes. Em *H. salinarum* NRC-1, por exemplo, o simples crescimento em meio complexo é capaz de promover alterações nos níveis transcricionais de até 63% dos genes não redundantes (Facciotti *et al.*, 2010). Dessa forma, considerando também a existência de um grande desbalanço no número de *reads* obtidos em uma das réplicas do experimento de dRNA-seq realizado para identificação dos TSSaRNAs (Tabela 6), o que poderia enviesar as análises posteriores, optamos pela realização de novos experimentos de dRNA-seq a partir de amostras obtidas em diferentes pontos da curva de crescimento de *H. salinarum* NRC-1 (Seção 3.2). Para ampliarmos o número de genes identificados foram escolhidos momentos representativos do crescimento, os momentos de 17 horas (17h), representando a primeira fase de crescimento, 37 horas (37h) quando as células se preparam para sair da fase estacionária e 86 horas (86h) quando o crescimento está estagnado, além de realizar adquirir novos dados para a condição referência de crescimento (Facciotti *et al.*, 2010) (Figura 7).

Figura 7 Curva de crescimento de *H. salinarum* NRC-1. Os círculos pretos indicam o tempo da curva de crescimento em que foram coletados pellets para posterior extração de RNA total e preparação das bibliotecas de *drNA-seq* (17h, 37h e 86h).

Após a extração de RNAs e tratamento com a enzima *Terminator Exonuclease* (Epicentre), as amostras enriquecidas com RNAs primários (TEX +) e as amostras controle (TEX -) foram utilizadas para preparação de bibliotecas de pequenos RNAs para sequenciamento no equipamento Illumina MiSeq, disponível na Faculdade de Medicina de Ribeirão Preto. Ao final do sequenciamento foram produzidos 30.096.003 *reads*. Como uma parte *reads* gerados são formados integral ou parcialmente por sequenciais oriundas dos adaptadores, utilizamos a ferramenta *FASTX-Toolkit* (http://hannonlab.cshl.edu/fastx_toolkit/) para remoção dessas sequências. Após esse tratamento, a quantidade final de *reads* foi de 28.366.040, com tamanho médio de aproximadamente 65 nucleotídeos, dos quais, em média, 61,02% foram mapeados ao genoma referência. O número de *reads* obtidos e as porcentagens de alinhamento para cada biblioteca é apresentado na Tabela 6. Para a identificação dos inícios de transcrição, optamos por utilizar também os dados balanceados produzidos anteriormente para a condição referência de crescimento (REF 2 2014). Dessa forma as

análises foram conduzidas em duplicata biológica para os três diferentes pontos na curva de crescimento e em triplicata para a condição referência de crescimento.

Tabela 6: Número de reads sequenciados, processados e mapeados para cada biblioteca

Bibliotecas	# reads	# reads após trimagem	# reads mapeados	% reads mapeados
REF 1 + (2014)	5.239.354	4.420.459	3.546.868	80,2%
REF 1 - (2014)	1.878.153	1.132.854	852.323	75,2%
REF 2 + (2014)	1.950.604	1.265.838	891.916	70,5%
REF -2 (2014)	2.378.435	1.580.141	1.114.406	70,5%
17h 1 +	1.922.014	1.874.168	1.401.212	74,8%
17h 1 -	2.175.004	2.137.572	1.567.107	73,3%
17h 2+	2.641.063	2.617.955	2.003.906	76,5%
17h 2-	1.891.117	1.630.334	1.162.493	71,3%
37h 1+	2.647.902	2.634.169	1.944.688	73,8%
37h 1-	3.238.764	3.141.112	2.157.941	68,7%
37h 2+	2.069.923	1.891.024	1.060.217	56,1%
37h 2-	2.264.181	2.247.138	1.629.073	72,5%
86h 1+	1.593.528	1.441.896	588.042	40,8%
86h 1-	1.532.836	1.306.610	596.032	45,6%
86h 2+	1.845.982	1.518.038	611.363	40,3%
86h 2-	1.570.241	1.486.808	687.382	46,2%
REF 1+ (2015)	1.468.041	1.356.827	726.703	53,6%
REF 1- (2015)	1.245.709	1.116.957	581.423	52,1%
REF 2+ (2015)	1.044.202	1.029.723	673.246	65,4%
REF 2- (2015)	945.496	935.709	611.522	65,4%

4.1.3 Identificação automatizada de TSS em dados de dRNA-seq

Para identificação dos inícios de transcrição os dados foram analisados através da ferramenta TSSAR (<http://nibiru.tbi.univie.ac.at/TSSAR/>), um *webservice* desenvolvido para a identificação de TSSs através da análise estatística de dados de dRNA-seq em busca de posições enriquecidas nas bibliotecas de RNAs primários (Amman *et al.*, 2014). Os dados provenientes de cada réplica biológica foram analisados de forma conjunta para cada condição investigada. Para ser definido como um início de transcrição, a posição enriquecida na biblioteca TEX+ deveria apresentar 10 ou mais *reads* iniciando na mesma posição e

apresentar um p-valor menor que 0,001. TSSs identificados em uma região menor que 10pb foram reunidos em uma única posição (Seção 3.6). De acordo com esses critérios, a ferramenta identificou um total de 4540 posições envolvidas com o início de transcrição de RNAs em pelo menos umas das condições analisadas (Anexo A). Destas, 19,37% (880) foram identificadas em todas as condições investigadas e 41,79% (1898) foram identificadas em apenas uma das condições. O ponto de 17 horas foi a condição com maior número de TSSs únicos, destacando assim a importância de analisar os inícios de transcrição em diferentes condições (Figura 8). Em relação a localização no genoma de *H. salinarum* NRC-1, 3880 TSSs foram identificados no cromossomo principal, 477 no plasmídeo pNRC200 e 183 plasmídeo pNRC100 (Tabela 7).

Figura 8: TSSs identificados em cada ponto da curva de crescimento. Diagrama de Venn representando a sobreposição de TSS identificados em cada uma das condições de crescimento analisadas

Tabela 7: Distribuição dos TSSs nas três regiões genômicas de *H. salinarum* NRC-1

Bibliotecas	Cromossomo	Plasmídeo pNRC100	Plasmídeo pNRC200	Total
17h 1 + 17h 2	2021	61	208	2290
37h 1 + 37h 2	2326	112	229	2667
86h 1 + 86h 2	1825	98	237	2160
REF 1 + REF 2 + REF (2014)	2351	92	219	2662

4.1.4 Anotação dos TSSs

A partir da identificação das posições genômicas associadas ao início de transcrição, realizamos uma classificação dessas posições de acordo com sua localização em relação aos genes de *H. salinarum* NRC-1. Dessa forma, cada TSS foi classificado da seguinte maneira: primário (pTSS), caso estivesse localizado até 250 pb *upstream* a uma região codificante; interno (iTSS), caso estivesse localizado internamente e na mesma orientação de uma região gênica; antisense (aTSS), caso fosse localizado internamente a uma região gênica mas na orientação oposta; órfão (oTSS), todos aqueles TSSs localizados em regiões intergênicas e não relacionados a qualquer gene (Figura 9).

Figura 9: Disposição dos TSSs em relação a anotação do genoma. pTSS: início de transcrição primário; iTSS: início de transcrição interno; aTSS: início de transcrição antisense; oTSS: início de transcrição órfão. A direção das setas indica a orientação no genoma. Adaptado de (Sass et al., 2015)

De acordo com essa classificação foi possível identificar um total de 1978 pTSSs, 1365 iTSSs, 737 aTSSs e 732 oTSSs no genoma de *H. salinarum* NRC-1, sendo que alguns dos inícios de transcrição podem pertencer a mais de uma classe, dependendo da posição onde estão localizados (Anexo A) (Figura 10).

Figura 10: Distribuição dos inícios de transcrição em quatro diferentes classes e as intersecções entre essas classes

Analisando a distribuição de TSSs em cada uma das condições investigadas é possível perceber um aumento no número de TSSs primários em 17 horas em relação a condição referência de crescimento, indo de 1097 posições identificadas para 1354, e se mantendo estável nas demais condições. O mesmo fenômeno se repete com os inícios de transcrição órfãos (Figura 11). Por outro lado, aTSSs e iTSSs apresentam um perfil de diminuição do número de posições identificadas. O número de inícios de transcrição antisense a regiões codificantes se mantêm estáveis nas primeiras etapas da curva de crescimento e diminuem na etapa final, já o número de inícios de transcrição internos apresentam uma queda ao longo da curva de crescimento (Figura 11). Essa diminuição do número de inícios de transcrição intragênicos, *sense* e *antisense*, pode estar relacionada com aumento na compactação do DNA, e consequentemente uma menor disponibilidade de sítios de interação a fatores de transcrição. Em *H. salinarum* já foi demonstrado que o cromossomo tende a estar em um estado livre de proteína nas fases iniciais de crescimento, aumentando sua compactação em

estruturas similares a nucleossomos na fase estacionária (Takayanagi *et al.*, 1992). Em *E. coli* e *Neisseria gonorrhoeae*, a proteína H-NS é responsável pela diminuição dos níveis de transcrição intragênica (Masters *et al.*, 2016; Singh *et al.*, 2014).

Figura 11: Número de TSSs identificados em cada classe de acordo com a condição analisada

Para verificarmos a conservação das sequências em torno dos TSSs identificados realizamos a análise da frequência de nucleotídeos através do *webservice* Weblogo (Crooks *et al.*, 2004), utilizando sequências correspondentes a região 40 pb *upstream* e 10 *downstream* ao TSS, desconsiderando da análise todos os TSSs agrupados em mais de uma classe (Figura 121). Todas as classes analisadas apresentaram alta frequência de nucleotídeos específicos em torno dos TSSs identificados, com atenção especial a presença de purinas na posição de início de transcrição e pirimidinas na posição -1. Essa característica já havia sido identificada em archaea halofílicas (Soppa, 1999) e também archaea do filo Crenarchaeota, como *Sulfolobus solfataricus* (Wurtzel *et al.*, 2010) e *Pyrobaculum aerophilum* (Slupska *et al.*, 2001), além de bactérias (Kim *et al.*, 2012) e eucariotos (Zhang, 2005; Carninci *et al.*, 2006), corroborando a

identificação de inícios de transcrição em *H. salinarum* NRC-1 através do dRNA-seq. Essa conservação está relacionada com o reconhecimento da posição de início de transcrição, sendo que alterações nessa composição de dinucleotídeos pode diminuir os níveis de expressão do gene correspondente (Carninci *et al.*, 2006). Além da conservação de purinas da posição +1, com preponderância de adenina, os TSS primários apresentam também alta frequência de nucleotídeos timina e guanina nas posições +2 e +3 respectivamente, indicando a presença de um códon de início de tradução junto a posição de início de transcrição, e consequentemente a ausência de uma região 5' UTR para a maioria dos pTSS analisados. Em relação a região promotora, é possível detectar a presença de um elemento de reconhecimento do fator de transcrição TFIIB (*B recognition element* – BRE) caracterizado pela alta frequência de nucleotídeos CG na posição -36, além de um TATA-box na posição -26. Esses elementos já foram amplamente descritos como responsáveis por promover o início de transcrição em archaea (Bell & Jackson, 2001; Seitzer *et al.*, 2012).

Figura 12: Análise da frequência nucleotídica nas regiões promotoras e de início de transcrição. Todas as classes de TSS analisadas apresentam alta frequência de purinas na posição de início de transcrição (0) e pirimidinas na posição -1. Nos inícios de transcrição primários (pTSS) é possível identificar também a alta frequência de nucleotídeos que compõem o start codon ATG, demonstrando que a maioria dos transcritos com TSS identificados nessa classe são leaderless. Entre as posições -26 a -37 é possível identificar o sítio de reconhecimento do fator de transcrição IIB e o TATA box.

4.1.5 Inícios de transcrição primários (pTSS)

Analisando as regiões 250 pb *upstream* ao início de genes de *H. salinarum* NRC-1 foram identificados 1978 inícios de transcrição primários (pTSS) nas quatro condições analisadas (Figura 9). Esses inícios de transcrição correspondem a 1545 genes, ou 58,8% dos genes anotados. Esse número é muito próximo dos 1518 genes diferencialmente expressos em dados de *microarray* obtidos ao longo da curva de crescimento de *H. salinarum* NRC-1, sugerindo a existência de um grupo de genes responsável pela coordenação do crescimento normal do organismo (Facciotti *et al.*, 2010). Para 79% dos genes com início de transcrição, apenas um pTSS foi identificado enquanto que 21% apresentam de 2 até 7 pTSS identificados (Tabela 8). Esse número é similar nas bactérias *E. coli* e *Klebsiella pneumoniae*, onde 26,5% e 22,8% dos genes, respectivamente, apresentam mais de um início de transcrição (Kim *et al.*, 2012); já na haloarchaea *H. volcanii*, apenas 7,5% dos genes analisados apresentaram mais de um TSS (Babski *et al.*, 2016). Apesar de alguns pTSS estarem possivelmente relacionados com a transcrição de pequenos genes não anotados nas regiões analisadas, a presença de múltiplos inícios de transcrição indica a existência de diferentes transcritos primários para um mesmo gene. O gene VNG_RS10065, por exemplo, apresenta 5 diferentes pTSS associados, sendo que o mais distante ao codon de início de tradução está localizado a 210 pb e o mais próximo, a 4 pb (Figura 13). Esse gene codifica a enzima *Can*, responsável pela isomerização do citrato em isocitrato, no ciclo do ácido cítrico. A região promotora desse gene apresenta 4 diferentes regiões regulatórias (GREs – *gene regulatory elements*) definidas através da análise da região promotora de genes corregulados em condições específicas de crescimento (Brooks *et al.*, 2014). A presença dessas diferentes regiões regulatórias em conjunto com múltiplos TSSs indica que algumas das isoformas podem ser transcritas em condições específicas do crescimento de *H. salinarum* NRC-1. A existência de diferentes inícios de transcrição indica

também a produção de isoformas de transcritos com diferentes regiões 5' UTRs. Variações nessa região podem ter importantes implicações na regulação pós-transcricional via interação com ncRNAs ou através do reconhecimento de diferentes ligantes por meio de *riboswitches* (Babski *et al.*, 2014).

Tabela 8: Número de genes com um ou mais pTSSs associados

# pTSS	# Genes
1	1226
2	245
3	46
4	20
5	5
6	2
7	1

Figura 13: Inícios de transcrição e elementos regulatórios associados a região promotora do gene VNG_RS10065. A região upstream ao gene VNG_RS10065 apresenta 5 diferentes inícios de transcrição (setas pretas) primários e 4 elementos regulatórios (GRE), indicando que o uso de diferentes regiões regulatórias podem ser responsáveis pela produção de isoformas de transcritos para uma mesma região gênica. Essas elementos foram identificadas através da análise da região promotora de genes coregulados em condições específicas de crescimento (Brooks *et al.*, 2014). A direção das setas indica a orientação do gene.

Quanto à distância em relação a região codificante, a maioria dos pTSSs estão localizados próximos ao códon de início de tradução, indicando a ausência ou pequeno tamanho de regiões 5' UTRs nos transcritos de *H. salinarum* NRC-1 (Figura 14). 775 pTSS

(39,2%) estão colocados com o códon de início de tradução, e 396 (20%) estão localizados entre 1 a 10 nucleotídeos de distância da região codificante. Os demais (40,8%) se estendem na região entre 11 e 250 nucleotídeos de distância em relação a região codificante. Esse resultado corrobora a alta frequência do códon ATG nas análises de sequência em torno da região promotora dos pTSSs (Figura 12) e valida o caráter *leaderless* da maioria dos transcritos em *H. salinarum*.

Figura 14: Distância dos TSS primários até códon de início de tradução. O eixo *x* representa a distância dos pTSSs até o primeiro start códon downstream, indicando o tamanho da região 5' UTR. A alta frequência de pTSS próximos ao start codon indica que a maioria dos transcritos analisados são *leaderless*.

Apesar da maioria dos transcritos de haloarchaea não apresentarem regiões 5' UTRs, análises experimentais em um conjunto limitado de genes de *H. volcanii* determinou que o tamanho médio da região 5' UTR nessa espécie é de 19 nucleotídeos enquanto em *H. salinarum* R1 essa região apresenta em média 33 nucleotídeos (Brenneis *et al.*, 2007). A

identificação em larga escala dos inícios de transcrição através do dRNA-seq em *H. volcanii* revisou o tamanho médio das regiões 5' UTRs para 23 nucleotídeos nesse organismo (Babski *et al.*, 2016), enquanto nossos dados mostram que as regiões 5' UTRs apresentam tamanho médio de 46 nucleotídeos em *H. salinarum* NRC-1. Essas informações sugerem a existência de um mecanismo diferenciado de regulação pós-transcricional entre essas espécies, com uma possível preponderância de mecanismos de regulação via região 5' UTRs em *H. salinarum* NRC-1, de maneira similar ao que ocorre em bactérias (Waters & Storz, 2009). Análises futuras da composição dessas regiões poderão desvendar mecanismos de regulação pós-transcricional, como o uso diferenciado de sítios de início de transcrição em alguns genes, e também a identificação de sítios de interação a ligantes, como pequenas moléculas ou ncRNAs, com efeito regulatório no mRNA (Brenneis & Soppa, 2009).

4.1.6 Genes com pTSS associados

Dos 1545 genes com pTSS associados, 985 foram encontrados na condição 17 horas, 1172 na condição 37 horas, 1139 na condição 86 horas e 1188 na condição referência de crescimento, com alto grau de sobreposição entre as condições analisadas. 639 genes tiveram inícios de transcrição identificados em todas as condições analisadas e a condição de 86 horas foi o momento do crescimento com o maior número de genes específicos, com 91 genes identificados somente nessa condição (Figura 15).

Analisando os genes associados a inícios de transcrição primários em cada um dos pontos da curva de crescimento é possível identificar 5 genes responsáveis pela produção de proteínas ribossomais na condição de 17 horas; nas condições 37 horas e 86 horas foram identificados pTSSs para 7 genes codificantes de transposases, 14 genes codificantes de proteínas relacionados a transporte de membrana, uma superóxido dismutase (VNG_RS05200), uma proteína de choque térmico, uma proteína flagelina (VNG_RS03940),

uma proteína relacionada a motilidade (*twitching motility protein PilT* – VNG_RS11145) e um gene codificante de vesícula de gás (VNG_RS11395). Esse resultado está em acordo com trabalhos anteriores que demonstraram que a fase inicial de crescimento está mais associada com a expressão de genes envolvidos com transcrição e tradução, enquanto a fase estacionária é caracterizada principalmente por genes codificadores de proteínas de transporte e genes relacionados a captação de energia através de outras fontes energéticas (Facciotti *et al.*, 2010; Lange *et al.*, 2007).

Para identificar as funções preponderantes dos genes com início de transcrição identificados em cada ponto do crescimento realizamos uma análise de enriquecimento funcional através do pacote topGO, disponível na plataforma *Bioconductor* (Alexa & Rahnenfuhrer, 2016). Essa análise permitiu a identificação de 19 ontologias enriquecidas nas diferentes condições de crescimento, sendo que algumas estão enriquecidas em mais de uma condição (Figura 16). De forma geral, na primeira fase do crescimento é possível detectar o enriquecimento de genes relacionados a ligação de magnésio e proteínas com função de aminopeptidase e oxidoreductase, já nas fases finais do crescimento é possível identificar o enriquecimento de fatores de transcrição e proteínas relacionadas a tradução, o que difere dos achados anteriores e está em desacordo com informações da literatura (Facciotti *et al.*, 2010; Lange *et al.*, 2007). Essa incongruência pode estar relacionada com o fato de que os genes analisados foram identificados através de sua associação com inícios de transcrição, sem considerar os níveis de expressão em si. Isso pode acarretar na inclusão de genes pouco expressos na análise, mas para os quais ainda foi possível identificar um início de transcrição. Com isso, funções características de pontos específicos do crescimento não se mostraram enriquecidas. Além disso, a existência de uma grande quantidade de genes hipotéticos ou sem um termo GO associado em *H. salinarum* NRC-1 pode comprometer a análise de enriquecimento funcional.

Figura 15: Número de genes com pTSS identificados nas diferentes condições analisadas

Figura 16: enriquecimento funcional dos genes com pTSS associado em cada uma das condições de crescimento avaliadas. A barra vermelha representa a porcentagem de genes de *H. salinarum* NRC-1 que pertencem ao GO em questão.

4.1.7 TSS antisense (aTSS)

Analisando os inícios de transcrição localizados na orientação oposta de regiões gênicas, identificamos um total de 737 posições classificadas de TSSs *antisense* (aTSSs) em *H. salinarum* NRC-1 (Figura 10; Anexo A). 53 destes foram classificados também como primários de genes *downstream*. Dados recentes de aTSS em *Haloferax volcanii* e *Thermococcus kodakarensis* indicam a presença de um número maior de aTSS nessas archaea, com 1244 e 1018 posições identificadas respectivamente (Jäger *et al.*, 2014; Babski *et al.*, 2016). O maior número de aTSS em *H. volcanii* pode ser explicado pelo maior número de genes anotados nessa espécie, a qual apresenta em torno de 4000 genes (Hartman *et al.*, 2010). Já em *T. kodakarensis*, apesar do número de genes ser similar ao de *H. salinarum* NRC-1, a região considerada para classificação de um aTSS incluiu uma possível região 3' UTR de até 100 pb. *downstream* aos genes, o que deve ter contribuído para o maior número de aTSS identificados (Jäger *et al.*, 2014). Além disso, o genoma de *H. salinarum* apresenta grandes duplicações de porções dos plasmídeos pNRC100 e pNRC200, diminuindo o número de genes únicos disponíveis para análise (Ng *et al.*, 2000).

Esses inícios de transcrição devem dar origem a pequenos RNAs capazes de interagir com o mRNA produzido pelo gene na fita *sense*, regulando de forma positiva ou negativa a expressão dos genes alvos, chamados de RNAs antisense (asRNAs) (Waters & Storz, 2009). A expressão diferencial dos possíveis asRNAs, cujos TSS foram mapeados e dos genes na fita oposta foi avaliada baseada em dados de *tiling array* nas condições de: i) alterações dos níveis de oxigênio (dados não publicados); ii) curva de crescimento (Koide *et al.*, 2009); iii) alteração dos níveis de salinidade (Beer *et al.*, 2014); iv) deleção de RNase (Wurtmann *et al.*, 2014); v) estresse oxidativo (Plaisier *et al.*, 2014); vi) deleção de fatores de transcrição (Turkarslan *et al.*, 2014); vii) deleção da chaperona de RNA Lsm (dados não publicados)

(Tabela 4). Analisamos todos os inícios de transcrição antisense que não possuíam sobreposição com a classe de TSS primários, totalizando 684 aTSS. Para realizar o cálculo de expressão diferencial definimos uma região 100 pb *downstream* ao aTSS como correspondente a um possível pequeno RNA antisense (asRNA). A partir disso calculamos a expressão mediana dessa região, e também do gene na fita oposta. A partir dos valores de expressão em cada uma das condições calculamos a correlação entre a expressão do asRNA e identificamos os genes e asRNAs significativamente correlacionados ($p < 0.05$). O maior número de asRNA correlacionados com o gene cognato foi identificado nos dados de expressão gênica obtidos ao longo da curva de crescimento, com 157 casos (23%), dos quais 95 estavam positivamente correlacionados e 62 estavam anti-correlacionados (Figura 17).

Para as demais condições analisadas foram identificados: 127 asRNAs correlacionados com os genes na fita oposta na linhagem que não apresenta a chaperona de RNAs Lsm; 115 na condição de variação de salinidade; 105 na condição de alteração de concentração de oxigênio; 50 na linhagem contendo a deleção de fatores de transcrição; 50 na linhagem com deleção da RNase codificada pelo gene VNG2099C e 34 na condição de crescimento sob estresse oxidativo. Alguns asRNAs apresentaram-se correlacionados com seus genes cognatos em mais de uma condição ou linhagem. Dessa forma, considerando todos os dados analisados, foram identificados 430 asRNAs correlacionados com a expressão na fita oposta, o que corresponde a 62,9% de todos os aTSS investigados (Anexo B). Esses inícios de transcrição estão localizados internamente a 322 genes distintos. O alto número de aTSS que apresentam sua região *downstream* correlacionada com a expressão do gene na fita oposta indica que muitos desses inícios de transcrição devem estar envolvidos com a produção de pequenos RNAs com funções regulatórias.

Figura 17: *Expressão de 684 asRNAs e dos genes na fita sense ao longo da curva de crescimento. Os triângulos indicam o crescimento do organismo.*

4.1.8 TSS órfãos (oTSS)

Além dos aTSSs, outra classe de inícios de transcrição que pode ser responsável pela produção de pequenos RNAs regulatórios são os oTSS, inícios de transcrição presentes em regiões intergênicas e classificados como órfãos. Além de ncRNAs, esses TSSs podem estar relacionados com o início de transcrição de genes codificantes não anotados, como as pequenas ORFs, sequências codificadoras de proteínas menores de 100 aa que usualmente são

desconsideradas em processos de anotação (Su *et al.*, 2013). Com a anotação automatizada de inícios de transcrição, foram identificados 732 inícios de transcritos órfãos durante o crescimento de *H. salinarum* NRC-1 (Anexo A; Figura 10)

Recentemente o genoma de procariotos hospedados no banco de dados *RefSeq* (NCBI – www.ncbi.nlm.nih.gov/refseq/) passou por um processo de reanotação visando eliminar redundâncias e estabelecer uma padronização no processo de anotação de genomas (O’Leary *et al.*, 2016). A linhagem NRC-1 de *H. salinarum* foi reanotada tomando por base as informações disponíveis para o genoma da linhagem R1, o que acarretou a exclusão de diversos genes. Comparando os TSSs identificados de acordo com anotação mais recente com a anotação previamente disponível, identificamos 159 oTSS localizados na região 250pb *upstream* de genes removidos na anotação atual de *H. salinarum* NRC-1. Esses oTSS representam 73 genes descontinuados, indicando que essas regiões são genes autênticos, responsáveis pela produção de transcritos codificadores (Anexo C). Além disso, 377 oTSS estão localizados *downstream* e na fita oposta a regiões gênicas, *antisense* a possíveis regiões 3’ UTRs (Anexo D). Essas regiões são conhecidos alvos para interação com pequenos RNAs responsáveis pela regulação pós-transcricional dos mRNA alvos (Babski *et al.*, 2014), dessa forma, esses oTSS podem representar o início de transcrição de RNAs não codificantes com importância regulatória.

Desconsiderando os oTSSs *upstream* a genes removidos da anotação atual de *H. salinarum* e também aqueles localizados *antisense* a regiões 3’ UTRs, temos um total de 196 inícios de transcrição intergênicos que não estão relacionados a qualquer outra anotação (Anexo E). Para investigar o perfil de expressão dos transcritos gerados a partir dos oTSSs, realizamos o cálculo de expressão em diferentes condições de crescimento, de forma similar a que foi realizada para os aTSSs. Todos os transcritos gerados a partir dos oTSSs que apresentaram uma variação de expressão de no mínimo duas vezes, em pelo menos 1/3 dos

pontos analisados em cada condição, foram definidos como diferencialmente expressos. De acordo com esses critérios foram identificados 132 RNAs diferencialmente expressos em pelo menos uma das condições analisadas, o que corresponde a 67,3% dos oTSS analisados (Anexo E). Essa informação corrobora a hipótese de que esses oTSSs devem dar origem a transcritos com importância funcional.

Analisando variações no perfil de transcrição ao longo da curva de crescimento de *H. salinarum* NRC-1, Koide *et al* 2009 identificaram 61 RNAs diferencialmente expressos, sem uma região codificante definida e que não apresentam correspondência com os dados disponíveis de proteômica, sendo, dessa forma, classificados como ncRNAs. Os dados de dRNAseq permitiram identificar o início de transcrição para 29 desses transcritos: 17 destes são precedidos por oTSSs, 4 estão relacionados a aTSSs, 2 a iTSSs e 6 a pTSSs (Anexo F). Um dos ncRNAs que está associado a um pTSS está atualmente anotado como uma proteína hipotética. Em relação aos ncRNAs que não apresentaram inícios de transcrição associados, 8 estão localizados *antisense* a sequências de inserção repetitivas e 14 estão localizados nas regiões duplicadas dos plasmídeos. Essas regiões são desconsideradas das análises uma vez que não é possível definir com precisão a origem dos *reads* produzidos a partir delas (Seção 3.6 - Sequenciamento e processamento dos dados). Para 2 ncRNAs localizados *antisense* a extremidade 5' de transposases de cópia única, existe enriquecimento de início de *reads* na biblioteca TEX + em posições próximas, e embora o número de *reads* alinhados às posições, ou o p-valor, tenham ficado próximos ao limiar definido na ferramenta TSSAR, não foram suficientes para classificação dessas posições como TSSs. Essas regiões provavelmente estão envolvidas com a produção de pequenos RNAs responsáveis pela regulação pós-transcricional das transposases, como já descrito para *E. coli* (Ma & Simons, 1990) e também para nas archaea *S. sulfataticus* (Tang *et al.*, 2005) e *Methanosarcina mazei* (Jager *et al.*, 2009). Para os demais 8 ncRNAs não houve nenhum sinal de início de transcrição.

Para verificar se alguns dos transcritos associados a oTSS poderiam gerar pequenas proteínas, utilizamos como base a anotação da linhagem R1 de *H. salinarum*. A aplicação de um protocolo específico para extração e análise de pequenas proteínas identificou 90 genes codificadores de proteínas menores que 10 kDa nessa linhagem (Klein et al, 2007). 18 destes genes não estão presentes na anotação atual da linhagem NRC-1, no entanto, buscando por sequências homólogas, utilizando a ferramenta BLAST (Altschul *et al.*, 1990), identificamos 30 pequenas ORFs (smORFs) com 100 % de similaridade a esses genes. O gene OE1231R, uma sequência de inserção pertencente a família ISH2, apresentou 13 cópias no genoma da linhagem NRC-1 (Tabela 9).

Analisando essas regiões em conjunto com os dados oriundos do dRNA-seq foi possível identificar o início de transcrição para 16 smORFs não anotadas em NRC-1, das quais 12 estão associadas a oTSS, duas são precedidas por TSSs primários e duas por TSS internos a genes *upstream* (Tabela 9).

Quatro smORFS associadas a oTSSs, homólogas aos genes OE_1789R, OE_3268F, OE_1435R, OE_4440F de *H. salinarum* R-1, codificam proteínas com o motivo CpxCG, similar a motivos dedos de zinco e associado com a regulação da transcrição gênica (Klein *et al.*, 2007) (Figura 18). Em *H. salinarum* R-1 esse motivo é usualmente encontrado em proteínas menores de 20 kDa (Klein et al, 2007). Já na linhagem NRC-1, está presente na proteína codificada pelo gene VNG_RS05710, a qual atua na regulação dos genes *bop*, codificante da bacteriorodopsina e *crtB1*, codificante de uma importante enzima da biossíntese de carotenoides (Tarasov *et al.*, 2008). Três dessas smORFs são superexpressas ao longo da curva de crescimento de *H. salinarum* NRC-1, enquanto a expressão da smORF homóloga ao gene OE_3268F é reprimida (Figura 18). Essas informações sugerem que essas regiões apresentam importantes papéis funcionais ao longo do crescimento desse micro-organismo.

Tabela 9: Pequenas proteínas de *H. salinarum* R1 não anotados no genoma de *H. salinarum* NRC-1

Gene R1	Cromossomo	Início	Fim	Orientação	TSS associado	Massa [Da]
OE1120F	chr	66054	66251	+	-	7556
OE3047F	chr	1060045	1060191	+	-	6054
OE3268F	chr	1189961	1190134	+	oTSS	6358
OE3859B1F	chr	1505330	1505425	+	iTSS	3487
OE4347F	chr	1789808	1789984	+	oTSS	6600
OE4440F	chr	1838364	1838513	+	oTSS	4125
OE4491F	chr	1864948	1865109	+	oTSS	6177
OE1435R	chr	226021	226179	-	pTSS	5869
OE1594R	chr	309533	309685	-	oTSS	5779
OE1789R	chr	409836	409991	-	oTSS	5670
OE2521R	chr	796243	796389	-	oTSS	5319
OE3009R	chr	1043476	1043619	-	oTSS	5046
OE3141R	chr	1110237	1110413	-	iTSS	6353
OE3365R	chr	1247170	1247331	-	-	5464
OE3750R	chr	1447469	1447639	-	pTSS	6666
OE5075F	plasmid_pNRC200	162134	162286	+	oTSS	5857
OE5177F	plasmid_pNRC200	219985	220194	+	oTSS	8281
OE1231R	chr	1230496	1230699	-	-	7655
OE1231R	chr	759667	759870	-	-	7655
OE1231R	chr	177093	177296	-	oTSS	7655
OE1231R	chr	771722	771925	-	oTSS	7655
OE1231R	plasmid_pNRC200	70849	71052	+	- (região repetitiva)	7655
OE1231R	plasmid_pNRC200	74812	75015	-	- (região repetitiva)	7655
OE1231R	plasmid_pNRC200	112954	113157	-	- (região repetitiva)	7655
OE1231R	plasmid_pNRC200	121887	122090	-	-	7655
OE1231R	plasmid_pNRC200	308896	309099	-	-	7655
OE1231R	plasmid_pNRC100	70849	71052	+	- (região repetitiva)	7655
OE1231R	plasmid_pNRC100	153698	153901	+	- (região repetitiva)	7655
OE1231R	plasmid_pNRC100	74812	75015	-	- (região repetitiva)	7655
OE1231R	plasmid_pNRC100	150412	150615	-	- (região repetitiva)	7655

Figura 18: Pequenas proteínas de *H. salinarum* R1 presentes na linhagem NRC-1 e associadas a inícios de transcrição. Proteínas com peptídeos identificados na linhagem R1 de *H. salinarum*, todas apresentam 100% de similaridade com o genoma de NRC-1. Análise de sequência indicam a presença de um motivo dedos de zinco CpxCG, o qual está presente em fatores de regulação da expressão gênica. O perfil de expressão gênica ao longo da curva de crescimento indica que todas são diferencialmente expressas. A região correspondente ao gene OE1435R está localizada na região 5' UTR do gene VNG_RS01125, o qual codifica a proteína de choque térmico *hsp20*. As setas indicam os inícios de transcrição identificados por meio do dRNA-seq. Seta quadriculada: 17 horas; seta cinza: condição referência; seta preta: 37 horas; seta branca: 86 horas.

4.1.9 TSS internos (iTSS)

O quarto grupo de inícios de transcrição identificado está localizado internamente e na mesma orientação de regiões codificantes, sendo denominado de inícios de transcrição internos (iTSS). A anotação automatizada de inícios de transcrição identificou um total de 1365 iTSS em *H. salinarum* NRC-1 (Figura 10 - Anexo A). Apesar desse número ser similar à quantidade de iTSS identificados recentemente na haloarchaea *H. volcanii* (1153) (Babski *et al*, 2016), é importante ressaltar que o número de genes nessa espécie é aproximadamente 1,5 vezes maior do que em *H. salinarum* NRC-1 (Hartman *et al*, 2010). O maior número de iTSS identificados através de nossos dados em *H. salinarum* NRC-1 deve estar relacionado com a

diversidade de amostras sequenciadas em *H. salinarum* NRC-1, obtidas em 4 momentos distintos da curva de crescimento (Figura 7), enquanto que em *H. volcanii* somente a fase exponencial de crescimento foi avaliada (Babski *et al.*, 2016).

Muitos desses inícios de transcrição estão localizados até 250 pb *upstream* a genes adjacentes, sendo dessa forma, classificados também como inícios de transcrição primários. Desconsiderando esses casos e também iTSS localizados em genes correspondentes a tRNAs ou rRNAs, foi possível identificar 1060 iTSS autênticos, localizados dentro de genes codificadores de proteínas.

A existência de inícios de transcrição dentro de regiões codificantes é usualmente associada a artefatos da técnica de dRNA-seq ou falhas na anotação gênica. Porém, trabalhos recentes em *Synechocystis* sp. PCC6803 (Mitschke *et al.*, 2011), *N. gonorrhoeae* (Wachter & Hill, 2015), *C. crescentus* (Schrader *et al.*, 2014) e *E. coli* (Bilusic *et al.*, 2014) têm demonstrado a existência de transcritos independentes dentro de regiões gênicas. Além disso, a análise de iTSSs em diferentes espécies do gênero *Shewanella* demonstrou que essas posições são conservadas nesse grupo, sugerindo a existência de um importante papel funcional (Shao *et al.*, 2014). Assim, utilizamos filtros para selecionar iTSS que poderiam dar origem a RNAs ainda não conhecidos em *H. salinarum*, excluindo casos relacionados a reanotação e posteriormente, casos relacionados a formação de estrutura secundária (Seção 4.1.10 - iTSS estruturados).

Os iTSS localizados próximos a extremidade 5' dos genes podem estar relacionados a erros na anotação do início da região codificante, já os iTSS localizados mais próximos a região 3' podem estar relacionados com a transcrição de genes codificantes *downstream* (Mitschke *et al.*, 2011). Em *H. salinarum* NRC-1, 1/3 dos iTSS (354) estão localizados nas regiões 10% iniciais ou finais das regiões codificantes (Figura 19) e foram então desconsiderados nas análises posteriores. 3 TSS internos as transposases VNG_RS00050,

VNG_RS00105 e VNG_RS00160 também foram desconsiderados, visto que são TSS para sotRNAs, RNAs sobrepostos a região 3' de transposases do tipo IS1341 identificados por nosso grupo de pesquisa (Gomes-Filho et al, 2015).

Para os 703 iTSS restantes localizados na região 80% central dos genes, realizamos uma curagem manual utilizando um compêndio de anotações provenientes de 7 diferentes bancos de dados (Seção 3.7 - Dados públicos analisados). Com essa análise 48 iTSS foram reclassificados como primários em relação ao próprio gene onde estavam localizados ou primários a genes *downstream* presentes em outras anotações (Figura 20). Assim, realizamos as análises posteriores baseado num conjunto de 655 iTSS, que não estão relacionados a problemas de anotação.

Figura 19: Distribuição de iTSSs ao longo da região codificante. Grande parte dos inícios de transcrição está localizada na região 10% inicial, indicando possíveis erros de anotação.

Figura 20: Curadoria de iTSS frente a anotações disponíveis para *H. salinarum* NRC-1. Os inícios de transcrição internos foram curados manualmente quanto a sua localização em anotações disponíveis em 7 diferentes bancos de dados. As setas pretas representam os inícios de transcrição identificados com o dRNA-seq. As linhas tracejadas vermelhas indicam iTSS que podem atuar como pTSS do gene onde estão localizados ou de genes downstream (piTSS). A linha tracejada verde representa um iTSS verdadeiro.

4.1.10 iTSS estruturados

Uma das explicações para a existência de inícios de transcrição internos a regiões codificantes é a presença de estruturas no mRNA que protegem os transcritos processados da degradação. No dRNA-seq, essas estruturas poderiam gerar um sinal interpretado como início de transcrição (Jäger *et al.*, 2014). Em *Thermococcus kodakarensis*, por exemplo, o transcrito produzido pelo gene *rpoL*, que codifica uma subunidade da RNA Polimerase, apresenta um *hairpin* predito em sua extremidade 3', o que gera um enriquecimento no sinal da biblioteca tratada com a enzima TEX (Jäger *et al.*, 2014). Em nossos dados de dRNA-seq também é possível identificar uma alta concentração de inícios de transcrição em genes que produzem transcritos sabidamente estruturados, como tRNAs e rRNAs.

Para identificar os iTSSs que podem estar relacionados com a pausa da degradação pela enzima TEX devido a presença de estruturas internas no mRNA, realizamos a identificação de possíveis regiões estruturadas nas regiões codificantes. Para isso, implementamos um algoritmo para realizar o cálculo do *z-score* da mínima energia livre

(MFE) da região 100 pb. e 50 pb. *downstream* aos iTSSs. O *z-score* representa o quanto determinado valor varia em relação a média de valores; nesse caso, o quanto a MFE de uma determinada sequência varia em relação a média da MFE de sequências randômicas, geradas a partir da sequência analisada (Le & Maizel, 1989) (Figura 21). Assim, se o *z-score* for mais positivo, a sequência nativa possui uma MFE maior do que as sequências randômicas e é classificada como possivelmente não estruturada. Por outro lado, se o *z-score* for mais negativo, a MFE da sequência original é menor do que a MFE das sequências permutadas e a região analisada é classificada como possivelmente estruturada.

$$Z(\mathbf{x}) := \frac{E(\mathbf{x}) - \langle X_{\text{shuffled}}(\mathbf{x}) \rangle}{\sigma(X_{\text{shuffled}}(\mathbf{x}))}$$

Figura 21 Equação para o cálculo do z-score. $\langle \rangle$ e σ representam a média e o desvio padrão, respectivamente, da MFE do subconjunto de 1000 sequências permutadas. $E(x)$ representa a MFE da sequência nativa

Como a estabilidade da ligação entre dois nucleotídeos em uma estrutura sofre influência dos pareamentos vizinhos, é importante que a comparação da mínima energia livre seja realizada entre sequências com a mesma composição de dinucleotídeos. Análises com sequências randômicas de mRNAs com a mesma composição de dinucleotídeos demonstraram que estas tendem a apresentar uma MFE menor que as sequências originais, enquanto sequências randômicas geradas da mesma maneira para RNAs estruturados apresentam uma MFE maior que as sequências originais (Workman & Krogh, 1999; Clote et al, 2005). Para calcular o *z-score* foram geradas 10001 sequências permutadas com a mesma composição de dinucleotídeos para cada sequência analisada, através do software *uShuffle* (Jiang et al., 2008). A MFE das sequências nativas foi comparada com a média de MFE de 1000 sequências permutadas escolhidas aleatoriamente no conjunto de 10001 sequências, através da equação apresentada na Figura 21.

Primeiramente realizamos o cálculo da variação de MFE para 52 RNAs estruturados de *H. salinarum* NRC-1 (47 tNRAs, 3 rRNAs, RNase P e RNA 7S), os quais apresentaram valores de *z-score* variando entre -1,08 até -10,8 (Figura 22A). A mesma análise conduzida com a região 100 pb. *downstream* aos 655 iTSS retornou valores de *z-score* entre -4,82 até 2,41 (Figura 22C), enquanto na análise da região 50 pb. *downstream* aos iTSSs esses valores variaram entre -5,14 e 2,53.

Uma vez calculado o *z-score* para as sequências relacionadas aos iTSS, resta saber se o valor identificado é significativamente diferente do *z-score* de sequências randômicas. Utilizando o mesmo conjunto de sequências permutadas geradas na análise anterior, repetimos o procedimento, dessa vez selecionando uma sequência aleatória no conjunto de sequências permutadas e calculando seu *z-score* frente a um subconjunto de outras 1000 sequências permutadas; tal procedimento foi repetido 10.000 vezes. Dessa forma, a fração de *z-scores* das sequências selecionadas aleatoriamente que são menores que o *z-score* da sequência nativa nos dá uma aproximação da probabilidade que um *z-score* ocorra randomicamente. Essa fração foi definida como valor-p do *z-score* (Freyhult et al., 2005). Aplicando essa metodologia aos 52 RNAs estruturados de *H. salinarum* NRC-1, 33 RNAs apresentaram um p-valor abaixo de 0,01 e 47 apresentaram um p-valor abaixo de 0,05 (Figura 22B). Definindo o limite de significância do p-valor em 0,05 identificamos 70 regiões possivelmente estruturadas quando analisamos a sequência de 100 pb *downstream* ao iTSS (Figura 22D), e 54 quando a região analisada foi de 50 pb. Algumas das regiões analisadas foram consideradas como estruturadas em ambos os casos, assim, ao final foram identificadas 101 posições com sinal de início de transcrição (15,5%) resultantes de estruturas presentes no mRNA. As demais 554 sequências nativas não apresentam *z-score* e p-value compatível com a formação de estrutura. Dessa forma, embora parte dos iTSS possa ter sido gerada através de estruturas internas ao mRNA que impedem sua degradação pela enzima TEX, a análise das

regiões estruturadas demonstra que a grande maioria dos iTSS analisados (554) representam inícios de transcrição genuínos e estão envolvidos com a produção de transcritos. As análises posteriores foram conduzidas nesse subconjunto de iTSS.

Figura 22: Cálculo do z-score da mínima energia livre (MFE) dos iTSSs. A) Distribuição dos z-scores de RNA estruturais de *H. salinarum* e **C)** da região 100 pb. downstream aos iTSSs. **B)** Adotando um limite de p-valor de 0,05 apenas 4 tRNAs e a RNase P são classificados como não estruturados (pontos pretos). **D)** Esse mesmo limite aplicado as regiões downstream dos iTSS classifica 70 deles como estruturados (pontos vermelhos).

4.1.11 Análise de enriquecimento funcional

Para identificar as funções preponderantes associadas aos genes que apresentam iTSS na porção central da sua região codificante, realizamos uma análise de enriquecimento funcional através dos termos *Gene Ontology* (GO). 408 genes apresentam pelo menos um início de transcrição interno nas condições descritas. Destes, 287 apresentam pelo menos uma ontologia associada. Para identificação dos termos enriquecidos utilizamos o pacote topGO (Alexa *et al.*, 2006). No total, quatro termos apresentaram-se enriquecidos ($p < 0,01$), sendo que dois estão relacionadas a atividade de oxidoreductase atuando em grupos CH-NH₂ (GO:0016638; GO:0016639), no qual três genes são glutamato desidrogenases (VNG_RS00665, VNG_RS02485 e VNG_RS04705) e duas são glicina desidrogenases (VNG_RS06210 e VNG_RS06215). O terceiro GO enriquecido está relacionado a atividade de fator de alongação da tradução (GO:0003746) e o quarto está relacionado a hidrólise de ligações éster formados por ácido sulfúrico GO:0008484 (Tabela 10).

Tabela 10: Termos GO enriquecidos nos genes que apresentam iTSS

GO	Termo	Anotados	Significantes	Esperado	p-valor
GO:0016638	oxidoreductase activity, acting on the CH-NH ₂ group of donors	8	6	1,55	0,00098
GO:0016639	oxidoreductase activity, acting on the CH-NH ₂ group of donors, NAD or NADP as acceptor	4	4	0,77	0,00137
GO:0003746	translation elongation factor activity	7	5	1,35	0,00389
GO:0008484	sulfuric ester hydrolase activity	3	3	0,58	0,00716
GO:0022804	active transmembrane transporter activity	55	18	10,64	0,01125
GO:0004672	protein kinase activity	18	8	3,48	0,01271
GO:0036094	small molecule binding	321	76	62,09	0,01453
GO:0001882	nucleoside binding	231	56	44,68	0,0243
GO:0032549	ribonucleoside binding	231	56	44,68	0,0243
GO:0022892	substrate-specific transporter activity	89	25	17,21	0,02446

4.1.12 Fatores de transcrição associados a iTSS

A análise da frequência de nucleotídeos da região promotora das quatro classes de TSSs demonstrou que os inícios de transcrição internos, assim como as demais classes,

apresentaram uma alta frequência de pirimidinas na posição -1 e purinas no TSS, além de um sítio BRE na posição -37. No entanto, a conservação da região promotora dos iTSSs é bem mais baixa quando comparada as demais classes de início de transcrição, e a região correspondente ao sítio TATA é praticamente ausente (Figura 12). Para identificar os iTSS associados a sítios de ligação de fatores de transcrição realizamos uma análise comparativa dos dados provenientes do dRNA-seq com dados públicos sobre sítios de interação com fatores de transcrição proveniente de experimentos de ChIP-Chip (Koide *et al.*, 2009; Facciotti *et al.*, 2007; Plaisier *et al.*, 2014). 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. Considerando uma região de 50 pb. *upstream* aos inícios de transcrição, identificamos um total de 147 iTSS (26,5 %) associados a fatores de transcrição. No entanto, como as posições de interação com os fatores de transcrição são obtidas a partir de estimativas de dados de *microarray*, a posição exata de interação pode variar, em média, 50 nucleotídeos (Koide *et al.*, 2009). Dessa forma, realizamos a busca estendendo a região para 100 pb, sendo possível identificar 214 iTSS (38,6 %) associados a fatores de transcrição. Analisando a frequência de nucleotídeos na região *upstream* dos iTSS colocalizados com sítios de interação de fatores de transcrição, podemos identificar um aumento na frequência de nucleotídeos pirimidinas e purinas nas posições -1 e 0 respectivamente e também de um aumento na frequência de nucleotídeos que compõem o sítio BRE (Figura 23).

Figura 23: Frequência de nucleotídeos na região promotora de iTSS colocados com sítios de ligação de fatores de transcrição. Em comparação com a mesma análise conduzida com o conjunto total de iTSS (Figura 1) é possível perceber um aumento da frequência de nucleotídeos em torno do início de transcrição e também na região referente ao sítio BRE.

4.2 Transcritos intragênicos (intraRNAs)

4.2.1 iTSS dão origem a intraRNAs

Analisando cada uma das hipóteses que pudessem fornecer explicações para a existência de inícios de transcrição internamente a regiões codificantes encontramos um subconjunto de 554 iTSS (40,5%) que não possuem relação com nenhum dos elementos anteriormente citados.

Para verificarmos se os iTSS poderiam estar envolvidos com a produção de transcritos intragênicos (intraRNAs) foram realizados experimentos de *Northern blot* com sondas desenhadas *downstream* aos inícios de transcrição internos. Para identificação dos intraRNAs foram selecionados iTSSs presentes nos genes VNG1339C (atual VNG_RS05220), VNG0045C (atual VNG_RS00175), VNG0060G e VNG1654G. Os genes VNG0060G e VNG1654G foram descontinuados na anotação mais recente de *H. salinarum* NRC-1. Esses experimentos foram conduzidos em colaboração com a Dr^a Livia Soares Zaramela.

Dos quatro alvos testados somente o gene VNG_RS05220, codificante de uma proteína hipotética com alta similaridade com a enzima Acyl-CoA sintetase, apresentou um sinal de tamanho aproximado de ~ 270 pb (Figura 24). O tamanho desse transcrito

corresponde a distância entre o segundo iTSS do gene VNG_RS05220 e o término da região codificante, indicando que o intraRNA produzido equivale a porção final do gene. Essa região codifica o domínio proteico PF13193, um pequeno domínio encontrado na região C terminal de sintetases dependentes de AMP (<http://pfam.xfam.org/family/PF13193.4>) (Figura 24A). O domínio é encontrado também em proteínas codificadas por outros 4 genes de *H. salinarum* NRC-1. Analisando os inícios de transcrição identificados nos demais genes é possível detectar a presença de um iTSS na mesma região no gene VNG_RS00670 (*alkK*), o qual também codifica para uma Acyl-CoA sintetase (Figura 25).

Figura 24: Northern blot do gene VNG_RS05220 e seu intraRNA. A) O gene apresenta um pTSS e dois iTSS, localizados a 580 e 269 pb de distância do stop codon (setas pretas). A sonda utilizada interage com uma região downstream ao segundo iTSS, permitindo a detecção de possíveis transcritos expressos a partir dos três inícios de transcrição. B) No experimento de Northern blot é possível identificar as bandas correspondentes ao transcritos original, de aproximadamente 1566 nt, e um intraRNA de ~270 nt correspondente a porção final do gene (*). O experimento foi conduzido com réplicas biológicas (VNG_RS05220-1 e 2).

Para avaliar se esse início de transcrição está presente também em genes homólogos na espécie *H. volcanii* analisamos os dados de dRNA-seq recentemente publicados para esse organismo (Babski *et al.*, 2016). O pré-processamento dos dados de sequenciamento e anotação de inícios de transcrição foi realizado de maneira similar aos nossos dados (Seção 3.6 - Sequenciamento e processamento dos dados). *H. volcanii* DS2 possui nove genes codificantes de proteínas que apresentam o domínio PF13193. Três desses genes apresentam iTSS localizados próximos ao início da região associado ao domínio PF13193 (Figura 25). HVO_RS08985 apresenta um início de transcrição localizado 314 pb *upstream* ao *stop* códon do gene, HVO_RS02560 apresenta um iTSS 320 pb *upstream* do término da região codificante e HVO_RS11320 tem um de seus iTSSs localizado a 281 pb *upstream* ao *stop* códon do gene (Figura 25). Assim como visto em bactérias do gênero *Shewanella*, onde 30% dos iTSSs estavam conversados em outras espécies do gênero (Shao *et al.*, 2014), a identificação de iTSS associados as regiões codificantes do domínio PF13193 em diferentes genes, e também em outra espécie, indica que o intraRNA produzido a partir desse iTSS pode apresentar uma funcionalidade.

Figura 25: iTSS associado ao domínio proteico PF13139. Dois genes de *H. salinarum* NRC-1 e três de *H. volcanii* DS2 apresentam iTSS (setas pretas) próximo ao início da região codificante do domínio proteico PF13139 (azul). A direção das setas indica a orientação dos genes analisados. Caixa rosa: PF00501; Caixa amarela: PF16177. A barra preta no canto superior esquerdo corresponde a 100 nucleotídeos.

4.2.2 intraRNAs diferencialmente expressos

Para compreendermos o comportamento transcricional dos intraRNAs de *H. salinarum* NRC-1 realizamos uma análise comparativa do perfil de expressão desses RNAs em relação ao seu gene cognato, buscando identificar todos aqueles que apresentassem um perfil diferenciado de expressão em relação ao gene onde estão localizados. A análise foi conduzida com o mesmo conjunto de dados utilizados para analisar o perfil de expressão dos RNAs gerados a partir dos aTSSs e oTSSs (Seção 3.7 - Dados públicos analisados). A identificação dos intraRNAs diferencialmente expressos em relação ao seu gene cognato foi realizada de forma automática por meio de um *script* implementado na linguagem de programação R. Os intraRNAs foram definidos como a região entre um iTSS e o *stop* codon do gene; a partir disso foi calculada a mediana da expressão gênica do intraRNA e também da região do gene sem sobreposição com o intraRNA. O intraRNA foi classificado como diferencialmente expresso se: (i) a região do intraRNA apresentasse uma diferença de expressão maior do que 2 vezes em relação a região exclusiva do gene cognato, em pelo menos metade das condições analisadas ou (ii) a região do intraRNA apresentasse uma diferença de expressão de 3 vezes, em pelo menos 1/3 das condições.

Utilizando essa abordagem para analisar os dados de expressão gênica provenientes de 8 condições ou linhagens distintas pudemos identificar um total de 117 intraRNAs diferencialmente expressos (21,1 %) (Anexo G). 18 apresentaram uma expressão diferencial em mais de uma condição e dois foram diferencialmente expressos em relação ao seu gene cognato em 3 condições. 89 genes apresentaram pelo menos um intraRNA diferencialmente expresso. Esse resultado indica que parte dos intraRNAs sofre uma regulação diferencial em relação aos genes onde estão localizados, demonstrando um possível papel funcional na fisiologia de *H. salinarum* NRC-1.

Um exemplo de *intraRNA* diferencialmente expresso em relação ao gene cognato está localizado no gene VNG_RS03660 (Figura 26). O iTSS está colocalizado com o início da região de mudança no perfil de expressão ao longo da curva de crescimento, sendo que a porção anterior do gene tem sua expressão diminuída em relação ao primeiro ponto do crescimento enquanto a região correspondente ao *intraRNA* tem sua expressão aumentada. Outro ponto importante é a associação entre o iTSS e sítios de ligação a fatores de transcrição, sendo que a região de 100 pb *upstream* ao início de transcrição interage com os fatores de transcrição tfbG, tfbA e tfbF. Esse gene codifica a proteína *acs3*, uma acetil-CoA sintase que pertence ao mesmo grupo de enzimas codificadas pelo gene VNG_RS05220, no qual foi demonstrada experimentalmente a existência de um *intraRNA* (Figura 24). No entanto, o domínio proteico codificado na região do *intraRNA* do gene VNG_RS03660, PF13549, é diferente do domínio encontrado no gene VNG_RS05220, sendo associado a atividade de ligação amino-carboxilato dependente de ATP.

Figura 26: *intraRNA* diferencialmente expresso em relação ao gene VNR_RS03660. O iTSS (seta preta) do gene VNR_RS03660 (barra cinza) está localizado na mesma região onde ocorre uma mudança no perfil de expressão, indicando a existência de um transcrito regulado de forma diferencial. Os triângulos verdes representam sítios de interação com fatores de transcrição. A direção da seta indicação a orientação do gene. As caixas laranja, verde e lilás indicam a região codificante dos domínios PF13380, PF13607 E PF13549, respectivamente.

4.2.3 intraRNAs estão associados a domínios proteicos e são conservados em procariotos

Algumas evidências sugerem a associação de iTSS a regiões codificadoras de domínios proteicos específicos. Entre elas podemos destacar a produção de um intraRNA no gene VNG_RS05520 na mesma região codificante do domínio proteico PF13139 e sua conservação em outro gene de *H. salinarum* NRC-1 e também em três genes da archaea *Halofeax volcanii* (Figura 25). Além disso, a região diferencialmente expressa do gene VNR_RS03660 é precedida por um iTSS e correspondente a região codificante do domínio PF13549 (Figura 26).

Para analisarmos a associação entre iTSS e domínios proteicos, realizamos uma busca de todos os inícios de transcrição que estivessem em uma região até 250 pb *upstream* a um domínio conservado. Para isso utilizamos a anotação de domínios proteicos para *H. salinarum* NRC-1 realizada através do software *Sifter-T* (Almeida-E-Silva & Vêncio, 2015) e baseada na versão 28.0 do banco de dados Pfam (Punta *et al.*, 2012). Essa anotação foi realizada em colaboração com o doutorando Danilo Silva, sob orientação do Prof. Dr. Ricardo Z. N. Vêncio.

H. salinarum NRC-1 apresenta 2453 domínios proteicos conservados, localizados em 1730 genes codificantes. Entre os 554 iTSSs analisados, 86 estão localizados *upstream* aos domínios proteicos, indicando uma possível relação entre o início de transcrição e essas regiões. No total, foram identificados 66 domínios pFAM únicos associados aos iTSSs. 16 desses domínios apresentam função de interação com ácidos nucleicos, atuando no reconhecimento de regiões específicas do DNA ou RNA, 13 domínios apresentam a função de interação com ligantes (NAD, nucleotídeos, ferro, zinco, cobre) e desempenham papel regulatório ou de transmissão de um sinal através da ligação com pequenas moléculas

(Anantharaman *et al.*, 2001). Os demais domínios estão envolvidos em diversas funções, tais como: ligação a proteínas, protease, metiltransferase, transporte intracelular, domínios catalíticos e também domínios de função desconhecida (Anexo H).

Analisando em conjuntos as informações de (i) intraRNAs com sobreposição a regiões relacionadas a domínios proteicos, (ii) intraRNAs diferencialmente expressos e (iii) intraRNAs associados a sítios de ligação de fatores de transcrição, identificamos um subconjunto de 10 intraRNAs que apresentam todas essas características (Tabela 11). Além do exemplo do intraRNA do gene VNG_RS03660 citado anteriormente (Figura 26), outros dois intraRNAs que chamam a atenção estão localizados nos genes VNG_RS00420 e VNG_RS06165 (Figura 27).

Tabela 11: intraRNAs colocados com domínios pFAM, associados a sítios de interação a fatores de transcrição e diferencialmente expressos

Nome	Gene	Cromossomo	Início	Orientação	Domínio pFAM	Fator de transcrição	Condição Dif. Expresso
TSS-166	VNG_RS00420	chr	90004	forward	PF01420 – Metilase	tfbD; trh4; tfbA; tbpB	GSE31308
TSS-713	VNG_RS03660	chr	717906	forward	PF13549 – Ligação a ATP	tfbG; tfbA; tfbF	GSE12923
TSS-846	VNG_RS04150	chr	812824	forward	PF05050 – Metiltransferase	tfbF; tfbC	GSE53544; GSE112133
TSS-1488	VNG_RS08330	chr	1580440	forward	PF08241 – Metiltransferase	tfbA; tfbD; tfbC; tfbG	Δ lsm
TSS-2304	VNG_RS01270	chr	251687	reverse	PF13412 – Ligação ao DNA	tfbB; tfbD; tfbC; tfbG	GSE112133
TSS-2490	VNG_RS02445	chr	470203	reverse	PF02080 – Ligação a NAD	tfbA	GSE61975
TSS-2618	VNG_RS03395	chr	657231	reverse	PF00382 – Ligação à proteínas	trh4; tfbF; tfbD	Δ lsm; GSE15788; GSE112133
TSS-2894	VNG_RS04770	chr	918779	reverse	PF00571 – Ligação a moléculas com grupo adenosil	tfbF; tfbD	Δ lsm; GSE12923; GSE112133
TSS-3095	VNG_RS06165	chr	1183688	reverse	PF07510 – Ligação ao DNA	tfbD	GSE53544; GSE112133
TSS-3513	VNG_RS08560	chr	1627823	reverse	PF00127 – Ligação a cobre	tfbF	Δ lsm

O gene VNG_RS00420 codifica a proteína *rmeS*, a subunidade S de uma enzima de restrição do Tipo I. O domínio associado a esse intraRNA (PF01420) possui função de metilação de regiões do genoma do próprio organismo, prevenindo a clivagem por enzimas de restrição (Janscak & Bickle, 1998). O intraRNA apresenta um perfil diferenciado de expressão em relação ao gene cognato, aumentando seus níveis nas condições de superexpressão de um fator de transcrição sintético (Turkarslan *et al.*, 2014), enquanto o gene permanece inalterado. Já o gene VNG_RS06165 codifica uma proteína hipotética. Contudo, o domínio produzido na região correspondente ao intraRNA (PF07510) contém um motivo conservado HxxP, normalmente encontrando em famílias de proteínas pertencentes a superfamília de endonucleases His-Metal e associado ao reconhecimento e ligação ao DNA. A região correspondente ao intraRNA é menos expressa do que o restante do gene em condições de crescimento com baixos níveis de NaCl (Beer *et al.*, 2014).

Figura 27: intraRNAs diferencialmente expressos, associados a sítios de interação de fatores de transcrição e com sobreposição a domínios proteicos. Os inícios de transcrição são indicados pela posição das setas, a direção indica a orientação do gene. O início de transcrição associado ao intraRNA diferencialmente expresso é indicado pela estrela, a caixa pontilhada representa a região correspondente ao intraRNA. Os triângulos verdes representam os sítios de interação com fatores de transcrição detectados através de experimentos de Chip-ChIP (Koide *et al.*, 2009). A expressão diferencial do intraRNA do gene VNG_RS00420 foi identificada na condição de superexpressão de um fator de transcrição sintético (Turkarslan *et al.*, 2014), enquanto o intraRNA do gene VNG_RS06165 foi diferencialmente expresso no crescimento em baixos níveis de salinidade (Beer *et al.*, 2014).

Um dos argumentos a favor da ausência de funcionalidade de transcritos gerados de forma pervasiva pelo genoma é o seu baixo nível de conservação em diferentes espécies (Wade & Grainger, 2014). A análise de inícios de transcrição internos às regiões codificantes de 8 espécies de bactérias do gênero *Shewanella* demonstrou que 30% dos iTSS analisados eram conservados em outras espécies, indicando que essas posições podem apresentar funcionalidade (Shao *et al.*, 2014).

Para verificar a conservação dos iTSS de *H. salinarum* NRC-1 associados aos domínios proteicos buscamos os inícios de transcrição que apresentassem a mesma associação em outros organismos com dados de dRNA-seq disponíveis. Os dados de sequenciamento das bibliotecas de dRNA-seq para cada organismo foram pré-processados como descrito anteriormente e os inícios de transcrição anotados com a ferramenta TSSAR (Seção 3.6 - Sequenciamento e processamento dos dados). Nessa análise utilizamos informações provenientes das bactérias *Campylobacter jejuni* RM1221 (Dugar *et al.*, 2013), *Helicobacter pylori* (Sharma *et al.*, 2010), *Neisseria lactamica* (Zhang *et al.*, 2013) e *Shewanella oneidensis* MR-1 (Shao *et al.*, 2014), além das archaea *Thermococcus kodakarensis* KOD1 (Jäger *et al.*, 2014) e *Haloferax volcanii* (Babski *et al.*, 2016) (Seção 3.7 - Dados públicos analisados). Após a identificação dos inícios de transcrição, realizamos uma busca pelos genes que apresentassem algum dos 66 domínios proteicos que estavam associados a iTSSs em *H. salinarum* NRC-1, utilizando como referência a anotação disponível no banco de dados UniProt (Bateman *et al.*, 2015). A partir dos genes identificados buscamos pelos iTSS desses organismos que estivessem na região até 250 pb *upstream* aos domínios em questão. Todas as espécies apresentaram pelo menos um iTSS associado a um domínio proteico de maneira similar a *H. salinarum* NRC-1, sendo possível identificar 36 genes com essa característica nos diferentes organismos analisados. No total foram identificados intraRNAs associados a 19

domínios proteicos presentes também em outras espécies, sendo que em 5 casos a conservação pode ser identificada em mais de uma espécie (Anexo I).

Dois exemplos de intraRNAs associado a regiões codificantes de domínio proteicos específicos estão localizados no gene VNG_RS10410, que codifica a proteína *FusA*, o fator de alongação EF-2 (EF-G), responsável pela translocação do tRNA e mRNA no ribossomo (Andersen *et al.*, 2003) (Figura 28A). Um dos iTSS desse gene está localizado *upstream* ao domínio PF00679, correspondente a região C-terminal do fator de alongação 2. No fator de alongação G, homólogo ao fator EF-2, esse domínio está relacionado com a interação do fator com o ribossomo e é responsável por sua alteração conformacional, permitindo a translocação do tRNA e subsequente liberação do fator (Savelsbergh *et al.*, 2000). Essa mesma associação de iTSS *upstream* ao domínio PF00679 é encontrado no gene CJE_RS02715 de *C. jejuni* RM1221 e nos genes SO_0228 e SO_0842 de *S. oneidensis* MR-1, todos codificadores do fator de alongação G. O segundo iTSS conservado do gene VNG_RS10410 está localizado *upstream* ao domínio PF03144 e demais domínios *downstream*. Apesar dessa associação entre o iTSS e o domínio PF03144 ser encontrada em genes de 4 espécies distintas, existe uma grande variação na região C-terminal desses genes. Em *H. volcanii* DS2 o gene HVO_RS06420, o qual apresenta esse iTSS, codifica uma proteína similar proteína *fusA* de *H. salinarum* NRC-1. Já em *T. kodakarensis* - KOD1 o gene TK_RS01520 codifica o fator de alongação 1 – EF1A, o qual apresenta somente três domínios proteicos, sendo que somente os dois primeiros são encontrados também no gene de *H. salinarum* NRC-1. Já o gene NLA_14640 da bactéria *N. lactamica* 020-06, codifica a proteína *lepA*, um fator de alongação (EF4) responsável pela translocação reversa do tRNA após uma translocação errônea (Qin *et al.*, 2006).

Outro grupo de intraRNAs está relacionado ao domínio CBS (cystathionine- β -synthase) – PF00571, o qual é evolutivamente conservado e presente em uma ampla

variedade de proteínas em diferentes organismos (Ignoul *et al.*, 2011). Usualmente ocorrendo na forma de repetições, esse domínio interage com ligantes contendo adenosina (ATP, AMP ou S-Adenosilmetionina) e sua função está relacionada com a regulação das proteínas onde estão presentes através da indução de alterações conformacionais após o reconhecimento de tais ligantes (Ignoul *et al.*, 2011). Em *H. salinarum* NRC-1 esse domínio é encontrado em 12 genes, sendo que em 3 é possível identificar um iTSS localizado *upstream* (Figura 28B) Os genes VNG_RS02575 (*imd1*) e VNG_RS06445, dois possíveis reguladores da transcrição, apresentam unicamente esse domínio, enquanto no gene VNG_RS09025 ele está associado ao domínio transmembrana PF01595 (DUF21) na extremidade N-terminal e ao domínio PF03471, relacionado ao transporte de íons, na extremidade C-terminal. Essa arquitetura é encontrada também no gene SO_2815 de *S. oneidensis* MR-1, embora nesse caso o iTSS esteja localizado mais próximo ao primeiro domínio CBS. Já em *T. kodakarensis* KOD1 o gene TK_RS08115 apresenta quatro repetições desse domínio, e iTSSs podem ser encontrados na região *upstream* à terceira e quarta repetições.

Figura 28: intraRNAs associados a domínios pFAM e conservados em outras espécies de archaea e bactérias. A) O gene VNG_RS10410 de *H. salinarum* NRC-1 codifica a proteína fusA, a qual apresenta 5 domínios conservados (caixas coloridas). Internamente a região codificante do domínio PF03764 existe um início de transcrição (seta preta), o qual está upstream a região codificante do domínio PF00679. Esse iTSS é encontrado também em um gene de *C. jejuni* RM1221 e dois genes de *S. oneidensis* MR-1 (tracejado laranja). O mesmo gene apresenta um iTSS na região codificante do domínio PF00009 e upstream ao domínio PF03144, uma arquitetura similar de domínios e inícios de transcrição é encontrada nas espécies *H. volcanii* DS2, *T. kodakarensis* KOD1 e *N. lactamica* 020-06 (tracejado azul claro). **B)** Essa conservação é encontrada também nos genes VNG_RS02575, VNG_RS06445 e VNG_RS09025 de *H. salinarum* NRC-1, onde um início de transcrição é encontrado upstream ao domínio PF03471 (caixa azul). A mesma disposição de inícios de transcrição é encontrada nos genes SO_2815 de *S. oneidensis* MR-1 e TK_RS08115 de *T. kodakarensis* KOD1. A direção das setas indica a orientação do gene. A barra no canto inferior direito representa 100 pb.

4.2.4 RNAs intragênicos produzem proteínas alternativas

Grande parte dos trabalhos envolvendo a análise de RNAs intragênicos destaca seu caráter não codificante e com função regulatória através da interação com transcritos alvo (Bilusic *et al.*, 2014; Mitschke *et al.*, 2011; Chao *et al.*, 2012). Por outro lado, a análise integrada de sítios de início de transcrição e regiões ativamente traduzidas da bactéria *Caulobacter crescentus* levou a identificação de proteínas traduzidas a partir de transcritos originários de regiões internas aos genes CCNA_00823 e *FtsW* nesse organismo (Schrader *et al.*, 2014). Para o primeiro gene foi possível identificar a existência de um sítio de interação dos ribossomos na mesma região com sinal de início de transcrição interno à região codificante. Nessa região também foram identificados peptídeos produzidos em dois *frames* de leitura distintos, sugerindo a produção de uma proteína alternativa a partir de uma troca do *frame* de leitura no transcrito intragênico. Para o gene *FtsW* foi detectado também um enriquecimento da ligação de ribossomos na porção final do gene, além de um sítio de interação com o fator de transcrição σ^{70} 35 pb *upstream* a região de interação, fornecendo indícios da produção de uma proteína alternativa. A superexpressão dessa região produziu alterações no crescimento da bactéria, demonstrando seu papel funcional (Schrader *et al.*, 2014).

Diferentes evidências encontradas em nossos dados também apontam para o potencial codificante de alguns intraRNAs de *H. salinarum* NRC-1. Entre elas podemos destacar a produção de um transcrito a partir de um dos iTSSs do gene VNG_RS05220, com tamanho correspondendo a distância entre o iTSS e o códon de parada de tradução do gene (Figura 24), e a associação de alguns iTSSs com domínios proteicos específicos e sua conservação em outras espécies de bactérias e archaea (Figura 28). Além disso, analisando em conjunto os iTSSs e os códons de início de tradução, é possível identificar 43 inícios de transcrição

internos colocalizados com um *start* codon na mesma fase de leitura do gene. Considerando uma região de 10 pb, o número de iTSSs nessa condição sobe para 188. Analisando as três fases de leitura é possível identificar 60 iTSSs na mesma posição de um *start* codon, e 293 com um início de tradução até 10 pb de distância.

Para verificar o potencial codificante dos intraRNAs realizamos experimentos de *Western Blot* com linhagens de *H. salinarum* NRC-1 transformadas para adição um TAG Cromossomal na extremidade 3' das regiões codificantes alvo. Essa construção permite a detecção da proteína original e também de uma possível versão alternativa, produzida a partir da tradução do intraRNA no mesmo *frame* de leitura do gene (Seção 3.9 - TAG Cromossomal). Os intraRNAs escolhidos para avaliação do potencial codificante estão localizados nos genes VNG_RS07995, VNGRS05220 e o gene VNG_RS11810. O intraRNA do gene VNG_RS05220 foi experimentalmente validado através de *Northern blot* (Figura 24).

- **VNG_RS07995**

O gene VNG_RS07995 codifica uma proteína de membrana transportadora de sódio e prótons em um sistema antiporter. Ele apresenta três iTSS, os quais estão localizados a uma distância de 354, 504 e 663 pares de base *upstream* ao códon de parada de tradução. O resultado do *Western blot* demonstra a existência de uma proteína de aproximadamente 60 kDa, o que está de acordo com o peso molecular predito para a proteína total. Além disso a existência de uma banda abaixo de 35 kDa evidencia a produção de uma proteína que equivale a tradução de um intraRNA transcrito a partir do iTSS localizado a 663 pares de base do *stop* codon do gene e na mesma posição de um códon de início de tradução (Figura 29A). O peso molecular predito para a proteína traduzida a partir do intraRNA é aproximadamente 10 kDa menor do que o identificado no *Western blot*, no entanto, diferentes trabalhos têm

demonstrado que proteínas ácidas tendem a ser detectadas em pesos moleculares maiores do que o predito quando analisadas via *Western blot* (Whitfield *et al.*, 1995; Matsumoto-Taniura *et al.*, 1996; Guan *et al.*, 2015). Isso poderia explicar a discrepância observada, uma vez que a proteína produzida a partir do intraRNA apresenta um pI de 4,11 (característica comum das proteínas de *H. salinarum*). Além disso, a adição do TAG, o qual apresenta ~1 kDa, e modificações pós-traducionais devem introduzir pequenas variações no peso molecular identificado.

A versão menor da proteína detectada apresenta os domínios proteicos trkA-N (PF02254) e trkA-C (PF02080), os quais são estruturalmente distintos e em conjunto constituem o domínio trkA. O domínio trkA, junto ao domínio CBS anteriormente discutido (Figura 28B), faz parte de um grupo de domínios denominado domínios de ligação a pequenas moléculas (*small-molecule-binding domains* – SMBDs). Esses domínios estão presentes em archaea, bactérias e eucariotos e desempenham papel de regulação do funcionamento das proteínas onde são encontrados, através da interação com seus ligantes (Anantharaman *et al.*, 2001). Em *E. coli* o domínio trkA se liga a NAD e regula o transporte de potássio (Schlosser *et al.*, 1993). No gene VNG_RS07995 é provável que esse domínio seja responsável por regular a entrada de sódio de maneira similar. Apesar de o domínio ser composto pelas porções trkA-N e C, elas podem ser encontradas de forma separada em determinadas proteínas. Em *Methanobacterium thermoautotrophicum*, a proteína MTH1402 e seus ortólogos em outras bactérias apresentam a porção trkA-N degenerada, enquanto a porção trkA-C é altamente conservada. Em outros casos, a situação é inversa, e a porção trkA-C pode ser altamente degenerada, como em proteínas de efluxo de potássio em bactérias, archaea e plantas (Anantharaman *et al.*, 2001). Essa característica, aliada as diferenças estruturais, indica que essas duas regiões, apesar de normalmente serem encontradas em conjunto, devem

apresentar funções distintas e podem atuar de maneira independente (Anantharaman *et al.*, 2001).

Além do gene VNG_RS07995, o domínio *trkA* é encontrado em 8 genes de *H. salinarum* NRC-1. Em 7 esse é o único domínio presente, sendo denominadas de proteínas *trkA* (Tabela 12). Em outros três genes, o domínio *trkA*-N está presente em conjunto com outros domínios proteicos, enquanto *trkA*-C aparece associado a outros domínios em 3 genes. Na proteína codificada pelo gene *trh2* (VNG_RS05030) o domínio o *trkA*-C está fusionado a um domínio de ligação ao DNA (PF13412) constituindo uma proteína reguladora da transcrição.

Além do iTSS que dá origem ao intraRNA traduzido e detectado via *Western blot*, o gene VNG_RS07995 apresenta outros dois iTSS *upstream* ao domínio *trkA*-C. Uma disposição similar de iTSS pode ser encontrada também nos genes VNG_RS07475, VNG_RS02445, VNG_RS05015 e VNG_RS11170, além disso, a análise comparativa com os dados de dRNA-seq identificou iTSS *upstream* a esse domínio também nos genes TK1575 e TK1595 de *T. kodakarensis* (Anexo I), indicando que esse domínio também pode estar relacionado com intraRNAs em outros genes de *H. salinarum* NRC-1 e também em outras espécies. Apesar disso, os resultados do *Western blot* não demonstraram a existência de uma proteína correspondente a esse domínio específico.

Tabela 12: Genes codificante de proteínas com o domínio *trkA*

Gene	Nome	Domínios	Cromossomo	iTSS
VNG_RS00720	trkA1	PF02254 / PF02080	chr	-
VNG_RS11170	trkA2	PF02254 / PF02080	plasmid_pNRC200	<i>upstream</i> PF02080
VNG_RS11310	trkA3	PF02254 / PF02080	plasmid_pNRC200	-
VNG_RS11415	trkA4	PF02254 / PF02080	plasmid_pNRC200	-
VNG_RS05015	trkA5	PF02254 / PF02080	chr	<i>upstream</i> PF02080
VNG_RS07475	trkA6	PF02254 / PF02080	chr	<i>upstream</i> PF02080
VNG_RS11330	trkA (7??)	PF02254 / PF02080	plasmid_pNRC200	-
VNG_RS07995	-	PF00999 / PF02254 / PF02080	chr	<i>upstream</i> PF02254 <i>upstream</i> PF02080
VNG_RS02445	-	PF07885 / PF02254 / PF02080 PF02254 / PF02080	chr	<i>upstream</i> PF02080
VNG_RS03825	-	PF02254 / PF01368	chr	-
VNG_RS02570	-	PF02254 / PF01368 / PF02272	chr	-
VNG_RS07170	pchA	PF07885 / PF02254	chr	-
VNG_RS05030	trh2	PF13412 / PF02080	chr	-
VNG_RS08150	pchB	PF01895 / PF02080 / PF01895 PF02080	chr	-
VNG_RS10175	arsB	PF03600 / PF02080 / PF02080	chr	-

- **VNG_RS05220**

O gene VNG_RS05220 codifica uma proteína hipotética com alta similaridade com a enzima Acyl-CoA sintetase e apresenta dois iTSS, sendo que o início de transcrição localizado a 269 pb do *stop* codon do gene dá origem a um intraRNA de tamanho equivalente, detectado em experimento de *Northen blot* (Figura 24). O peso molecular predito para a proteína codificada por esse gene é de ~56,5 kDa, sendo que o resultado do *Western blot* apresenta uma banda entre os marcadores de ~70 e ~55 kDa, confirmando a presença da proteína original do gene VNG_RS05220. Além disso, é possível detectar uma proteína abaixo do marcador de ~15 kDa, o que corrobora a existência de uma proteína alternativa produzida a partir da tradução do intraRNA, iniciando a 269 pb de distância do codon de parada do gene (Figura 29B). O possível sítio de início de tradução do intraRNA é um códon alternativo

(TTG) e está localizado a 18 pb *downstream* ao iTSS. A discrepância entre o peso molecular predito e identificado também deve estar relacionado com a presença de aminoácidos ácidos na proteína alternativa, uma vez que ela apresenta pI de 4,82.

Com discutido anteriormente a região correspondente a proteína traduzida a partir do intraRNA apresenta um domínio proteico PF13193 o qual corresponde a um pequeno domínio na região C-terminal de Acyl-CoA sintetases. A região de ligação entre os domínios N-terminal e C-terminal forma uma fenda e é o sítio catalítico da enzima. Após a interação com os ligantes (ATP e ácido graxo) os dois domínios se aproximam, diminuindo a exposição da região ao solvente (Hisanaga *et al.*, 2004; Black *et al.*, 2000). Como discutido anteriormente, esse domínio está presente em 5 genes de *H. salinarum* NRC-1 e em todos os casos ele está associado ao domínio PF00501 na extremidade N-terminal da proteína. No entanto, o gene VNG_RS00670 (*alkK*) também apresenta um iTSS na mesma posição do gene VNG_RS05220, e além disso, análise de dados de dRNA-seq de *H. volcanii* DS2 demonstrou a presença de inícios de transcrição internos a 3 genes homólogos nessa espécie, indicando que a geração de intraRNAs e posterior tradução em uma proteína alternativa pode estar presente também em outros organismos (Figura 25).

- **VNG_RS11810**

Por fim, o terceiro gene para o qual foi demonstrado a produção de uma proteína alternativa foi o VNG_RS11810. Esse gene codifica a proteína de controle de divisão celular *orc4* a qual faz parte da família de proteínas de divisão celular Cdc6. A anotação automatizada de inícios de transcrição em dados de dRNA-seq, através da ferramenta TSSAR (Amman *et al.*, 2014), não identificou inícios de transcrição internos a esse gene em nenhuma das condições analisadas. Porém, analisando as posições enriquecidas para o início de *reads*, o que pode indicar locais de início de transcritos, é possível identificar a existência de um pico

de contagens na condição de 17h em uma posição 423 pb *usptream* ao códon de parada de tradução do gene. Esse pico representa a contagem de 64 inícios de *reads* na biblioteca enriquecida para RNAs primários, enquanto a biblioteca controle apresenta 43,7 contagens após a normalização dos dados pelo número de *reads* alinhados ao genoma. Isso representa um aumento no número de início de *reads* de 1,46 vezes nessa posição na biblioteca enriquecida. Dessa forma optamos por classificar a posição como um verdadeiro início de transcrição. O resultado do *Western blot* confirma a existência de uma proteína entre os marcadores de 55 e 40 kDa, o que está de acordo com a predição da proteína original do gene VNG_RS11810, a qual apresentou 40,6 kDa. Além disso a presença de uma banda abaixo de 15 kDa corrobora a existência de uma proteína traduzida a partir de um códon alternativo de início de tradução (GTG) a 34 pb *downstream* do início do intraRNA (Figura 29C).

A versão menor da proteína codificada pelo gene VNG_RS11810 corresponde ao domínio PF09079 (CDC6-C) o qual apresenta uma estrutura *winged-helix*. Essa estrutura tem similaridade com estruturas encontradas em histonas e a função mais provável do domínio é de promover a interação com o DNA e também entre proteínas (Liu *et al.*, 2000). No gene VNG_RS11810 esse domínio está fusionado a um módulo do tipo AAA⁺, o qual tem função de ligação de ATP. Essa arquitetura de domínios é característica da família de proteínas de divisão celular Cdc6/Cdc18 e é encontrada em mais 4 genes de *H. salinarum* NRC-1 (VNG_RS09430, VNG_RS00175, VNG_RS11490 e VNG_RS04790). Em outro gene da família o domínio CDC6-C está associado como o domínio PF13191 o qual também faz parte do grupo AAA⁺ e possui atividade de ATPase. Por fim, no gene (VNG6187G – *orc3*), que também faz parte da família *cdc6*, ele é o único domínio presente, apesar desse gene ter sido descontinuado na anotação mais recente de *H. salinarum* NRC-1. Em 3 dos 10 genes que pertencem a família *cdc6* o domínio CDC6-C não está presente.

Além de estar presente em proteínas de controle de divisão celular o domínio de ligação ao DNA, *winged-helix*, é encontrado em uma ampla variedade de proteínas, como helicases (Yamada *et al.*, 2002), endonucleases (Wah *et al.*, 1997) e transposases (Clubb *et al.*, 1996). Além disso parece ser um domínio ancestral nos sistemas de transcrição, uma vez que é encontrado em elementos basais da transcrição de procariotos e eucariotos, com fatores transcrição de archaea, no fator sigma 70 em bactérias e também na RNA Polimerase do tipo I, II e III em *Saccharomyces* (Teichmann *et al.*, 2012). Essa característica destaca o comportamento modular desse domínio em diferentes organismos.

Figura 29: Proteínas alternativas detectados por western blot para três genes com iTSS. Para os três genes foi possível detectar tanto a versão original da proteína (#) quando sua versão alternativa (*) produzida a partir da tradução do intraRNA. Na representação da região codificante dos genes (caixa cinza) as regiões correspondentes aos domínios proteicos são indicadas pelas caixas coloridas. As setas indicam a posição e a orientação do início de transcrição. O tamanho aproximado de cada proteína e sua versão alternativa são apresentados abaixo de cada gene. A seta verde indica o início de transcrição do intraRNA traduzido para a produção da proteína alternativa. O códon de início de tradução é indicado após a seta verde, junto com a distância em relação ao iTSS. A estrutura das proteínas representa as regiões correspondentes a cada domínio. A barra preta abaixo de cada esquema representa 100nt.

Além dos três genes para os quais identificamos a presença de uma proteína alternativa, equivalente a tradução dos intraRNAs, dados da literatura sugerem a existência de um fenômeno similar na archaea metanogênica *Methanococcus jannaschii*. Analisando informações obtidas a partir da espectrometria de massa de proteínas separadas por meio de eletroforese em gel 2D Mukhopadhyay et al, 2000 identificaram peptídios provenientes de uma proteína de ~16 kDa equivalentes a região C-terminal da proteína FlaD, uma proteína constituinte do flagelo, com peso molecular predito de 39,95 kDa. A superexpressão de proteínas do flagelo de *Methanococcus voltae* em *E. coli* permitiram a identificação tanto da proteína original FlaD como também de uma versão alternativa com 16 kDa. O sequenciamento N-terminal demonstrou que se tratava de uma proteína traduzida a partir de uma códon de tradução interno ao gene e não um produto de degradação como previamente proposto (Thomas & Jarrell, 2001). Em *H. salinarum* NRC-1 o gene codificante da proteína FlaD encontra-se fusionado ao gene da proteína FlaC e na sua porção final é possível identificar um iTSS *upstream* ao domínio proteico PF04659, encontrado também na região C-terminal da proteína FlaD em *M. jannaschii* e *M. voltae*. O peso molecular predito para uma possível proteína traduzida a partir de um intraRNA nessa região é de ~14,4 kDa. Apesar de não existirem informações a respeito de inícios de transcrição em *M. jannaschii* e *M. voltae* a existência de um iTSSs na mesma região em *H. salinarum* NRC-1 sugere que as versões menores da proteína FlaD, identificadas nesses organismos, possam ser traduzidas a partir de um intraRNA.

A demonstração experimental de que intraRNAs correspondendo a domínios proteicos específicos são capazes de produzir proteínas alternativas, em conjunto com a informação de que muitos desses domínios podem se fusionar a outros para compor novas proteínas, ou até mesmo formar proteínas de domínio único (Anexo I), destaca o comportamento modular dessas regiões.

Os domínios constituem as unidades básicas de formação de proteínas, atuando como peças de Lego passíveis de combinação para a formação de proteínas com novas funções (Das & Smith, 2000). Em eucariotos, os principais mecanismos para a geração de novas proteínas, a partir de domínios já estabelecidos, envolve recombinação de éxons (*exon shuffling*) (Patthy, 1999) e também o *splicing* alternativo (Babushok *et al.*, 2007). Já em procariotos, os quais apresentam genomas compactados e não dispõem de genes interrompidos por íntrons, a geração de novas proteínas com multidomínios está relacionada com fusão ou fissão de genes (Pasek *et al.*, 2006).

Nossos dados indicam também a associação dos iTSS com sítios de interação de fatores de transcrição localizados internamente as regiões codificantes (Figura 23). Essa característica já havia sido associada a transcrição de subunidades de operons de *H. salinarum* NRC-1 em condições específicas do crescimento dessa archaea, aumento o repertório de transcritos relacionados a uma mesma região genômica (Koide *et al.*, 2009). Dados similares tem sido reportados em uma ampla variedade de procariotos, demonstrando que a existência de sítios de ligação a fatores de transcrição dentro de genes é amplamente distribuído (Bonocora *et al.*, 2015; Zhu *et al.*, 2006; Tanay, 2006; Shimada *et al.*, 2008).

Dessa forma, a existência de um grande número de regiões promotoras internamente as regiões codificantes, aliada a uma evolução modular das proteínas, poderiam contribuir para a existência de domínios proteicos sendo transcritos e traduzidos de maneira independente, mesmo após processos de fusão a outros domínios para a constituição de proteínas multidomínios. A conservação da associação de iTSSs a regiões codificantes de determinados domínios proteicos em diferentes archaea e bactérias sugere que esse poderia ser um mecanismo amplamente distribuído entre procariotos.

4.2.5 Superexpressão de domínios através de intraRNAs

Para avaliar os possíveis efeitos fenotípicos da superexpressão de domínios proteicos individuais no desenvolvimento de *H. salinarum* NRC-1 foram realizadas curvas de crescimento com linhagens superexpressando os intraRNAs dos genes VNG_RS07995, VNG_RS05220, VNG_RS11810, além do intraRNA do gene VNG_RS00175, o qual também codifica uma proteína de controle de ciclo celular (Seção 3.11 - Superexpressão de intraRNAs).

A curva de crescimento das linhagens superexpressando os intraRNAs em meio de crescimento CM (*complex medium*) não apresentou diferença significativa em relação a linhagem controle contendo o vetor de superexpressão vazio (Figura 30A). As pequenas diferenças de crescimento não foram reproduzíveis entre as réplicas biológicas utilizadas no experimento. Reproduções do experimento em outros momentos apresentaram resultados similares (dados não apresentados). Um dos possíveis motivos para ausência de fenótipo relacionado a superexpressão desses intraRNAs pode estar relacionado com a grande redundância de genes que apresentam os domínios presentes na região correspondente aos intraRNAs. Como discutido anteriormente os domínios relacionados ao gene VNG_RS07995 (trkA-N e C) estão presentes, em conjunto, em 9 genes de *H. salinarum* NRC-1 (Tabela 12) enquanto que os domínios relacionados aos intraRNAs dos genes VNG_RS05220 (PF13193) e VNG_RS00175 (PF09079) estão presentes em 5 e 7 genes, respectivamente.

Como a proteína produzida a partir da tradução do intraRNA do gene VNG_RS07995 apresenta o domínio trkA, o qual está relacionado com o transporte de potássio (Schlosser et al, 1993; Coker et al, 2007), decidimos avaliar o crescimento da linhagem superexpressando esse intraRNA em um meio de cultura deficiente para potássio (Wurtmann *et al.*, 2014). O crescimento da linhagem controle e a linhagem superexpressando o intraRNA apresenta

diferenças na fase inicial, sendo que a linhagem superexpressando o intraRNA atinge a DO de 0,5 me torno de 21 horas após a inoculação, enquanto na linhagem controle essa DO é alcançada somente após 27 horas ou mais de crescimento (Figura 30B e C). Após a fase estacionária ambas as linhagens apresentam comportamentos similares no crescimento, mesmo após a adição de 100mM de KCl no momento 37 horas para que as linhagens retomem o crescimento (Figura 30C).

O potássio apresenta um papel fundamental na osmorregulação de *H. salinarum* NRC-1, uma vez que a alta concentração externa de sódio (4-5 M) é contrabalanceada com uma alta concentração interna de potássio (Christian & Waltho, 1962). Apesar de apresentar diferentes genes relacionados com o transporte desse íon, concentrações externas de potássio abaixo de 1 mM tendem a inibir o crescimento de *Halobacterium* após um determinado período (Strahl & Greie, 2008). O crescimento mais acelerado da linhagem superexpressando o do domínio *trkA* deve estar relacionado com uma maior facilidade na captação de potássio quando esse está em níveis mais baixo no meio de cultura, possibilitando assim um crescimento mais acelerado em relação a linhagem controle. Após essa fase o crescimento não progride em ambas as linhagens, se mais potássio não for adicionado, uma vez que a disponibilidade desse íon no meio de cultura é baixa. Se quantidades maiores de potássio forem adicionadas ao meio de cultura as duas linhagens apresentam respostas similares na fase final de crescimento, devido à presença de diversos genes responsáveis pela captação de potássio.

Figura 30: Curva de crescimento com linhagens de *H. salinarum* NRC-1 superexpressando *intraRNAs*. A superexpressão dos *intraRNAs* dos genes VNG_RS07995, VNG_RS05220 e VNG_RS00175 e VNG_RS11810 não apresentou efeito fenotípico no crescimento de *H. salinarum* NRC-1 em meio complexo (A). É possível perceber um crescimento mais acelerado da linhagem superexpressando o *intraRNA* do gene VNG_RS07995, a qual atinge D.O. ligeiramente mais elevadas nas primeiras 24 horas de crescimento em meio de cultura deficiente de potássio (B/C). Em baixas concentrações de potássio as linhagens superexpressando o *intraRNA* do gene VNG_RS07995 e a linhagem controle não conseguem retomar o crescimento após a fase estacionária (B). Quando 100mM de potássio são adicionados ao meio de cultura as linhagens voltam a crescer, no entanto não apresentam diferença no crescimento final (C).

A capacidade da superexpressão do intraRNA do gene VNG_RS07995 afetar a fase inicial do crescimento de *H. salinarum* NRC-1 em condição de baixa concentração de potássio, indica que as proteínas traduzidas a partir desses transcritos podem apresentar papéis funcionais. Essa informação tem implicações na essencialidade de genes, uma vez que a existência de uma região gênica transcrita e traduzida de maneira independente possibilitaria que domínios essenciais continuassem a ser produzidos mesmo que alterações em outras regiões do gene resultassem na produção de proteínas truncadas e sem funcionalidade (Lu *et al.*, 2015).

A determinação de regiões essenciais no desenvolvimento dos organismos envolve o uso de experimentos de deleção dirigidos a determinados genes, mutagênese por meio de mobilização de elementos de transposição e inibição de transcritos via RNAs *antisense* (Glass *et al.*, 2006; Forsyth *et al.*, 2002; de Berardinis *et al.*, 2008). No entanto, o conjunto de genes essenciais identificados em diferentes organismos apresenta baixa sobreposição, indicando que determinados genes podem ou não ser classificados como essenciais de acordo com o organismo analisado (Lu *et al.*, 2015). Além disso, a análise experimental de genes altamente conservados em bactérias indica que menos de um quarto desses genes são essenciais, indicando que a uma alta conservação não necessariamente está relacionada com a essencialidade do gene (Zalacain *et al.*, 2003).

Trabalhos recentes têm apontado que a essencialidade dos genes pode estar relacionada com a presença de determinados domínios proteicos específicos nos genes analisados (Lu *et al.*, 2014). Resultados nesse sentido foram obtidos com o uso de sistemas de mutagênese por meio de transposons nas bactérias *C. crescentus* e *M. pneumoniae*, onde a análise dos genes essenciais demonstrou a existência de sítios de inserção em regiões específicas dos genes, enquanto regiões correspondentes a determinados domínios específicos não foram interrompidos (Christen *et al.*, 2011; Lluch-Senar *et al.*, 2015). Além disso, a

análise de 355 proteínas essenciais em bactérias, identificou 238 domínios sem função conhecida (DUF), sendo que em muitas dessas proteínas esse era o único domínio presente. Esse resultado destaca que domínios sem função estabelecida podem apresentar importantes papéis funcionais no desenvolvimento de bactérias (Goodacre et al., 2014).

Analizadas em conjunto, as informações a respeito dos inícios de transcrição internos a regiões codificantes permitiram a identificação de uma nova classe de RNAs em *H. salinarum* NRC-1, os RNAs intragênicos (intraRNAs). A existência de 554 iTSS não relacionados a erros de anotação ou a artefatos da técnica de dRNA-seq mostra que a produção de transcritos internos aos genes é amplamente distribuído no genoma dessa archaea. Além disso, a associação de 89 desses intraRNAs com domínios proteicos específicos, e a demonstração experimental de que 3 deles estão envolvidos com a produção de proteínas alternativas, sugere que os intraRNAs podem estar associados com o comportamento modular de domínios proteicos, com importantes implicações na essencialidade de determinadas regiões do genoma. A conservação da associação entre os intraRNAs e determinados domínios proteicos em outros procariotos indica que esse fenômeno pode ser amplamente distribuído nesse grupo.

4.3 Identificação do término de transcrição (TTS)

Apesar da menor disponibilidade de informações a respeito do término de transcrição em Archaea, diferentes experimentos têm apontando que microorganismos pertencente a esse domínio devem utilizar um mecanismo intrínseco de terminação similar ao encontrado em Bactérias. Análises realizados na Archaea *Thermococcus kodakarensis* indicam que a presença de uma sequência poly(T) após o *stop* codon dos genes é suficiente para promover a desestabilização do complexo da RNA polimerase e interrupção da transcrição. *H. salinarum*

NRC-1 no entanto possui um genoma rico em conteúdo GC, o qual pode chegar a 62%, diminuindo a disponibilidade de sequências ricas em timina que podem estar associadas ao término de transcrição.

Para identificação dos termos de transcrição em *H. salinarum* NRC-1 analisamos dados de RNA-seq produzidos a partir do sequenciamento de RNAs total extraído de *H. salinarum* NRC-1 na condição referência do crescimento. Os *reads* foram alinhados ao genoma referência utilizando os parâmetros padrão da ferramenta, descartando todos os *reads* que alinhavam em mais de uma posição do genoma. Para identificação das posições associadas ao término de transcrição utilizamos uma abordagem similar a adotada para identificação de transcrição em *S. solfataricus*, onde o início de transcrição foi identificado através da identificação das posições associadas ao início de *reads* (Wurtzel *et al.*, 2010). Para a identificação do término de transcrição todos os *reads* de tamanho igual ao máximo permitido pelo protocolo de sequenciamento (150 nt) foram descartados. Para identificação das posições associadas ao término de transcrição foi produzido um sinal de expressão gênica contabilizando apenas as posições associadas ao término dos *reads* alinhados. A partir das contagens das posições associados ao término de *reads* implementamos um algoritmo visando a identificação das posições onde a essa contagem estivesse enriquecida em relação a contagem das regiões vizinhas (Figura 31). Para isso, o *script* implementado percorre todas as posições genômicas, identificando primeiramente as posições que apresentam um número de contagem igual ou acima de um *threshold* determinado pelo usuário. Quando uma posição que satisfaça esse critério é identificada, o *script* analisa a mediana e o desvio padrão da contagem de término de *reads* de um intervalo de posições anterior e posterior à posição analisada. Caso todas essas condições sejam satisfeitas o *script* identifica a posição como um ponto de quebra de sinal, ou *breakpoint*.

Figura 31. Metodologia para detecção das regiões com enriquecimento de término de reads. O algoritmo percorre as posições do genoma analisando o número de reads que iniciam em cada posição. Para definição de um breakpoint três condições devem ser satisfeitas: i) o número de contagens deve ser igual ou superior a um determinado threshold ($t = 10$), ii) a mediana das contagens de intervalo de posições antecedente a posição analisada deve ser menor ou igual a uma porcentagem das contagens nesta posição (5%), iii) o desvio padrão das contagens do intervalo de posições antecedentes não pode ser maior que o valor definido ($s = 5$).

Para identificação de possíveis termos de transcrição utilizamos o script implementado buscando pelas posições com, no mínimo, 10 contagens de término de reads, além disso, as regiões 50 pb. *upstream* e *downstream* à posição analisada deveriam apresentar valores médios de contagens de término de reads de, no máximo, 25% do valor na posição analisada. De acordo com esses critérios foram identificadas 1431 posições enriquecidas para o término de reads, ao longo de todo o genoma de *H. salinarum* NRC-1. Buscando pelos breakpoints localizados até 150 pb. *downstream* de um códon de parada da tradução, e em região intergênica, identificamos um conjunto de 135 posições possivelmente associadas ao

término de transcrição, as quais estão relacionadas a 134 genes de *H. salinarum* NRC-1, com uma distância média de 63 pb em relação ao stop códon (Anexo J). Como as posições enriquecidas para o término de *reads* podem ter sido geradas através de processamento da extremidade 3' realizamos uma busca das posições colocalizadas com assinaturas de 4 timinas em uma sequência de 5 nucleotídeos (polyT 5/4) ou 6 timinas em uma sequência de 10 nucleotídeos (polyT 10/6). Dentre as 135 posições analisadas 55 apresentaram a assinatura polyT 5/4 e 30 apresentaram a assinatura polyT 10/6, totalizando 58 genes onde a posição enriquecida para o término de *reads* estava colocalizada com uma assinatura polyT (Anexo J). 57,1% das assinaturas polyT 5/4 e 56,2% das assinaturas polyT 10/6 estão localizadas até 20 pb. *upstream* ao sítio enriquecido para o término de *reads* (Figura 32), indicando uma associação entre essas regiões. A distância média dos sítios de término de *reads* associados a assinaturas polyT em relação ao stop codon do gene *upstream* é de 62 pb, quando consideradas todas as sequências e de 60,75 pb quando analisamos apenas os *breakpoints* com uma assinatura polyT até 20pb *upstream* (Anexo J).

A preponderância de sequências polyT *upstream* as posições associados com uma quebra de sinal de RNA-seq (Figura 32) indica que essas regiões devem estar atuando no término de transcrição em *H. salinarum* NRC-1, de forma similar ao que ocorre em *T. kodakarensis* (Santangelo *et al.*, 2009). Essa informação é corroborada por análises experimentais realizadas em nosso laboratório nas quais a clonagem de sequências de aproximadamente 200 pb, contendo assinaturas poliT, entre o promotor e o gene codificante da proteína GFP, foi capaz de diminuir significativamente os níveis de fluorescência *in vivo* (dados não apresentados). Além disso, as informações obtidas em nossas análises sugerem que as regiões 3' UTRs de *H. salinarum* NRC-1, as quais apresentam em média 62 pb, são maiores do que na linhagem R1, a qual apresenta regiões 3' UTR de 50 nt. Esse tamanho está

mais próximo ao que foi detectado na haloarchaea *H. volcanii*, a qual apresenta regiões 3' UTRS em média de 66 pb (Brenneis et al., 2007).

Embora as informações obtidas a respeito do término de transcrição em *H. salinarum* NRC-1 sugiram a existência de um mecanismo associado a assinaturas polyT nessa archaea, a análise global através dos dados de RNA-seq permitiu a identificação de possíveis TTS para apenas 58 genes. Um dos principais motivos para o baixo número de TTSs identificados está possivelmente relacionado a variações nas extremidades 3' dos transcritos (Stazic & Voß, 2016), produzindo um sinal de expressão gênica com decaimento gradativo e dificultado a identificação de uma posição exata de término de transcrição (Koide *et al.*, 2009). Além disso, diferentemente da identificação de TSS através do dRNA-seq, os dados analisados não nos permitem diferenciar verdadeiros términos de transcrição de transcritos processados em suas extremidades 3'. O uso de técnicas específicas, como o sequenciamento em larga escala das extremidades de cDNA – RACE-seq (*rapid amplification of cDNA ends*) (Salehi-Ashtiani *et al.*, 2009; Goldfarb & Cech, 2013), pode ser uma alternativa para a identificação em larga escala dos términos de transcrição em *H. salinarum* NRC-1.

Figura 32. Distribuição das assinaturas polyT em relação ao breakpoint. Os gráficos indicam o número de assinaturas polyT identificadas na região 40 pb upstream e 40 pb downstream aos breakpoints. A grande maioria das assinaturas se concentra entre a região 15 pb downstream e posição do breakpoint. a) polyT 5/4 b) polyT 10/6

5 Conclusão

O fenômeno da transcrição pervasiva, segundo o qual todas as posições genômicas estão envolvidas em processos de transcrição em algum momento ou condição do desenvolvimento, traz à tona um cenário muito mais complexo da expressão gênica, revelando inúmeras espécies de novos RNAs ao longo do genoma de diferentes organismos (Wade & Grainger, 2014; Berretta & Morillon, 2009). Embora as funções da maioria desses transcritos ainda não seja completamente compreendida é possível relacionar alguns desses RNAs com importantes funções regulatórias nos organismos onde estão presentes, e identificar elementos evolutivamente conservados, indicando importantes papéis funcionais nos organismos onde são encontrados (Shao *et al.*, 2014). Um organismo muito atrativo para estudos de análise e caracterização do transcrito é a archaea halofílica *Halobacterium salinarum*. Além de pertencer a um domínio da vida pouco caracterizado, e ser de fácil manipulação em laboratório, esse organismo dispõe de informações acerca de sua resposta transcricional em uma ampla variedade de condições, permitindo a análise integrada de seus dados e identificação de condições específicas do desenvolvimento onde os transcritos de interesse apresentem comportamento diferenciado. A partir dos resultados obtidos nesse trabalho diferentes elementos do transcrito dessa archaea foram caracterizados.

O uso da técnica de dRNA-seq, a qual se baseia na comparação do sequenciamento de uma biblioteca enriquecida para RNAs primários e uma biblioteca controle, para identificação de inícios de transcrição de forma global (Sharma & Vogel, 2014), levou a identificação de 179 RNAs associados ao início de transcrição (TSSaRNAs). Esses RNAs foram identificados em mais 10 espécies de archaea, sendo evolutivamente conservados nos três domínios da vida, sugerindo uma importante função biológica (Zaramela *et al.*, 2014).

De maneira similar, o enriquecimento e sequenciamento de RNAs primários a partir de amostras extraídas em diferentes pontos da curva de crescimento de *H. salinarum* NRC-1 permitiu a identificação de 4540 TSS ao longo de todo o genoma desse organismo (Anexo A). Destes, 1978 estão *upstream* a regiões codificantes, permitindo a identificação do início de transcrição para 1545 genes (58,8%). A identificação de 737 inícios de transcrição *antisense* a genes (aTSSs) e 732 inícios de transcrição em regiões intergênicas sem relação a outras anotações (oTSSs) indica a existência de um grande número de transcritos envolvidos com a produção de RNAs regulatórios ou com potencial codificante de proteínas ainda não anotadas. A análise do perfil transcricional das regiões *downstream* a esses inícios de transcrição demonstrou que 62,9 % das regiões associadas com inícios de transcrição *antisense* apresentam expressão correlacionada com os genes na fita oposta (Anexo B), enquanto 67,3 % dos oTSS são diferencialmente expressos em, pelo menos, uma das condições analisadas (Anexo E). Além disso, a análise comparativa do oTSS com informações de proteômica indicam que alguns desses inícios de transcrição podem estar envolvidos com a produção de reguladores de transcrição.

Por fim, a presença de 1365 inícios de transcrição localizados internamente as regiões codificantes (iTSS) de diversos genes revela um cenário complexo de transcritos associados a genes. Embora muitos desses inícios de transcrição estejam relacionados com ncRNAs previamente caracterizados ou sejam gerados pela presença de estruturas no mRNA (Jäger *et al.*, 2014), a análise detalhada dos iTSSs revelou um conjunto de 554 inícios de transcrição que não podem ser explicados por esses fatores. Experimentos de *northern* e *western blot* revelaram que alguns desses iTSS estão envolvidos com a transcrição de RNAs intragênicos (intraRNAs) codificantes de proteínas alternativas. A sobreposição desses intraRNAs com domínios proteicos específicos, e sua conservação em outras espécies de procariotos, sugere que sua origem esteja relacionada com a modularidade dos domínios, os quais continuariam a

ser produzidos de forma independente mesmo quando fusionados a outros domínios em proteínas maiores. O crescimento mais acelerado de *H. salinarum* NRC-1 nas fases iniciais do desenvolvimento, através da superexpressão do intraRNA codificante de um domínio proteico relacionado ao transporte de potássio, demonstra que esses RNAs podem desempenhar importantes papéis funcionais no organismo (Figura 30).

Além dos iTSSs associados a produção de domínios proteicos específicos, existe um grande número de inícios de transcrição internos às regiões codificantes que não apresentam relação direta com essas regiões. Além de produzirem pequenos RNAs não codificantes (Figura 2), alguns desses iTSSs podem estar envolvidos com a produção de intraRNAs codificantes de versões truncadas das proteínas originais, ou até mesmo podem codificar proteínas em uma fase de leitura diferente, como visto na bactéria *C. crescentus* (Schrader *et al.*, 2014), ampliando assim o número de proteínas geradas a partir de uma mesma região genômica.

Por fim, através da análise das posições enriquecidas com o término de *reads*, obtidos a partir do sequenciamento de bibliotecas de pequenos RNAs, e colocadas com sítios de término de transcrição do tipo polyT foi possível identificar a posição de término de transcrição para 58 genes de *H. salinarum* NRC-1 (Anexo J). Além de demonstrar que sequências polyT podem atuar como terminadores intrínsecos da transcrição em genomas com alto conteúdo GC, de modo similar ao que ocorre na archaea *T. kodakarensis* (Santangelo *et al.*, 2009), esses dados fornecem informações importantes sobre regiões que podem ser alvo de regulação pós-transcricional através da interação com ncRNAs (Waters & Storz, 2009; Babski *et al.*, 2014).

Analisadas em conjunto, as informações obtidas a partir dos experimentos e análises conduzidos durante a execução desse trabalho ajudam a traçar um cenário mais completo dos elementos que compõem o transcrito de *H. salinarum* NRC-1. Além disso, nossos

resultados revelam a existência de um novo papel funcional para os RNAs intragênicos, os quais são responsáveis pela tradução de proteínas alternativas correspondente a domínios proteicos específicos e devem ser conservados nas demais archaea e bactérias.

Referências

- Alexa A, Rahnenführer J & Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**: 1600–1607
- Alexander RP, Fang G, Rozowsky J, Snyder M & Gerstein MB (2010) Annotating non-coding regions of the genome. *Nat. Rev. Genet.* **11**: 559–571
- Allers T & Mevarech M (2005) Archaeal genetics - the third way. *Nat. Rev. Genet.* **6**: 58–73
- Almeida-E-Silva DC & Vêncio RZN (2015) Sifter-T: A scalable and optimized framework for the SIFTER phylogenomic method of probabilistic protein domain annotation. *Biotechniques* **58**: 140–142
- Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–10
- Amman F, Wolfinger MT, Lorenz R, Hofacker IL, Stadler PF & Findeiß S (2014) TSSAR: TSS annotation regime for dRNA-seq data. *BMC Bioinformatics* **15**: 89
- Anantharaman V, Koonin E V & Aravind L (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.* **307**: 1271–92
- Andersen GR, Nissen P & Nyborg J (2003) Elongation factors in protein biosynthesis. *Trends Biochem. Sci.* **28**: 434–441
- Andrews SJ & Rothnagel JA (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* **15**: 193–204
- Babski J, Haas KA, Näther-Schindler D, Pfeiffer F, Förstner KU, Hammelmann M, Hilker R, Becker A, Sharma CM, Marchfelder A & Soppa J (2016) Genome-wide identification of transcriptional start sites in the haloarchaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* **17**: 629
- Babski J, Maier L-K, Heyer R, Jaschinski K, Prasse D, Jäger D, Randau L, Schmitz R a, Marchfelder A & Soppa J (2014) Small regulatory RNAs in Archaea. *RNA Biol.* **11**: 1–10
- Babushok D V., Ostertag EM & Kazazian HH (2007) Current topics in genome evolution: Molecular mechanisms of new gene formation. *Cell. Mol. Life Sci.* **64**: 542–554
- Bachellerie JP, Cavaillé J & Hüttenhofer A (2002) The expanding snoRNA world. *Biochimie* **84**: 775–790
- van Bakel H, Nislow C, Blencowe BJ & Hughes TR (2010) Most “dark matter” transcripts are associated with known genes. *PLoS Biol.* **8**: e1000371
- Baliga NS, Bjork SJ, Bonneau R & Pan M (2004) Systems level insights into the stress response to UV radiation in the halophilic archaeon *Halobacterium* NRC-1. *Genome Res.* **14**: 1025–1035
- Bánfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, Kundaje A, Gunawardena HP, Yu Y, Xie L, Krajewski K, Strahl BD, Chen X, Bickel P, Giddings MC, Brown JB & Lipovich L (2012) Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**: 1646–1657
- Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, Antunes R, Arganiska J, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Chavali G, Cibrian-Uhalte E, Da Silva A, De Giorgi M, Dogan T, Fazzini F, Gane P, et al. (2015) UniProt: A hub for protein information. *Nucleic Acids Res.* **43**: D204–D212
- Beer KD, Wurtmann EJ, Pinel N & Baliga NS (2014) Model organisms retain an “ecological memory” of complex ecologically relevant environmental variation. *Appl. Environ. Microbiol.* **80**: 1821–1831

- Bell SD & Jackson SP (2001) Mechanism and regulation of transcription in archaea. *Curr. Opin. Microbiol.* **4**: 208–213
- de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, Samair S, Lechaplais C, Gyapay G, Richez C, Durot M, Kreimeyer A, Le Fèvre F, Schächter V, Pezo V, Döring V, Scarpelli C, Médigue C, Cohen GN, Marlière P, et al. (2008) A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Mol. Syst. Biol.* **4**: 174
- Berretta J & Morillon A (2009) Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* **10**: 973–982
- Bilusic I, Popitsch N, Rescheneder P, Schroeder R & Lybecker M (2014) Revisiting the coding potential of the *E. coli* genome through Hfq co-immunoprecipitation. *RNA Biol.* **11**: 37–41
- Black PN, DiRusso CC, Sherin D, MacColl R, Knudsen J & Weimar JD (2000) Affinity labeling fatty acyl-CoA synthetase with 9-p-azidophenoxy nonanoic acid and the identification of the fatty acid-binding site. *J. Biol. Chem.* **275**: 38547–38553
- Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC, Longabaugh W, Vuthoori M, Whitehead K, Madar A, Suzuki L, Mori T, Chang D-E, Diruggiero J, Johnson CH, Hood L, et al. (2007) A predictive model for transcriptional control of physiology in a free living cell. *Cell* **131**: 1354–65
- Bonocora RP, Smith C, Lapierre P & Wade JT (2015) Genome-Scale Mapping of *Escherichia coli* σ^{54} Reveals Widespread, Conserved Intragenic Binding. *PLoS Genet.* **11**: 1–30
- Brenneis M, Hering O, Lange C & Soppa J (2007) Experimental characterization of Cis-acting elements important for translation and transcription in halophilic archaea. *PLoS Genet.* **3**: e229
- Brenneis M & Soppa J (2009) Regulation of translation in haloarchaea: 5'- and 3'-UTRs are essential and have to functionally interact in vivo. *PLoS One* **4**: e4484
- Brooks AN, Reiss DJ, Allard A, Wu W-J, Salvanha DM, Plaisier CL, Chandrasekaran S, Pan M, Kaur A & Baliga NS (2014) A system-level model for the microbial regulatory genome. *Mol. Syst. Biol.* **10**: 740
- Brosius J (2005) Waste not, want not - Transcript excess in multicellular eukaryotes. *Trends Genet.* **21**: 287–288
- Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple C a M, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**: 626–635
- Cavicchioli R (2002) Extremophiles and the search for extraterrestrial life. *Astrobiology* **2**: 281–292
- Cavicchioli R (2011) Archaea--timeline of the third domain. *Nat. Rev. Microbiol.* **9**: 51–61
- Chao Y, Papenfort K, Reinhardt R, Sharma CM & Vogel J (2012) An atlas of Hfq-bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs. *EMBO J.* **31**: 4005–4019
- Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS & Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**: 1149–54
- Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Collier JA, Fero MJ, McAdams HH & Shapiro L (2011) The essential genome of a bacterium. *Mol. Syst. Biol.* **7**: 1–7

- Christian JHB & Waltho J a (1962) Solute concentrations within cells of halophilic and non-halophilic bacteria. *Biochim. Biophys. Acta* **65**: 506–508
- Clark MB, Choudhary A, Smith MA, Taft RJ & Mattick JS (2013) The dark matter rises: the expanding world of regulatory RNAs. *Essays Biochem.* **54**: 1–16
- Clubb RT, Mizuuchi M, Huth JR, Omichinski JG, Savilahti H, Mizuuchi K, Clore GM & Gronenborn AM (1996) The wing of the enhancer-binding domain of Mu phage transposase is flexible and is essential for efficient transposition. *Proc. Natl. Acad. Sci.* **93**: 1146–1150
- Coker J a, DasSarma P, Kumar J, Müller J a & DasSarma S (2007) Transcriptional profiling of the model Archaeon Halobacterium sp. NRC-1: responses to changes in salinity and temperature. *Saline Systems* **3**: 6
- Consortium. IHGS (2001) Initial sequencing and analysis of the human genome. *Nature* **412**: 860–921
- Crooks G, Hon G, Chandonia J & Brenner S (2004) WebLogo: a sequence logo generator. *Genome Res.* **14**: 1188–1190
- Das S & Smith TF (2000) Identifying nature's protein Lego set. *Adv. Protein Chem.* **54**: 159–183
- DasSarma S, Robb FT, Place AR, Sowers KR, Schreier HJ & Fleischmann EM (1995) Archaea: a laboratory manual -- halophiles Plainview N.Y.: Cold Spring Harbor Laboratory Press
- David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW & Steinmetz LM (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U. S. A.* **103**: 5320–5325
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, et al. (2012) Landscape of transcription in human cells. *Nature* **489**: 101–108
- Dornenburg JE, DeVita AM, Palumbo MJ & Wade JT (2010) Widespread antisense transcription in Escherichia coli. *MBio* **1**: 1–4
- Durand S, Gilet L & Condon C (2012) The Essential Function of B. subtilis RNase III Is to Silence Foreign Toxin Genes. *PLoS Genet.* **8**:
- Engström PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Sin LT, Yang L, Kunarso G, Ng ELC, Batalov S, Wahlestedt C, Kai C, Kawai J, Carninci P, Hayashizaki Y, Wells C, Bajic VB, et al. (2006) Complex loci in human and mouse genomes. *PLoS Genet.* **2**: 564–577
- Facciotti MT, Pang WL, Lo F, Whitehead K, Koide T, Masumura K, Pan M, Kaur A, Larsen DJ, Reiss DJ, Hoang L, Kalisiak E, Northen T, Trauger S a, Siuzdak G & Baliga NS (2010) Large scale physiological readjustment during growth enables rapid, comprehensive and inexpensive systems analysis. *BMC Syst. Biol.* **4**: 64
- Facciotti MT, Reiss DJ, Pan M, Kaur A, Vuthoori M, Bonneau R, Shannon P, Srivastava A, Donohoe SM, Hood LE & Baliga NS (2007) General transcription factor specified global gene regulation in archaea. *Proc. Natl. Acad. Sci. U. S. A.* **104**: 4630–4635
- Falcon S & Gentleman R (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**: 257–258
- Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf, G. Hannon GJ, Kapranov P, Foissac S, Willingham AT, Dattagupta R, Dumais E & Gingeras TR (2009) Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**: 1028–1032

- Forsyth RA, Haselbeck RJ, Ohlsen KL, Yamamoto RT, Xu H, Trawick JD, Wall D, Wang L, Brown-Driver V, Froelich JM, Kedar GC, King P, McCarthy M, Malone C, Misiner B, Robbins D, Tan Z, Zhu ZY, Carr G, Mosca DA, et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**: 1387–1400
- Freyhult E, Gardner PP & Moulton V (2005) A comparison of RNA folding measures. *BMC Bioinformatics* **6**: 241
- Ge H, Walhout AJ. & Vidal M (2003) Integrating “omic” information: a bridge between genomics and systems biology. *Trends Genet.* **19**: 551–560
- Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison C a, Smith HO & Venter JC (2006) Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **103**: 425–430
- Goldfarb KC & Cech TR (2013) 3' terminal diversity of MRP RNA and other human noncoding RNAs revealed by deep sequencing. *BMC Mol. Biol.* **14**: 23
- Gomes-Filho JV, Zaramela LS, Italiani VC da S, Baliga NS, Vêncio RZN & Koide T (2015) Sense overlapping transcripts in IS1341 type transposase genes are functional non coding RNAs in archaea. *RNA Biol.* **12**: 490–500
- Goodacre NF, Gerloff DL & Uetz P (2014) Protein Domains of Unknown Function Are Essential in Bacteria. *MBio* **5**: e00744-13-e00744-13
- Grabowski B & Kelman Z (2003) Archeal DNA replication: eukaryal proteins in a bacterial context. *Annu. Rev. Microbiol.* **57**: 487–516
- Green M. & Sambrook D.W. J (2012) Molecular Cloning: A laboratory manual (Fourth Edition). *Cold Spring Harb. Press. Cold Spring Harb. New York.*: 2028 Available at: <http://www.amazon.com/Molecular-Cloning-Laboratory-Edition-Three/dp/1936113422>
- Gribaldo S, Poole AM, Daubin V, Forterre P & Brochier-Armanet C (2010) The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat. Rev. Microbiol.* **8**: 743–752
- Guan Y, Zhu Q, Huang D, Zhao S, Jan Lo L & Peng J (2015) An equation to estimate the difference between theoretically predicted and SDS PAGE-displayed molecular weights for an acidic peptide. *Sci. Rep.* **5**: 13370
- Güell M, Yus E, Lluch-Senar M & Serrano L (2011) Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat. Rev. Microbiol.* **9**: 658–669
- Hartman AL, Norais C, Badger JH, Delmas S, Haldenby S, Madupu R, Robinson J, Khouri H, Ren Q, Lowe TM, Maupin-Furlow J, Pohlschroder M, Daniels C, Pfeiffer F, Allers T & Eisen JA (2010) The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon. *PLoS One* **5**:
- Heinen TJAJ, Staubach F, Häming D & Tautz D (2009) Emergence of a New Gene from an Intergenic Region. *Curr. Biol.* **19**: 1527–1531
- Heyer R, Dörr M & Jellen-Ritter A (2012) High throughput sequencing reveals a plethora of small RNAs including tRNA derived fragments in *Haloferax volcanii*. *RNA Biol.*: 1011–1018
- Hisanaga Y, Ago H, Nakagawa N, Hamada K, Ida K, Yamamoto M, Hori T, Arii Y, Sugahara M, Kuramitsu S, Yokoyama S & Miyano M (2004) Structural basis of the substrate-specific two-step catalysis of long chain fatty acyl-CoA synthetase dimer. *J. Biol. Chem.* **279**: 31717–31726
- Ignoul S, Eggermont J, Karasawa A, Erkens GB, Berntsson RP, Otten R, Gea K, Mulder FAA, Poolman B, Russo BC, Horzempa J, Dee DMO, Schmitt DM, Brown MJ, Jr PEC, Xavier RJ, Nau GJ, Burg JS, Espenshade PJ, Mosberg JA, et al. (2011) CBS domains :

- structure , function , and pathology in human proteins CBS domains : structure , function , and pathology in human proteins.
- Irnov I, Sharma CM, Vogel J & Winkler WC (2010) Identification of regulatory RNAs in *Bacillus subtilis*. *Nucleic Acids Res.* **38**: 6637–51
- Jäger D, Förstner KU, Sharma CM, Santangelo TJ & Reeve JN (2014) Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis*. *BMC Genomics* **15**: 684
- Jager D, Sharma CM, Thomsen J, Ehlers C, Vogel J & Schmitz RA (2009) Deep sequencing analysis of the *Methanosarcina mazei* Go1 transcriptome in response to nitrogen availability. *Proc. Natl. Acad. Sci.* **106**: 21878–21882
- Jain R, Rivera MC & Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **96**: 3801–6
- Jain S, Caforio A & Driessen AJM (2014) Biosynthesis of archaeal membrane ether lipids. *Front. Microbiol.* **5**: 641
- Janscak P & Bickle T a (1998) The DNA recognition subunit of the type IB restriction-modification enzyme EcoAI tolerates circular permutations of its polypeptide chain. *J. Mol. Biol.* **284**: 937–48
- Jayaprakash AD, Jabado O, Brown BD & Sachidanandam R (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res.* **39**: e141
- Jiang M, Anderson J, Gillespie J & Mayne M (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics* **9**: 192
- Johnson JM, Edwards S, Shoemaker D & Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21**: 93–102
- Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H & Gingeras TR (2004) Novel RNAs identified from a in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**: 331–342
- Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**: 1484–1488
- Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD, Snyder MA & Basrai MA (2006) Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res.* **16**: 365–373
- Kawano M (2005) Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in *Escherichia coli*. *Nucleic Acids Res.* **33**: 1040–1050
- Kim D, Hong JS-J, Qiu Y, Nagarajan H, Seo J-H, Cho B-K, Tsai S-F & Palsson BØ (2012) Comparative Analysis of Regulatory Elements between *Escherichia coli* and *Klebsiella pneumoniae* by Genome-Wide Transcription Start Site Profiling. *PLoS Genet.* **8**: e1002867
- Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G & Greenberg ME (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182–7
- Klein C, Aivaliotis M, Olsen J V, Falb M, Besir H, Scheffer B, Bisle B, Tebbe A, Konstantinidis K, Siedler F, Pfeiffer F, Mann M & Oesterhelt D (2007) The Low

- Molecular Weight Proteome of *Halobacterium salinarum*. *J. Proteome Res.* **6**: 1510–1518
- Koga Y (2012) Thermal adaptation of the archaeal and bacterial lipid membranes. *Archaea* **2012**:
- Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo F-Y, Pratap A, Deutsch EW, Peterson A, Martin D & Baliga NS (2009) Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol. Syst. Biol.* **5**: 285
- Koonin E V. & Martin W (2005) On the origin of genomes and cells within inorganic compartments. *Trends Genet.* **21**: 647–654
- Kottemann M, Kish A, Iloanusi C, Bjork SJ & DiRuggiero J (2005) Physiological responses of the halophilic archaeon *Halobacterium* sp. strain NRC1 to desiccation and gamma irradiation. *Extremophiles* **9**: 219–27
- Kröger C, Colgan A, Srikumar S, Händler K, Sivasankaran SK, Hammarlöf DL, Canals R, Grissom JE, Conway T, Hokamp K & Hinton JCD (2013) An infection-relevant transcriptomic compendium for *salmonella enterica* serovar typhimurium. *Cell Host Microbe* **14**: 683–695
- Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A & Couso JP (2011) Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol.* **12**: R118
- Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foà R, Schliwka J, Fuchs U, Novosel A, Müller R-U, et al. (2007) A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell* **129**: 1401–1414
- Lange C, Zaigler A, Hammelmann M, Twellmeyer J, Raddatz G, Schuster SC, Oesterhelt D & Soppa J (2007) Genome-wide analysis of growth phase-dependent translational and transcriptional regulation in halophilic archaea. *BMC Genomics* **8**: 415
- Langmead B, Trapnell C, Pop M & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**: R25
- Lasa I, Toledo-Arana a., Dobin a., Villanueva M, de los Mozos IR, Vergara-Irigaray M, Segura V, Fagegaltier D, Penades JR, Valle J, Solano C & Gingeras TR (2011) Genome-wide antisense transcription drives mRNA processing in bacteria. *Proc. Natl. Acad. Sci.* **108**: 20172–20177
- Lee S, Liu B, Lee S, Huang S-X, Shen B & Qian S-B (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci.* **109**: E2424–E2432
- Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S & Kim VN (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**: 415–419
- Liu J, Smith CL, DeRyckere D, DeAngelis K, Martin GS & Berger JM (2000) Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. *Mol. Cell* **6**: 637–648
- Lluch-Senar M, Delgado J, Chen W-H, Llorens-Rico V, O'Reilly FJ, Wodke JA, Unal EB, Yus E, Martinez S, Nichols RJ, Ferrar T, Vivancos A, Schmeisky A, Stulke J, van Noort V, Gavin A-C, Bork P & Serrano L (2015) Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol. Syst. Biol.* **11**: 780–780
- Lu Y, Lu Y, Deng J, Peng H, Lu H & Lu LJ (2014) A novel essential domain perspective for exploring gene essentiality. *Bioinformatics* **31**: 2921–2929
- Lu Y, Lu Y, Deng J, Peng H, Lu H & Lu LJ (2015) A novel essential domain perspective for exploring gene essentiality. *Bioinformatics* **31**: 2921–2929

- Ma C & Simons RW (1990) The IS10 antisense RNA blocks ribosome binding at the transposase translation initiation site. *EMBO J.* **9**: 1267–74
- Mamanova L & Turner DJ (2011) Low-bias, strand-specific transcriptome Illumina sequencing by on-flowcell reverse transcription (FRT-seq). *Nat. Protoc.* **6**: 1736–47
- Martincorena I, Seshasayee ASN & Luscombe NM (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* **485**: 95–98
- Masters TL, Wachter S, Wachter J & Hill SA (2016) H-NS suppresses pilE intragenic transcription and antigenic variation in *Neisseria gonorrhoeae*. *Microbiol. (United Kingdom)* **162**: 177–190
- Matsumoto-Taniura N, Pirollet F, Monroe R, Gerace L & Westendorf JM (1996) Identification of novel M phase phosphoproteins by expression cloning. *Mol. Biol. Cell* **7**: 1455–69
- Mercer TR, Wilhelm D, Dinger ME, Solda G, Korbie DJ, Glazov EA, Truong V, Schwenke M, Simons C, Matthaei KI, Saint R, Koopman P & Mattick JS (2011) Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Res.* **39**: 2393–2403
- Metzker ML (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**: 31–46
- Miller C, Schwalb B, Maier K, Schulz D, Dümcke S, Zacher B, Mayer A, Sydow J, Marcinowski L, Dölken L, Martin DE, Tresch A & Cramer P (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* **7**: 458
- Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voss B, Steglich C, Wilde A, Vogel J & Hess WR (2011) An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proc. Natl. Acad. Sci. U. S. A.* **108**: 2124–9
- Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**: 621–8
- Mukhopadhyay B, Johnson EF & Wolfe RS (2000) A novel pH2 control on the expression of flagella in the hyperthermophilic strictly hydrogenotrophic methanarchaeon *Methanococcus jannaschii*. *Proc. Natl. Acad. Sci. U. S. A.* **97**: 11522–7
- Munro S & Pelham HR (1984) Use of peptide tagging to detect proteins expressed from cloned genes: deletion mapping functional domains of *Drosophila* hsp 70. *EMBO J.* **3**: 3087–93
- Nakaya HI, Amaral PP, Louro R, Lopes A, Fachel AA, Moreira YB, El-Jundi TA, da Silva AM, Reis EM & Verjovski-Almeida S (2007) Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol.* **8**: R43
- Nam J, Choi S & You B (2016) Incredible RNA : Dual Functions of Coding and Noncoding. *Mol. Cells* **39**: 367–374
- Ng W V, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl T a, Welti R, Goo Y a, Leithauser B, Keller K, Cruz R, Danson MJ, et al. (2000) Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. U. S. A.* **97**: 12176–81
- O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, et al. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**: D733–D745
- Omer AD (2000) Homologs of Small Nucleolar RNAs in Archaea. *Science (80-.).* **288**: 517–522

- Pasek S, Risler JL & Brézellec P (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* **22**: 1418–1423
- Patra AK (2012) Enteric methane mitigation technologies for ruminant livestock: A synthesis of current research and future directions. *Environ. Monit. Assess.* **184**: 1929–1952
- Pathy L (1999) Genome evolution and the evolution of exon-shuffling - A review. *Gene* **238**: 103–114
- Peters JM, Mooney RA, Grass JA, Jessen ED, Tran F & Landick R (2012) Rho and NusG suppress pervasive antisense transcription in *Escherichia coli*. *Genes Dev.* **26**: 2621–2633
- Pfeiffer F, Schuster SC, Broicher A, Falb M, Palm P, Rodewald K, Ruepp A, Soppa J, Tittor J & Oesterhelt D (2008) Evolution in the laboratory: the genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1. *Genomics* **91**: 335–46
- Plaisier CL, Lo F-Y, Ashworth J, Brooks AN, Beer KD, Kaur A, Pan M, Reiss DJ, Facciotti MT & Baliga NS (2014) Evolution of context dependent regulation by expansion of feast/famine regulatory proteins. *BMC Syst. Biol.* **8**: 122
- Ponting CP, Oliver PL & Reik W (2009) Evolution and functions of long noncoding RNAs. *Cell* **136**: 629–41
- Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH & Jensen TH (2008) RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science* (80-.). **322**: 1851–1854
- Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A & Finn RD (2012) The Pfam protein families database. *Nucleic Acids Res.* **40**: D290-301
- Qin Y, Polacek N, Vesper O, Staub E, Einfeldt E, Wilson DN & Nierhaus KH (2006) The Highly Conserved LepA Is a Ribosomal Elongation Factor that Back-Translocates the Ribosome. *Cell* **127**: 721–733
- Rivera MC, Jain R, Moore JE & Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U. S. A.* **95**: 6239–6244
- Robinson R (2010) Dark Matter Transcripts: Sound and Fury, Signifying Nothing? *PLoS Biol* **8**: e1000370
- Roscetto E, Angrisano T, Costa V, Casalino M, Förstner KU, Sharma CM, Di Nocera PP & De Gregorio E (2012) Functional characterization of the RNA chaperone Hfq in the opportunistic human pathogen *Stenotrophomonas maltophilia*. *J. Bacteriol.* **194**: 5864–74
- Salehi-Ashtiani K, Lin C, Hao T, Shen Y, Szeto D, Yang X, Ghamsari L, Lee H, Fan C, Murray RR, Milstein S, Svrtikapa N, Cusick ME, Roth FP, Hill DE & Vidal M (2009) Large-scale RACE approach for proactive experimental definition of *C. elegans* ORFeome. *Genome Res.* **19**: 2334–2342
- Santangelo TJ & Artsimovitch I (2011) Termination and antitermination: RNA polymerase runs a stop sign. *Nat. Rev. Microbiol.* **9**: 319–29
- Santangelo TJ, Cubonová L, Skinner KM & Reeve JN (2009) Archaeal intrinsic transcription termination in vivo. *J. Bacteriol.* **191**: 7102–8
- Santangelo TJ & Reeve JN (2006) Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *J. Mol. Biol.* **355**: 196–210
- Sass AM, Van Acker H, Förstner KU, Van Nieuwerburgh F, Deforce D, Vogel J & Coenye T (2015) Genome-wide transcription start site profiling in biofilm-grown *Burkholderia cenocepacia* J2315. *BMC Genomics* **16**: 775
- Savelsbergh a, Matassova NB, Rodnina M V & Wintermeyer W (2000) Role of domains 4 and 5 in elongation factor G functions on the ribosome. *J. Mol. Biol.* **300**: 951–961

- Savery NJ (2007) The molecular mechanism of transcription-coupled DNA repair. *Trends Microbiol.* **15**: 326–333
- Schlosser A, Hamann A, Bossemeyer D, Schneider E & Bakker EP (1993) NAD⁺ binding to the Escherichia coli K⁺-uptake protein TrkA and sequence similarity between TrkA and domains of a family of dehydrogenases suggest a role for NAD⁺ in bacterial transport. *Mol Microbiol* **9**: 533–543
- Schrader JM, Zhou B, Li G, Lasker K, Childers WS, Williams B, Long T, Crosson S, Mcadams HH, Weissman JS & Shapiro L (2014) The Coding and Noncoding Architecture of the Caulobacter crescentus Genome. *PLoS Genet.* **10**: e1004463
- Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn R a, Young R a & Sharp P a (2008) Divergent transcription from active promoters. *Science* **322**: 1849–51
- Seitzer P, Wilbanks EG, Larsen DJ & Facciotti MT (2012) A Monte Carlo-based framework enhances the discovery and interpretation of regulatory sequence motifs. *BMC Bioinformatics* **13**: 317
- Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ & Church GM (2000) RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nat. Biotechnol.* **18**: 1262–1268
- Sendler E, Johnson GD & Krawetz S a (2011) Local and global factors affecting RNA sequencing analysis. *Anal. Biochem.* **419**: 317–22
- Sesto N, Wurtzel O, Archambaud C, Sorek R & Cossart P (2012) The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. *Nat. Rev. Microbiol.* **11**: 75–82
- Shao W, Price MN, Deutschbauer AM, Romine MF & Arkin AP (2014) Conservation of Transcription Start Sites within Genes across a Bacterial Genus. *MBio* **5**:
- Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF & Vogel J (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464**: 250–5
- Sharma CM & Vogel J (2014) Differential RNA-seq: The approach behind and the biological insight gained. *Curr. Opin. Microbiol.* **19**: 97–105
- Shimada T, Ishihama A, Busby SJW & Grainger DC (2008) The Escherichia coli RutR transcription factor binds at targets within genes as well as intergenic regions. *Nucleic Acids Res.* **36**: 3950–3955
- Silva-Rocha R, Pontelli MC, Furtado GP, Zaramela LS & Koide T (2015) Development of New Modular Genetic Tools for Engineering the Halophilic Archaeon Halobacterium salinarum. *PLoS One* **10**: e0129215
- Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT & Grainger DC (2014) Widespread suppression of intragenic transcription initiation by H-NS. *Genes Dev.* **28**: 214–219
- Sinicropi D, Qu K, Collin F, Crager M, Liu M-L, Pelham RJ, Pho M, Rossi AD, Jeong J, Scott A, Ambannavar R, Zheng C, Mena R, Esteban J, Stephans J, Morlan J & Baker J (2012) Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. *PLoS One* **7**: e40092
- Skellam JG (1946) The Frequency Distribution of the Difference Between Two Poisson Variates Belonging to Different Populations. *J. R. Stat. Soc.* **109**: 296
- Slupska MM, King a G, Fitz-Gibbon S, Besemer J, Borodovsky M & Miller JH (2001) Leaderless transcripts of the crenarchaeal hyperthermophile Pyrobaculum aerophilum. *J. Mol. Biol.* **309**: 347–360

- Soppa J (1999) Normalized nucleotide frequencies allow the definition of archaeal promoter elements for different archaeal groups and reveal base-specific TFB contacts upstream of the TATA box. *Mol. Microbiol.* **31**: 1589–1592
- Stazic D & Voß B (2016) The complexity of bacterial transcriptomes. *J. Biotechnol.* **232**: 69–78
- Strahl H & Greie JC (2008) The extremely halophilic archaeon *Halobacterium salinarum* R1 responds to potassium limitation by expression of the K⁺-transporting KdpFABC P-type ATPase and by a decrease in intracellular K⁺. *Extremophiles* **12**: 741–752
- Su M, Ling Y, Yu J, Wu J & Xiao J (2013) Small proteins: Untapped area of potential biological importance. *Front. Genet.* **4**: 1–9
- Taft RJ, Glazov E a, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest ARR, Grimmond SM, Schroder K, Irvine K, Arakawa T, Nakamura M, Kubosaki A, Hayashida K, Kawazu C, Murata M, Nishiyori H, Fukuda S, Kawai J, et al. (2009) Tiny RNAs associated with transcription start sites in animals. *Nat. Genet.* **41**: 572–8
- Takayanagi S, Morimura S, Kusaoke H, Yokoyama Y, Kano K & Shioda M (1992) Chromosomal structure of the halophilic archaeobacterium *Halobacterium salinarum*. *J. Bacteriol.* **174**: 7207–7216
- Tanay A (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**: 962–972
- Tang TH, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachellerie JP & Hüttenhofer A (2005) Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* **55**: 469–481
- Tarasov VY, Besir H, Schwaiger R, Klee K, Furtwängler K, Pfeiffer F & Oesterhelt D (2008) A small protein from the bop-brp intergenic region of *Halobacterium salinarum* contains a zinc finger motif and regulates bop and crtB1 transcription. *Mol. Microbiol.* **67**: 772–780
- Teichmann M, Dumay-Odelot H & Fribourg S (2012) Structural and functional aspects of winged-helix domains at the core of transcription initiation complexes. *Transcription* **3**: 2–7
- Thomas NA & Jarrell KF (2001) Characterization of Flagellum Gene Families of Methanogenic Archaea and Localization of Novel Flagellum Accessory Proteins. *J. Bacteriol.* **183**: 7154–7164
- Thomason MK, Bischler T, Eisenbart SK, Förstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM & Storza G (2015) Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J. Bacteriol.* **197**: 18–28
- Toffano-Nioche C, Ott A, Crozat E, Nguyen AN, Zytnicki M, Leclerc F, Forterre P, Boulloc P & Gautheret D (2013) RNA at 92 degrees C The non-coding transcriptome of the hyperthermophilic archaeon *Pyrococcus abyssi*. *RNA Biol* **10**: 1211–1220
- Turkarslan S, Reiss DJ, Gibbins G, Su WL, Pan M, Bare JC, Plaisier CL & Baliga NS (2014) Niche adaptation by expansion and reprogramming of general transcription factors. *Mol. Syst. Biol.* **7**: 554–554
- Van PT, Schmid AK, King NL, Kaur A, Pan M, Whitehead K, Koide T, Facciotti MT, Goo YA, Deutsch EW, Reiss DJ, Mallick P & Baliga NS (2008) *Halobacterium salinarum* NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage. *J. Proteome Res.* **7**: 3755–3764
- Venter J & et al (2001) The sequence of the human genome. *Science (80-.).* **291**: 1304

- Villar E, Klopp C, Noirot C, Novaes E, Kirst M, Plomion C & Gion J-M (2011) RNA-Seq reveals genotype-specific molecular responses to water deficit in eucalyptus. *BMC Genomics* **12**: 538
- Wachter J & Hill SA (2015) Small transcriptome analysis indicates that the enzyme RppH influences both the quality and quantity of sRNAs in *Neisseria gonorrhoeae*. *FEMS Microbiol. Lett.* **362**: 1–7
- Wade JT & Grainger DC (2014) Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.* **12**: 647–653
- Wah DA, Hirsch JA, Dorner LF, Schildkraut I & Aggarwal AK (1997) Structure of the multimodular endonuclease FokI bound to DNA. *Nature* **388**: 97–100
- Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, Samudrala R, Yu J & Wong GK-S (2004) Mouse transcriptome: Neutral evolution of “non-coding” complementary DNAs. *Nature* **431**: 14–15
- Wang Z, Gerstein MB & Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**: 57–63
- Waters LS & Storz G (2009) Regulatory RNAs in Bacteria. *Cell* **136**: 615–28
- Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562
- White MF & Bell SD (2002) Holding it together: Chromatin in the Archaea. *Trends Genet.* **18**: 621–626
- Whitfield WG, Chaplin M a, Oegema K, Parry H & Glover DM (1995) The 190 kDa centrosome-associated protein of *Drosophila melanogaster* contains four zinc finger motifs and binds to specific sites on polytene chromosomes. *J. Cell Sci.* **108** (Pt 1: 3377–3387
- Wilbanks EG, Larsen DJ, Neches RY, Yao AI, Wu CY, Kjolby RAS & Facciotti MT (2012) A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq. *Nucleic Acids Res.* **40**:
- Woese CR & Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**: 5088–5090
- Woese CR, Kandler O & Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**: 4576–4579
- Wurtmann EJ, Ratushny A V., Pan M, Beer KD, Aitchison JD & Baliga NS (2014) An evolutionarily conserved RNase-based mechanism for repression of transcriptional positive autoregulation. *Mol. Microbiol.* **92**: 369–382
- Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons B a & Sorek R (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res.* **20**: 133–41
- Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Bécavin C, Archambaud C, Cossart P & Sorek R (2012) Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. *Mol. Syst. Biol.* **8**:
- Yamada K, Miyata T, Tsuchiya D, Oyama T, Fujiwara Y, Ohnishi T, Iwasaki H, Shinagawa H, Ariyoshi M, Mayanagi K & Morikawa K (2002) Crystal structure of the RuvA-RuvB complex: A structural basis for the holliday junction migrating motor machinery. *Mol. Cell* **10**: 671–681
- Yus E, Güell M, Vivancos AP, Chen W-H, Lluch-Senar M, Delgado J, Gavin A-C, Bork P & Serrano L (2012) Transcription start site associated RNAs in bacteria. *Mol. Syst. Biol.* **8**: 585

- Yutin N, Makarova KS, Mekhedov SL, Wolf YI & Koonin E V (2008) The deep archaeal roots of eukaryotes. *Mol. Biol. Evol.* **25**: 1619–30
- Zago M a, Dennis PP & Omer a D (2005) The expanding world of small RNAs in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* **55**: 1812–1828
- Zalacain M, Biswas S, Ingraham KA, Ambrad J, Bryant A, Chalker AF, Iordanescu S, Fan J, Fan F, Lunsford RD, O'Dwyer K, Palmer LM, So C, Sylvester D, Volker C, Warren P, McDevitt D, Brown JR, Holmes DJ & Burnham MKR (2003) A global approach to identify novel broad-spectrum antibacterial targets among proteins of unknown function. *J. Mol. Microbiol. Biotechnol.* **6**: 109–126
- Zaramela LS, Vêncio RZN, Ten-Caten F, Baliga NS & Koide T (2014) Transcription Start Site Associated RNAs (TSSaRNAs) Are Ubiquitous in All Domains of Life. *PLoS One* **9**: e107680.
- Zhang T, Datta S, Eichler J, Ivanova N, Axen SD, Kerfeld C a., Chen F, Kyrpides N, Hugenholtz P, Cheng J-F, Sale KL, Simmons B & Rubin E (2011) Identification of a haloalkaliphilic and thermostable cellulase with improved ionic liquid tolerance. *Green Chem.* **13**: 2083
- Zhang Z (2005) Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE. *Nucleic Acids Res.* **33**: 2838–2851
- Zhu X, Wirén M, Sinha I, Rasmussen NN, Linder T, Holmberg S, Ekwall K & Gustafsson CM (2006) Genome-Wide Occupancy Profile of Mediator and the Srb8-11 Module Reveals Interactions with Coding Regions. *Mol. Cell* **22**: 169–178

Anexos

Anexo A - 4540 inícios de transcrição identificados no genoma de H. salinarum NRC-1 através de dRNA-seq

Disponível em: <http://labisismi.fmrp.usp.br/index.php/br/anexos-tese-ftencaten>

Anexo B: asRNAs com expressão correlacionada com os genes na fita oposta em 7 diferentes condições

Disponível em: <http://labisismi.fmrp.usp.br/index.php/br/anexos-tese-ftencaten>

Anexo C: oTSSs upstream a genes descontinuados

Cromossomo	Posição	Orientação	Gene descontinuado	Nome
chr	12934	forward	VNG0014C	TSS-14
chr	15310	forward	VNG0019H	TSS-18
chr	29010	forward	VNG0032H	TSS-45
chr	29048	forward	VNG0032H	TSS-46
chr	29079	forward	VNG0032H	TSS-47
chr	29101	forward	VNG0032H	TSS-48
chr	29226	forward	VNG0032H	TSS-49
chr	29247	forward	VNG0032H	TSS-50
chr	29274	forward	VNG0032H	TSS-51
chr	41730	forward	VNG0049H	TSS-77
chr	41794	forward	VNG0049H	TSS-78
chr	41880	forward	VNG0049H	TSS-79
chr	41937	forward	VNG0049H	TSS-80
chr	42632	forward	VNG0049H	TSS-81
chr	43433	forward	VNG0050C	TSS-84
chr	43682	forward	VNG0050C	TSS-85
chr	44138	forward	VNG0050C	TSS-86
chr	116228	forward	VNG0139H	TSS-201
chr	116422	forward	VNG0139H	TSS-202
chr	116702	forward	VNG0139H	TSS-203
chr	177587	forward	VNG0212H	TSS-255
chr	177614	forward	VNG0212H	TSS-256
chr	177629	forward	VNG0212H	TSS-257
chr	177693	forward	VNG0212H	TSS-258
chr	394049	forward	VNG0509H	TSS-447
chr	424253	forward	VNG0553C	TSS-469
chr	635605	forward	VNG0836H	TSS-660
chr	635903	forward	VNG0836H	TSS-661
chr	651889	forward	VNG0863H	TSS-676
chr	753195	forward	VNG0987H	TSS-756
chr	753221	forward	VNG0987H	TSS-757
chr	753480	forward	VNG0987H	TSS-758
chr	754016	forward	VNG0988H	TSS-762
chr	754267	forward	VNG0988H	TSS-763
chr	754326	forward	VNG0988H	TSS-764
chr	770925	forward	VNG1006H	TSS-788
chr	771276	forward	VNG1006H	TSS-789
chr	1105675	forward	VNG1486H	TSS-1095
chr	1189718	forward	VNG1591H	TSS-1155
chr	1189938	forward	VNG1591H	TSS-1156
chr	1189961	forward	VNG1591H	TSS-1157
chr	1458992	forward	VNG1977H	TSS-1396
chr	1459030	forward	VNG1977H	TSS-1397
chr	1459111	forward	VNG1977H	TSS-1398
chr	1459473	forward	VNG1977H	TSS-1399
chr	1459724	forward	VNG1977H	TSS-1400
chr	1460102	forward	VNG1978H	TSS-1401
chr	1460248	forward	VNG1978H	TSS-1402
chr	1460323	forward	VNG1978H	TSS-1403
chr	1789748	forward	VNG2385H	TSS-1660
chr	1789770	forward	VNG2385H	TSS-1661
chr	1789805	forward	VNG2385H	TSS-1662
chr	1833647	forward	VNG2442H	TSS-1691
chr	1833667	forward	VNG2442H	TSS-1692
chr	1833725	forward	VNG2442H	TSS-1693
chr	1833745	forward	VNG2442H	TSS-1694
chr	1835896	forward	VNG2446H	TSS-1696
chr	1838364	forward	VNG2451H	TSS-1698
chr	1838380	forward	VNG2451H	TSS-1699
chr	1960360	forward	VNG2619H	TSS-1888
chr	1960426	forward	VNG2619H	TSS-1889
chr	22976	reverse	VNG0027H	TSS-2004
chr	57934	reverse	VNG0060G	TSS-2093
chr	58115	reverse	VNG0060G	TSS-2094
chr	58186	reverse	VNG0060G	TSS-2095
chr	58375	reverse	VNG0060G	TSS-2096
chr	58453	reverse	VNG0060G	TSS-2097
chr	58610	reverse	VNG0060G	TSS-2098
chr	58659	reverse	VNG0060G	TSS-2099
chr	58734	reverse	VNG0060G	TSS-2100

chr	58768	reverse	VNG0060G	TSS-2101
chr	69354	reverse	VNG0076H	TSS-2114
chr	103077	reverse	VNG0119H	TSS-2156
chr	177466	reverse	VNG0210H	TSS-2231
chr	179718	reverse	VNG0214C	TSS-2238
chr	224350	reverse	VNG0280H	TSS-2282
chr	309685	reverse	VNG0402H	TSS-2356
chr	409132	reverse	VNG0529H	TSS-2428
chr	460171	reverse	VNGr01	TSS-2476
chr	460192	reverse	VNGr01	TSS-2477
chr	576015	reverse	VNG0762H	TSS-2570
chr	712447	reverse	VNG0934H	TSS-2656
chr	772116	reverse	VNG1007H	TSS-2744
chr	796389	reverse	VNG1047H	TSS-2759
chr	796406	reverse	VNG1047H	TSS-2760
chr	953554	reverse	VNG1270H	TSS-2919
chr	953676	reverse	VNG1270H	TSS-2920
chr	954198	reverse	VNG1270H	TSS-2921
chr	954315	reverse	VNG1270H	TSS-2922
chr	1026759	reverse	VNG1376H	TSS-2980
chr	1143724	reverse	VNG1534H	TSS-3076
chr	1221094	reverse	VNG1642H	TSS-3139
chr	1233217	reverse	VNG1654G	TSS-3149
chr	1233273	reverse	VNG1654G	TSS-3150
chr	1233568	reverse	VNG1654G	TSS-3151
chr	1233817	reverse	VNG1654G	TSS-3152
chr	1233841	reverse	VNG1654G	TSS-3153
chr	1397994	reverse	VNG1890H	TSS-3280
chr	1566074	reverse	VNG2129H	TSS-3446
chr	1607770	reverse	VNG2174H	TSS-3486
chr	1607794	reverse	VNG2174H	TSS-3487
chr	1607933	reverse	VNG2174H	TSS-3488
chr	1608100	reverse	VNG2174H	TSS-3489
chr	1609547	reverse	VNG2176H	TSS-3490
chr	1609563	reverse	VNG2176H	TSS-3491
chr	1609847	reverse	VNG2176H	TSS-3492
chr	1609864	reverse	VNG2176H	TSS-3493
chr	1609919	reverse	VNG2176H	TSS-3494
chr	1609964	reverse	VNG2176H	TSS-3495
chr	1609982	reverse	VNG2176H	TSS-3496
chr	1610015	reverse	VNG2176H	TSS-3497
chr	1610033	reverse	VNG2176H	TSS-3498
chr	1610516	reverse	VNG2177H	TSS-3499
plasmid_pNRC200	133165	forward	VNG6156H	TSS-3899
plasmid_pNRC200	133202	forward	VNG6156H	TSS-3900
plasmid_pNRC200	143759	forward	VNG6171H	TSS-3912
plasmid_pNRC200	162134	forward	VNG6194H	TSS-3941
plasmid_pNRC200	163771	forward	VNG6197H	TSS-3943
plasmid_pNRC200	164127	forward	VNG6198H	TSS-3944
plasmid_pNRC200	219774	forward	VNG6287H	TSS-3984
plasmid_pNRC200	219797	forward	VNG6287H	TSS-3985
plasmid_pNRC200	228687	forward	VNG6297C	TSS-4000
plasmid_pNRC200	267433	forward	VNG6334H	TSS-4035
plasmid_pNRC200	278540	forward	VNG6347H	TSS-4047
plasmid_pNRC200	328139	forward	VNG6432H	TSS-4095
plasmid_pNRC200	135966	reverse	VNG6157H	TSS-4121
plasmid_pNRC200	136724	reverse	VNG6159H	TSS-4122
plasmid_pNRC200	139913	reverse	VNG6164G	TSS-4126
plasmid_pNRC200	205842	reverse	VNG6266H	TSS-4193
plasmid_pNRC200	206051	reverse	VNG6266H	TSS-4194
plasmid_pNRC200	206069	reverse	VNG6266H	TSS-4195
plasmid_pNRC200	213060	reverse	VNG6276H	TSS-4203
plasmid_pNRC200	222535	reverse	VNG6291H	TSS-4211
plasmid_pNRC200	222757	reverse	VNG6291H	TSS-4212
plasmid_pNRC200	222827	reverse	VNG6291H	TSS-4213
plasmid_pNRC200	223545	reverse	VNG6292C	TSS-4214
plasmid_pNRC200	264330	reverse	VNG6329H	TSS-4271
plasmid_pNRC200	264349	reverse	VNG6329H	TSS-4272
plasmid_pNRC200	264368	reverse	VNG6329H	TSS-4273
plasmid_pNRC200	264391	reverse	VNG6329H	TSS-4274
plasmid_pNRC200	264412	reverse	VNG6329H	TSS-4275
plasmid_pNRC200	264949	reverse	VNG6329H	TSS-4276
plasmid_pNRC200	265015	reverse	VNG6329H	TSS-4277
plasmid_pNRC200	265058	reverse	VNG6329H	TSS-4278

plasmid_pNRC200	276249	reverse	VNG6344H	TSS-4296
plasmid_pNRC200	276343	reverse	VNG6344H	TSS-4297
plasmid_pNRC200	293831	reverse	VNG6364H	TSS-4317
plasmid_pNRC200	294041	reverse	VNG6364H	TSS-4318
plasmid_pNRC200	317263	reverse	VNG6407H	TSS-4332
plasmid_pNRC200	326531	reverse	VNG6429H	TSS-4345
plasmid_pNRC200	326826	reverse	VNG6429H	TSS-4346
plasmid_pNRC200	326968	reverse	VNG6429H	TSS-4347
plasmid_pNRC200	326987	reverse	VNG6429H	TSS-4348
plasmid_pNRC200	331536	reverse	VNG6439H	TSS-4354
plasmid_pNRC200	331579	reverse	VNG6439H	TSS-4355
plasmid_pNRC100	127949	forward	VNG7113	TSS-4399
plasmid_pNRC100	140392	forward	VNG7125	TSS-4424
plasmid_pNRC100	140599	forward	VNG7125	TSS-4425
plasmid_pNRC100	146377	forward	VNG7133	TSS-4440

Anexo D: oTSSs antisense a regiões 3' UTRs

Cromossomo	Início	Fim	Orientação	Nome
chr	11462	11762	forward	TSS-11
chr	13393	13693	forward	TSS-15
chr	21823	22123	forward	TSS-31
chr	22091	22391	forward	TSS-32
chr	24111	24411	forward	TSS-34
chr	24137	24437	forward	TSS-35
chr	24151	24451	forward	TSS-36
chr	24507	24807	forward	TSS-37
chr	29649	29949	forward	TSS-52
chr	29677	29977	forward	TSS-53
chr	32416	32716	forward	TSS-59
chr	32950	33250	forward	TSS-60
chr	47153	47453	forward	TSS-93
chr	49324	49624	forward	TSS-98
chr	54653	54953	forward	TSS-108
chr	57659	57959	forward	TSS-114
chr	57978	58278	forward	TSS-115
chr	58246	58546	forward	TSS-116
chr	58376	58676	forward	TSS-117
chr	58470	58770	forward	TSS-118
chr	58516	58816	forward	TSS-119
chr	60427	60727	forward	TSS-123
chr	60546	60846	forward	TSS-124
chr	69255	69555	forward	TSS-142
chr	69268	69568	forward	TSS-143
chr	95060	95360	forward	TSS-175
chr	102730	103030	forward	TSS-185
chr	103150	103450	forward	TSS-186
chr	103223	103523	forward	TSS-187
chr	124955	125255	forward	TSS-212
chr	126162	126462	forward	TSS-213
chr	179482	179782	forward	TSS-262
chr	179908	180208	forward	TSS-263
chr	180665	180965	forward	TSS-267
chr	180709	181009	forward	TSS-268
chr	183857	184157	forward	TSS-275
chr	196697	196997	forward	TSS-290
chr	230623	230923	forward	TSS-330
chr	257868	258168	forward	TSS-351
chr	307249	307549	forward	TSS-387
chr	329542	329842	forward	TSS-409
chr	382456	382756	forward	TSS-438
chr	384482	384782	forward	TSS-440
chr	413435	413735	forward	TSS-466
chr	417812	418112	forward	TSS-467
chr	417886	418186	forward	TSS-468
chr	449394	449694	forward	TSS-489
chr	471604	471904	forward	TSS-516
chr	501128	501428	forward	TSS-540
chr	580294	580594	forward	TSS-616
chr	580315	580615	forward	TSS-617
chr	580344	580644	forward	TSS-618
chr	580359	580659	forward	TSS-619
chr	580379	580679	forward	TSS-620
chr	580402	580702	forward	TSS-621
chr	680447	680747	forward	TSS-698
chr	728894	729194	forward	TSS-721
chr	751841	752141	forward	TSS-752
chr	756547	756847	forward	TSS-766
chr	759419	759719	forward	TSS-773
chr	811031	811331	forward	TSS-843
chr	811111	811411	forward	TSS-844
chr	916139	916439	forward	TSS-933
chr	920081	920381	forward	TSS-935
chr	925394	925694	forward	TSS-939
chr	925417	925717	forward	TSS-940
chr	953614	953914	forward	TSS-963
chr	953645	953945	forward	TSS-964
chr	954154	954454	forward	TSS-965
chr	963988	964288	forward	TSS-973

chr	983890	984190	forward	TSS-989
chr	983924	984224	forward	TSS-990
chr	990595	990895	forward	TSS-992
chr	990616	990916	forward	TSS-993
chr	990639	990939	forward	TSS-994
chr	990748	991048	forward	TSS-995
chr	1026789	1027089	forward	TSS-1017
chr	1026832	1027132	forward	TSS-1018
chr	1060250	1060550	forward	TSS-1044
chr	1060288	1060588	forward	TSS-1045
chr	1060315	1060615	forward	TSS-1046
chr	1060331	1060631	forward	TSS-1047
chr	1068607	1068907	forward	TSS-1059
chr	1068627	1068927	forward	TSS-1060
chr	1128053	1128353	forward	TSS-1109
chr	1154201	1154501	forward	TSS-1127
chr	1154231	1154531	forward	TSS-1128
chr	1154249	1154549	forward	TSS-1129
chr	1154270	1154570	forward	TSS-1130
chr	1154287	1154587	forward	TSS-1131
chr	1204646	1204946	forward	TSS-1168
chr	1229342	1229642	forward	TSS-1190
chr	1232857	1233157	forward	TSS-1195
chr	1232987	1233287	forward	TSS-1196
chr	1233545	1233845	forward	TSS-1197
chr	1279674	1279974	forward	TSS-1251
chr	1281626	1281926	forward	TSS-1253
chr	1281670	1281970	forward	TSS-1254
chr	1304013	1304313	forward	TSS-1274
chr	1307217	1307517	forward	TSS-1275
chr	1309301	1309601	forward	TSS-1276
chr	1311911	1312211	forward	TSS-1277
chr	1333436	1333736	forward	TSS-1293
chr	1356471	1356771	forward	TSS-1320
chr	1397366	1397666	forward	TSS-1346
chr	1421195	1421495	forward	TSS-1376
chr	1495213	1495513	forward	TSS-1421
chr	1508899	1509199	forward	TSS-1436
chr	1553841	1554141	forward	TSS-1471
chr	1555327	1555627	forward	TSS-1473
chr	1556160	1556460	forward	TSS-1474
chr	1565954	1566254	forward	TSS-1480
chr	1609564	1609864	forward	TSS-1510
chr	1609734	1610034	forward	TSS-1511
chr	1609898	1610198	forward	TSS-1512
chr	1610028	1610328	forward	TSS-1513
chr	1610539	1610839	forward	TSS-1515
chr	1627093	1627393	forward	TSS-1531
chr	1658519	1658819	forward	TSS-1556
chr	1779792	1780092	forward	TSS-1654
chr	1791659	1791959	forward	TSS-1663
chr	1804650	1804950	forward	TSS-1671
chr	1853330	1853630	forward	TSS-1708
chr	1853362	1853662	forward	TSS-1709
chr	1864948	1865248	forward	TSS-1715
chr	1882377	1882677	forward	TSS-1793
chr	1907560	1907860	forward	TSS-1821
chr	1921264	1921564	forward	TSS-1835
chr	1926813	1927113	forward	TSS-1842
chr	1937219	1937519	forward	TSS-1859
chr	1937256	1937556	forward	TSS-1860
chr	1937297	1937597	forward	TSS-1861
chr	1952631	1952931	forward	TSS-1878
chr	1952651	1952951	forward	TSS-1879
chr	1987773	1988073	forward	TSS-1917
chr	2002621	2002921	forward	TSS-1933
chr	11448	11148	reverse	TSS-1976
chr	11469	11169	reverse	TSS-1977
chr	17723	17423	reverse	TSS-1990
chr	17770	17470	reverse	TSS-1991
chr	17928	17628	reverse	TSS-1992
chr	17971	17671	reverse	TSS-1993
chr	23980	23680	reverse	TSS-2007
chr	29160	28860	reverse	TSS-2017

chr	29253	28953	reverse	TSS-2018
chr	29293	28993	reverse	TSS-2019
chr	32400	32100	reverse	TSS-2024
chr	38046	37746	reverse	TSS-2036
chr	38061	37761	reverse	TSS-2037
chr	42031	41731	reverse	TSS-2045
chr	42354	42054	reverse	TSS-2046
chr	42568	42268	reverse	TSS-2047
chr	42629	42329	reverse	TSS-2048
chr	42948	42648	reverse	TSS-2049
chr	42969	42669	reverse	TSS-2050
chr	43050	42750	reverse	TSS-2051
chr	43068	42768	reverse	TSS-2052
chr	43437	43137	reverse	TSS-2054
chr	43563	43263	reverse	TSS-2055
chr	43921	43621	reverse	TSS-2056
chr	44059	43759	reverse	TSS-2057
chr	49326	49026	reverse	TSS-2066
chr	50457	50157	reverse	TSS-2072
chr	53088	52788	reverse	TSS-2079
chr	53152	52852	reverse	TSS-2080
chr	54538	54238	reverse	TSS-2085
chr	54559	54259	reverse	TSS-2086
chr	64870	64570	reverse	TSS-2108
chr	65959	65659	reverse	TSS-2109
chr	66641	66341	reverse	TSS-2111
chr	68570	68270	reverse	TSS-2113
chr	94485	94185	reverse	TSS-2146
chr	94554	94254	reverse	TSS-2147
chr	94570	94270	reverse	TSS-2148
chr	99356	99056	reverse	TSS-2151
chr	108684	108384	reverse	TSS-2163
chr	116187	115887	reverse	TSS-2172
chr	116301	116001	reverse	TSS-2173
chr	116393	116093	reverse	TSS-2174
chr	116461	116161	reverse	TSS-2175
chr	116520	116220	reverse	TSS-2176
chr	116574	116274	reverse	TSS-2177
chr	119073	118773	reverse	TSS-2183
chr	167506	167206	reverse	TSS-2222
chr	177819	177519	reverse	TSS-2232
chr	177962	177662	reverse	TSS-2233
chr	178103	177803	reverse	TSS-2234
chr	178228	177928	reverse	TSS-2235
chr	179368	179068	reverse	TSS-2236
chr	179392	179092	reverse	TSS-2237
chr	223359	223059	reverse	TSS-2281
chr	267277	266977	reverse	TSS-2317
chr	409814	409514	reverse	TSS-2429
chr	409994	409694	reverse	TSS-2430
chr	445236	444936	reverse	TSS-2456
chr	445258	444958	reverse	TSS-2457
chr	478447	478147	reverse	TSS-2498
chr	495576	495276	reverse	TSS-2509
chr	522312	522012	reverse	TSS-2528
chr	522329	522029	reverse	TSS-2529
chr	522361	522061	reverse	TSS-2530
chr	522409	522109	reverse	TSS-2531
chr	522423	522123	reverse	TSS-2532
chr	522444	522144	reverse	TSS-2533
chr	522491	522191	reverse	TSS-2534
chr	522511	522211	reverse	TSS-2535
chr	522534	522234	reverse	TSS-2536
chr	522548	522248	reverse	TSS-2537
chr	537950	537650	reverse	TSS-2543
chr	568960	568660	reverse	TSS-2563
chr	579785	579485	reverse	TSS-2572
chr	580251	579951	reverse	TSS-2573
chr	591281	590981	reverse	TSS-2583
chr	620364	620064	reverse	TSS-2596
chr	635511	635211	reverse	TSS-2605
chr	635794	635494	reverse	TSS-2606
chr	635924	635624	reverse	TSS-2607
chr	678490	678190	reverse	TSS-2629

chr	678562	678262	reverse	TSS-2630
chr	734896	734596	reverse	TSS-2684
chr	748896	748596	reverse	TSS-2699
chr	751546	751246	reverse	TSS-2702
chr	752940	752640	reverse	TSS-2707
chr	753166	752866	reverse	TSS-2708
chr	753612	753312	reverse	TSS-2709
chr	753847	753547	reverse	TSS-2710
chr	754272	753972	reverse	TSS-2712
chr	754359	754059	reverse	TSS-2713
chr	759418	759118	reverse	TSS-2727
chr	770821	770521	reverse	TSS-2739
chr	771244	770944	reverse	TSS-2740
chr	794452	794152	reverse	TSS-2756
chr	806459	806159	reverse	TSS-2775
chr	833961	833661	reverse	TSS-2806
chr	853136	852836	reverse	TSS-2829
chr	860265	859965	reverse	TSS-2841
chr	889900	889600	reverse	TSS-2865
chr	890563	890263	reverse	TSS-2866
chr	916278	915978	reverse	TSS-2890
chr	955344	955044	reverse	TSS-2926
chr	982986	982686	reverse	TSS-2953
chr	1023043	1022743	reverse	TSS-2979
chr	1043513	1043213	reverse	TSS-2998
chr	1043619	1043319	reverse	TSS-2999
chr	1055298	1054998	reverse	TSS-3006
chr	1089075	1088775	reverse	TSS-3032
chr	1089100	1088800	reverse	TSS-3033
chr	1183143	1182843	reverse	TSS-3092
chr	1185871	1185571	reverse	TSS-3099
chr	1186009	1185709	reverse	TSS-3100
chr	1186031	1185731	reverse	TSS-3101
chr	1199914	1199614	reverse	TSS-3118
chr	1218497	1218197	reverse	TSS-3134
chr	1232547	1232247	reverse	TSS-3148
chr	1279366	1279066	reverse	TSS-3179
chr	1296954	1296654	reverse	TSS-3192
chr	1342995	1342695	reverse	TSS-3239
chr	1398200	1397900	reverse	TSS-3281
chr	1457330	1457030	reverse	TSS-3331
chr	1459031	1458731	reverse	TSS-3334
chr	1459050	1458750	reverse	TSS-3335
chr	1459682	1459382	reverse	TSS-3336
chr	1459932	1459632	reverse	TSS-3337
chr	1460448	1460148	reverse	TSS-3338
chr	1495328	1495028	reverse	TSS-3373
chr	1538814	1538514	reverse	TSS-3416
chr	1541981	1541681	reverse	TSS-3420
chr	1584052	1583752	reverse	TSS-3474
chr	1607677	1607377	reverse	TSS-3485
chr	1646118	1645818	reverse	TSS-3528
chr	1678559	1678259	reverse	TSS-3558
chr	1693133	1692833	reverse	TSS-3574
chr	1765393	1765093	reverse	TSS-3626
chr	1765426	1765126	reverse	TSS-3627
chr	1833825	1833525	reverse	TSS-3689
chr	1838390	1838090	reverse	TSS-3695
chr	1877024	1876724	reverse	TSS-3734
chr	1880719	1880419	reverse	TSS-3742
chr	1924043	1923743	reverse	TSS-3777
chr	1924090	1923790	reverse	TSS-3778
chr	1926530	1926230	reverse	TSS-3781
chr	1926558	1926258	reverse	TSS-3782
chr	1930366	1930066	reverse	TSS-3787
chr	1932493	1932193	reverse	TSS-3788
chr	1959158	1958858	reverse	TSS-3814
chr	1960473	1960173	reverse	TSS-3816
chr	1984229	1983929	reverse	TSS-3830
chr	1989672	1989372	reverse	TSS-3842
plasmid_pNRC200	135810	136110	forward	TSS-3903
plasmid_pNRC200	136366	136666	forward	TSS-3904
plasmid_pNRC200	136386	136686	forward	TSS-3905
plasmid_pNRC200	141136	141436	forward	TSS-3909

plasmid_pNRC200	157354	157654	forward	TSS-3932
plasmid_pNRC200	167031	167331	forward	TSS-3945
plasmid_pNRC200	219368	219668	forward	TSS-3983
plasmid_pNRC200	223350	223650	forward	TSS-3993
plasmid_pNRC200	223370	223670	forward	TSS-3994
plasmid_pNRC200	227270	227570	forward	TSS-3999
plasmid_pNRC200	254351	254651	forward	TSS-4018
plasmid_pNRC200	254497	254797	forward	TSS-4019
plasmid_pNRC200	258438	258738	forward	TSS-4022
plasmid_pNRC200	258467	258767	forward	TSS-4023
plasmid_pNRC200	264284	264584	forward	TSS-4029
plasmid_pNRC200	264599	264899	forward	TSS-4030
plasmid_pNRC200	264613	264913	forward	TSS-4031
plasmid_pNRC200	268330	268630	forward	TSS-4037
plasmid_pNRC200	272936	273236	forward	TSS-4041
plasmid_pNRC200	275737	276037	forward	TSS-4045
plasmid_pNRC200	275755	276055	forward	TSS-4046
plasmid_pNRC200	280381	280681	forward	TSS-4050
plasmid_pNRC200	283114	283414	forward	TSS-4054
plasmid_pNRC200	283147	283447	forward	TSS-4055
plasmid_pNRC200	293809	294109	forward	TSS-4063
plasmid_pNRC200	299189	299489	forward	TSS-4070
plasmid_pNRC200	301470	301770	forward	TSS-4073
plasmid_pNRC200	315133	315433	forward	TSS-4082
plasmid_pNRC200	325552	325852	forward	TSS-4091
plasmid_pNRC200	326039	326339	forward	TSS-4092
plasmid_pNRC200	326478	326778	forward	TSS-4093
plasmid_pNRC200	327772	328072	forward	TSS-4094
plasmid_pNRC200	332551	332851	forward	TSS-4103
plasmid_pNRC200	332669	332969	forward	TSS-4104
plasmid_pNRC200	132389	132089	reverse	TSS-4117
plasmid_pNRC200	133494	133194	reverse	TSS-4118
plasmid_pNRC200	144371	144071	reverse	TSS-4133
plasmid_pNRC200	153083	152783	reverse	TSS-4146
plasmid_pNRC200	153785	153485	reverse	TSS-4147
plasmid_pNRC200	153807	153507	reverse	TSS-4148
plasmid_pNRC200	159780	159480	reverse	TSS-4162
plasmid_pNRC200	169287	168987	reverse	TSS-4174
plasmid_pNRC200	196617	196317	reverse	TSS-4184
plasmid_pNRC200	208781	208481	reverse	TSS-4197
plasmid_pNRC200	208883	208583	reverse	TSS-4198
plasmid_pNRC200	217638	217338	reverse	TSS-4207
plasmid_pNRC200	222510	222210	reverse	TSS-4210
plasmid_pNRC200	228764	228464	reverse	TSS-4221
plasmid_pNRC200	228901	228601	reverse	TSS-4222
plasmid_pNRC200	229196	228896	reverse	TSS-4223
plasmid_pNRC200	262641	262341	reverse	TSS-4267
plasmid_pNRC200	264112	263812	reverse	TSS-4270
plasmid_pNRC200	301431	301131	reverse	TSS-4325
plasmid_pNRC200	328345	328045	reverse	TSS-4350
plasmid_pNRC200	328699	328399	reverse	TSS-4351
plasmid_pNRC200	329650	329350	reverse	TSS-4352
plasmid_pNRC200	329679	329379	reverse	TSS-4353
plasmid_pNRC100	114954	115254	forward	TSS-4367
plasmid_pNRC100	114973	115273	forward	TSS-4368
plasmid_pNRC100	114999	115299	forward	TSS-4369
plasmid_pNRC100	115018	115318	forward	TSS-4370
plasmid_pNRC100	131019	131319	forward	TSS-4409
plasmid_pNRC100	135739	136039	forward	TSS-4418
plasmid_pNRC100	137155	137455	forward	TSS-4420
plasmid_pNRC100	137186	137486	forward	TSS-4421
plasmid_pNRC100	142494	142794	forward	TSS-4429
plasmid_pNRC100	114405	114105	reverse	TSS-4455
plasmid_pNRC100	114919	114619	reverse	TSS-4456
plasmid_pNRC100	114980	114680	reverse	TSS-4457
plasmid_pNRC100	116502	116202	reverse	TSS-4467
plasmid_pNRC100	116545	116245	reverse	TSS-4468
plasmid_pNRC100	116572	116272	reverse	TSS-4469
plasmid_pNRC100	116610	116310	reverse	TSS-4470
plasmid_pNRC100	116734	116434	reverse	TSS-4471
plasmid_pNRC100	120568	120268	reverse	TSS-4477
plasmid_pNRC100	125673	125373	reverse	TSS-4487
plasmid_pNRC100	126965	126665	reverse	TSS-4489
plasmid_pNRC100	127228	126928	reverse	TSS-4490

plasmid_pNRC100	129833	129533	reverse	TSS-4496
plasmid_pNRC100	140586	140286	reverse	TSS-4516
plasmid_pNRC100	140807	140507	reverse	TSS-4517
plasmid_pNRC100	141541	141241	reverse	TSS-4518
plasmid_pNRC100	147033	146733	reverse	TSS-4524
plasmid_pNRC100	147051	146751	reverse	TSS-4525
plasmid_pNRC100	147079	146779	reverse	TSS-4526
plasmid_pNRC100	147096	146796	reverse	TSS-4527
plasmid_pNRC100	149180	148880	reverse	TSS-4534
plasmid_pNRC100	149192	148892	reverse	TSS-4535
plasmid_pNRC100	149210	148910	reverse	TSS-4536

Anexo E: 196 oTSS não relacionados a outras regiões do genoma de *H. salinarum* NRC-1

Cromossomo	Início	Fim	Orientação	Condição diferencialmente expresso	Nome
chr	17751	18051	forward	Δ lsm	TSS-23
chr	17789	18089	forward	Δ lsm	TSS-24
chr	17832	18132	forward	Curva de crescimento ; Δ lsm	TSS-25
chr	17943	18243	forward	Curva de crescimento; Δ lsm	TSS-26
chr	25151	25451	forward	Δ VNG2099 (RNase)	TSS-38
chr	34133	34433	forward	-	TSS-61
chr	35982	36282	forward	-	TSS-64
chr	45468	45768	forward	-	TSS-91
chr	45546	45846	forward	-	TSS-92
chr	68525	68825	forward	-	TSS-140
chr	68930	69230	forward	-	TSS-141
chr	94591	94891	forward	Curva de crescimento ; Δ VNG2099 (RNase)	TSS-174
chr	119176	119476	forward	Curva de crescimento; Variação salinidade; Δ VNG2099 (RNase)	TSS-204
chr	119203	119503	forward	Curva de crescimento; Variação salinidade; Δ VNG2099 (RNase); Estresse oxidativo	TSS-205
chr	119216	119516	forward	Curva de crescimento; Δ VNG2099 (RNase); Estresse oxidativo	TSS-206
chr	119235	119535	forward	Curva de crescimento; Δ VNG2099 (RNase); Estresse oxidativo	TSS-207
chr	179359	179659	forward	Δ VNG2099 (RNase)	TSS-260
chr	179373	179673	forward	Δ VNG2099 (RNase)	TSS-261
chr	180502	180802	forward	Curva de crescimento	TSS-264
chr	180514	180814	forward	Curva de crescimento	TSS-265
chr	180575	180875	forward	Curva de crescimento	TSS-266
chr	184031	184331	forward	Curva de crescimento; Variação salinidade	TSS-276
chr	186763	187063	forward	Δ lsm	TSS-283
chr	453867	454167	forward	Curva de crescimento	TSS-494
chr	547618	547918	forward	Curva de crescimento; Δ lsm	TSS-577
chr	547735	548035	forward	Curva de crescimento; Δ lsm	TSS-578
chr	636122	636422	forward	Δ VNG2099 (RNase)	TSS-662
chr	752827	753127	forward	-	TSS-754
chr	752862	753162	forward	-	TSS-755
chr	753718	754018	forward	Curva de crescimento; Variação salinidade	TSS-759
chr	753837	754137	forward	Variação salinidade	TSS-760
chr	753898	754198	forward	Variação salinidade	TSS-761
chr	758160	758460	forward	Curva de crescimento; Variação salinidade; Δ VNG2099 (RNase)	TSS-768
chr	772219	772519	forward	-	TSS-791
chr	772756	773056	forward	-	TSS-792
chr	1068407	1068707	forward	Curva de crescimento; Δ lsm	TSS-1058
chr	1186064	1186364	forward	Variação salinidade	TSS-1148
chr	1188943	1189243	forward	-	TSS-1152
chr	1229080	1229380	forward	-	TSS-1189
chr	1233861	1234161	forward	-	TSS-1198
chr	1233875	1234175	forward	-	TSS-1199
chr	1233962	1234262	forward	Curva de crescimento; Variação salinidade	TSS-1200
chr	1234008	1234308	forward	Variação salinidade	TSS-1201
chr	1302449	1302749	forward	Variação salinidade	TSS-1266
chr	1303062	1303362	forward	-	TSS-1267
chr	1303253	1303553	forward	-	TSS-1268
chr	1303278	1303578	forward	-	TSS-1269
chr	1303311	1303611	forward	-	TSS-1270
chr	1303545	1303845	forward	Variação salinidade	TSS-1271
chr	1303572	1303872	forward	-	TSS-1272
chr	1303618	1303918	forward	Variação salinidade	TSS-1273
chr	1397976	1398276	forward	Curva de crescimento; Δ VNG2099 (RNase)	TSS-1349
chr	1398156	1398456	forward	Curva de crescimento; Variação salinidade	TSS-1350
chr	1398220	1398520	forward	Curva de crescimento	TSS-1351
chr	1610471	1610771	forward	Variação salinidade	TSS-1514
chr	1633290	1633590	forward	-	TSS-1535
chr	1633418	1633718	forward	-	TSS-1536
chr	1633583	1633883	forward	-	TSS-1537
chr	1633605	1633905	forward	-	TSS-1538
chr	1874589	1874889	forward	Curva de crescimento	TSS-1724
chr	1874652	1874952	forward	Alteração oxigênio; Curva de crescimento; Δ lsm	TSS-1725
chr	1874750	1875050	forward	Alteração oxigênio; Curva de crescimento	TSS-1726
chr	1874875	1875175	forward	Alteração oxigênio; Curva de crescimento; Variação salinidade; Δ VNG2099 (RNase)	TSS-1727
chr	1875128	1875428	forward	Alteração oxigênio; Curva de crescimento; Δ VNG2099 (RNase); Δ lsm	TSS-1728
chr	1957157	1957457	forward	-	TSS-1884

chr	2007558	2007858	forward	-	TSS-1940
chr	2007576	2007876	forward	-	TSS-1941
chr	18133	17833	reverse	-	TSS-1994
chr	18342	18042	reverse	-	TSS-1995
chr	23260	22960	reverse	-	TSS-2005
chr	25255	24955	reverse	-	TSS-2008
chr	25364	25064	reverse	-	TSS-2009
chr	25788	25488	reverse	Curva de crescimento; Variação salinidade	TSS-2010
chr	34279	33979	reverse	Variação salinidade	TSS-2028
chr	34599	34299	reverse	Curva de crescimento; Variação salinidade	TSS-2029
chr	34623	34323	reverse	Curva de crescimento	TSS-2030
chr	34643	34343	reverse	Curva de crescimento	TSS-2031
chr	180664	180364	reverse	Variação salinidade	TSS-2239
chr	180755	180455	reverse	Variação salinidade	TSS-2240
chr	180784	180484	reverse	Variação salinidade	TSS-2241
chr	180799	180499	reverse	Variação salinidade	TSS-2242
chr	180816	180516	reverse	Variação salinidade	TSS-2243
chr	522600	522300	reverse	Curva de crescimento; Variação salinidade	TSS-2538
chr	568171	567871	reverse	Curva de crescimento	TSS-2562
chr	634220	633920	reverse	-	TSS-2601
chr	754043	753743	reverse	Variação salinidade	TSS-2711
chr	756517	756217	reverse	-	TSS-2717
chr	756630	756330	reverse	Curva de crescimento; Variação salinidade	TSS-2718
chr	760152	759852	reverse	Variação salinidade	TSS-2728
chr	760167	759867	reverse	Variação salinidade	TSS-2729
chr	760204	759904	reverse	Variação salinidade	TSS-2730
chr	773152	772852	reverse	-	TSS-2745
chr	773326	773026	reverse	-	TSS-2746
chr	811148	810848	reverse	-	TSS-2785
chr	883085	882785	reverse	Curva de crescimento	TSS-2860
chr	914933	914633	reverse	Variação salinidade; Δ VNG2099 (RNAse)	TSS-2884
chr	1186192	1185892	reverse	Alteração oxigênio; Curva de crescimento; Variação salinidade	TSS-3102
chr	1186240	1185940	reverse	Alteração oxigênio; Curva de crescimento; Variação salinidade	TSS-3103
chr	1186480	1186180	reverse	Variação salinidade	TSS-3104
chr	1229287	1228987	reverse	Variação salinidade	TSS-3142
chr	1229367	1229067	reverse	Curva de crescimento; Variação salinidade	TSS-3143
chr	1229429	1229129	reverse	Curva de crescimento; Variação salinidade	TSS-3144
chr	1234359	1234059	reverse	Variação salinidade	TSS-3154
chr	1279613	1279313	reverse	Variação salinidade	TSS-3180
chr	1279663	1279363	reverse	Variação salinidade; Δ VNG2099 (RNAse)	TSS-3181
chr	1303222	1302922	reverse	-	TSS-3194
chr	1303441	1303141	reverse	Variação salinidade	TSS-3195
chr	1303617	1303317	reverse	Variação salinidade	TSS-3196
chr	1303878	1303578	reverse	-	TSS-3197
chr	1303946	1303646	reverse	-	TSS-3198
chr	1303987	1303687	reverse	-	TSS-3199
chr	1304070	1303770	reverse	Curva de crescimento; Variação salinidade	TSS-3200
chr	1304109	1303809	reverse	Curva de crescimento	TSS-3201
chr	1398534	1398234	reverse	Curva de crescimento; Variação salinidade	TSS-3282
chr	1460916	1460616	reverse	Curva de crescimento	TSS-3339
chr	1610750	1610450	reverse	Curva de crescimento	TSS-3500
chr	1633508	1633208	reverse	-	TSS-3518
chr	1874892	1874592	reverse	Curva de crescimento; Δ lsm	TSS-3727
chr	1875013	1874713	reverse	Δ lsm	TSS-3728
chr	1875139	1874839	reverse	-	TSS-3729
chr	1875353	1875053	reverse	-	TSS-3730
chr	1957391	1957091	reverse	-	TSS-3812
chr	1987865	1987565	reverse	Curva de crescimento; Variação salinidade	TSS-3839
plasmid_pNRC100	115621	115921	forward	Curva de crescimento ; Variação salinidade; Estresse oxidativo; Δ lsm	TSS-4372
plasmid_pNRC100	116557	116857	forward	Variação salinidade	TSS-4379
plasmid_pNRC100	116821	117121	forward	-	TSS-4380
plasmid_pNRC100	123900	124200	forward	Variação salinidade	TSS-4390
plasmid_pNRC100	127025	127325	forward	Variação salinidade	TSS-4394
plasmid_pNRC100	127570	127870	forward	Curva de crescimento; Δ VNG2099 (RNAse)	TSS-4395
plasmid_pNRC100	127870	128170	forward	Curva de crescimento	TSS-4396
plasmid_pNRC100	127892	128192	forward	Curva de crescimento	TSS-4397
plasmid_pNRC100	127917	128217	forward	Curva de crescimento	TSS-4398
plasmid_pNRC100	141359	141659	forward	Variação salinidade	TSS-4428
plasmid_pNRC100	148822	149122	forward	Estresse oxidativo	TSS-4445
plasmid_pNRC100	112874	112574	reverse	-	TSS-4451
plasmid_pNRC100	115778	115478	reverse	Alteração oxigênio; Curva de crescimento; Variação salinidade; gse31308	TSS-4462
plasmid_pNRC100	115796	115496	reverse	Alteração oxigênio; Curva de crescimento; Variação	TSS-4463

				salinidade;gse31308	
plasmid_pNRC100	116820	116520	reverse	-	TSS-4472
plasmid_pNRC100	116919	116619	reverse	Curva de crescimento; Δ lsm	TSS-4473
plasmid_pNRC100	116939	116639	reverse	Curva de crescimento; Δ lsm	TSS-4474
plasmid_pNRC100	121021	120721	reverse	Variação salinidade	TSS-4478
plasmid_pNRC100	123915	123615	reverse	-	TSS-4486
plasmid_pNRC100	127514	127214	reverse	Estresse oxidativo	TSS-4491
plasmid_pNRC100	133910	133610	reverse	Δ VNG2099 (RNase); Estresse oxidativo	TSS-4503
plasmid_pNRC100	140353	140053	reverse	Curva de crescimento; Variação salinidade	TSS-4515
plasmid_pNRC100	149001	148701	reverse	Estresse oxidativo	TSS-4533
plasmid_pNRC200	132461	132761	forward	Δ VNG2099 (RNase)	TSS-3897
plasmid_pNRC200	132508	132808	forward	-	TSS-3898
plasmid_pNRC200	133291	133591	forward	Alteração oxigênio; Curva de crescimento	TSS-3901
plasmid_pNRC200	134868	135168	forward	Curva de crescimento; Δ VNG2099 (RNase)	TSS-3902
plasmid_pNRC200	140903	141203	forward	-	TSS-3908
plasmid_pNRC200	152972	153272	forward	-	TSS-3924
plasmid_pNRC200	153831	154131	forward	Curva de crescimento	TSS-3926
plasmid_pNRC200	159923	160223	forward	Curva de crescimento; Variação salinidade	TSS-3936
plasmid_pNRC200	205006	205306	forward	Curva de crescimento	TSS-3972
plasmid_pNRC200	210084	210384	forward	-	TSS-3976
plasmid_pNRC200	210270	210570	forward	Curva de crescimento; Variação salinidade	TSS-3977
plasmid_pNRC200	223420	223720	forward	-	TSS-3995
plasmid_pNRC200	223437	223737	forward	Δ VNG2099 (RNase)	TSS-3996
plasmid_pNRC200	254751	255051	forward	Variação salinidade, Δ VNG2099 (RNase)	TSS-4020
plasmid_pNRC200	260320	260620	forward	Variação salinidade	TSS-4026
plasmid_pNRC200	267993	268293	forward	Variação salinidade	TSS-4036
plasmid_pNRC200	274319	274619	forward	Curva de crescimento	TSS-4042
plasmid_pNRC200	274350	274650	forward	-	TSS-4043
plasmid_pNRC200	280221	280521	forward	Curva de crescimento, Variação salinidade	TSS-4049
plasmid_pNRC200	328562	328862	forward	Curva de crescimento	TSS-4096
plasmid_pNRC200	332246	332546	forward	Curva de crescimento, Δ VNG2099 (RNase)	TSS-4100
plasmid_pNRC200	332271	332571	forward	Curva de crescimento; Δ VNG2099 (RNase)	TSS-4101
plasmid_pNRC200	332327	332627	forward	Curva de crescimento; Δ VNG2099 (RNase)	TSS-4102
plasmid_pNRC200	135222	134922	reverse	Δ VNG2099 (RNase)	TSS-4119
plasmid_pNRC200	135338	135038	reverse	-	TSS-4120
plasmid_pNRC200	141036	140736	reverse	Variação salinidade	TSS-4127
plasmid_pNRC200	143727	143427	reverse	Curva de crescimento; Variação salinidade	TSS-4132
plasmid_pNRC200	154324	154024	reverse	Variação salinidade	TSS-4149
plasmid_pNRC200	154355	154055	reverse	Variação salinidade	TSS-4150
plasmid_pNRC200	156273	155973	reverse	Δ VNG2099 (RNase); Estresse oxidativo	TSS-4155
plasmid_pNRC200	163760	163460	reverse	Curva de crescimento	TSS-4169
plasmid_pNRC200	206158	205858	reverse	Curva de crescimento; Variação salinidade	TSS-4196
plasmid_pNRC200	227499	227199	reverse	Curva de crescimento; Variação salinidade; Δ lsm	TSS-4218
plasmid_pNRC200	252806	252506	reverse	Variação salinidade	TSS-4254
plasmid_pNRC200	254460	254160	reverse	-	TSS-4256
plasmid_pNRC200	258528	258228	reverse	-	TSS-4260
plasmid_pNRC200	260423	260123	reverse	Variação salinidade, Estresse oxidativo	TSS-4262
plasmid_pNRC200	260470	260170	reverse	Variação salinidade	TSS-4263
plasmid_pNRC200	260723	260423	reverse	-	TSS-4264
plasmid_pNRC200	260998	260698	reverse	Variação salinidade	TSS-4265
plasmid_pNRC200	261020	260720	reverse	Curva de crescimento, Variação salinidade	TSS-4266
plasmid_pNRC200	272311	272011	reverse	-	TSS-4288
plasmid_pNRC200	272574	272274	reverse	-	TSS-4289
plasmid_pNRC200	272967	272667	reverse	Curva de crescimento	TSS-4290
plasmid_pNRC200	274585	274285	reverse	Variação salinidade	TSS-4292
plasmid_pNRC200	276647	276347	reverse	-	TSS-4298
plasmid_pNRC200	293012	292712	reverse	Curva de crescimento	TSS-4315
plasmid_pNRC200	293291	292991	reverse	-	TSS-4316
plasmid_pNRC200	324711	324411	reverse	-	TSS-4344
plasmid_pNRC200	332670	332370	reverse	-	TSS-4357

Anexo F - ncRNAs identificados em Koide et al 2009

Cromossomo	Início	Fim	Orientação	TSS associado	Região repetitiva
chr	21331	21631	forward	-	Não
chr	30251	30431	forward	-	Não
chr	95471	95871	forward	pTSS	Não
chr	119131	119271	forward	oTSS	Não
chr	510911	510971	forward	-	Não
chr	580231	580391	forward	oTSS	Não
chr	592071	592111	forward	aTSS	Não
chr	697771	697931	forward	-	Sim
chr	760211	760311	forward	pTSS	Não
chr	772451	772511	forward	oTSS	Não
chr	812431	812491	forward	-	Sim
chr	983851	984131	forward	oTSS	Não
chr	1104271	1104311	forward	-	Não
chr	1154131	1154311	forward	oTSS	Não
chr	1173131	1173931	forward	-	Não
chr	1356611	1356651	forward	oTSS	Não
chr	1413291	1413411	forward	pTSS	Não
chr	1413291	1413431	forward	pTSS	Não
chr	1926671	1926771	forward	pTSS	Não
chr	1937151	1937391	forward	oTSS	Não
chr	430041	430101	reverse	iTSS	Não
chr	522281	522521	reverse	oTSS	Não
chr	534441	534501	reverse	-	Não
chr	748801	748861	reverse	oTSS	Não
chr	760261	760361	reverse	-	Não
chr	1043301	1043601	reverse	oTSS	Não
chr	1107561	1107681	reverse	iTSS	Não
chr	1922181	1922241	reverse	-	Não
plasmid_pNRC100	115691	115838	forward	oTSS	Não
plasmid_pNRC100	116762	116909	forward	oTSS	Não
plasmid_pNRC100	127514	127997	forward	oTSS	Não
plasmid_pNRC100	133583	133646	forward	-	Sim
plasmid_pNRC100	133583	133667	forward	-	Sim
plasmid_pNRC100	136943	137006	forward	-	Não
plasmid_pNRC100	123786	123891	reverse	oTSS	Não
plasmid_pNRC100	146991	147075	reverse	aTSS	Não
plasmid_pNRC200	3414	3477	forward	-	Sim
plasmid_pNRC200	8874	9042	forward	-	Sim
plasmid_pNRC200	88401	88464	forward	-	Sim
plasmid_pNRC200	97662	97788	forward	-	Sim
plasmid_pNRC200	125004	125067	forward	aTSS	Não
plasmid_pNRC200	132396	132480	forward	aTSS	Não
plasmid_pNRC200	219714	219798	forward	oTSS	Não
plasmid_pNRC200	267993	268266	forward	oTSS	Não
plasmid_pNRC200	295776	295923	forward	pTSS	Não
plasmid_pNRC200	311064	311148	forward	-	Sim
plasmid_pNRC200	324252	324378	forward	-	Sim
plasmid_pNRC200	339204	339624	forward	-	Sim
plasmid_pNRC200	355017	355479	forward	-	Sim
plasmid_pNRC200	355269	355437	forward	-	Sim
plasmid_pNRC200	9136	9283	reverse	-	Sim
plasmid_pNRC200	36499	36898	reverse	-	Sim
plasmid_pNRC200	36751	36814	reverse	-	Sim
plasmid_pNRC200	71044	71170	reverse	-	Sim
plasmid_pNRC200	79717	79780	reverse	-	Sim
plasmid_pNRC200	89482	89587	reverse	-	Sim
plasmid_pNRC200	97609	97840	reverse	-	Sim
plasmid_pNRC200	101347	101515	reverse	-	Sim
plasmid_pNRC200	136984	137047	reverse	-	Não
plasmid_pNRC200	270565	270628	reverse	-	Sim
plasmid_pNRC200	326887	326971	reverse	oTSS	Não

Anexo G - *intra*RNAs diferencialmente expressos

Cromossomo	Orientação	Posição TSS	Nome	Gene	Condição
cromossomo	forward	10876	TSS-10	VNG_RS00045	GSE53544
cromossomo	forward	14922	TSS-17	VNG_RS00070	GSE53544; GSE112133
cromossomo	forward	48989	TSS-97	VNG_RS00210	Oxygen-Shift
cromossomo	forward	64061	TSS-133	VNG_RS00270	Δlsm
cromossomo	forward	90004	TSS-166	VNG_RS00420	GSE31308
cromossomo	forward	90107	TSS-167	VNG_RS00420	GSE31308
cromossomo	forward	90306	TSS-168	VNG_RS00420	GSE31308
cromossomo	forward	184746	TSS-280	VNG_RS00885	GSE112133
cromossomo	forward	311457	TSS-392	VNG_RS01570	GSE12923
cromossomo	forward	318991	TSS-398	VNG_RS01600	Δlsm
cromossomo	forward	399855	TSS-451	VNG_RS02015	GSE12923
cromossomo	forward	464901	TSS-508	VNG_RS02420	GSE112133
cromossomo	forward	542895	TSS-575	VNG_RS02820	GSE53544
cromossomo	forward	617480	TSS-646	VNG_RS03200	Δlsm
cromossomo	forward	634392	TSS-658	VNG_RS03280	GSE53544
cromossomo	forward	655842	TSS-678	VNG_RS03385	GSE112133
cromossomo	forward	717906	TSS-713	VNG_RS03660	GSE12923
cromossomo	forward	750438	TSS-749	VNG_RS03835	GSE12923
cromossomo	forward	750690	TSS-750	VNG_RS03835	GSE12923
cromossomo	forward	750791	TSS-751	VNG_RS03835	GSE12923
cromossomo	forward	774916	TSS-798	VNG_RS03945	Oxygen-Shift
cromossomo	forward	802022	TSS-827	VNG_RS04105	GSE12923
cromossomo	forward	802037	TSS-828	VNG_RS04105	GSE12923
cromossomo	forward	808070	TSS-841	VNG_RS04130	GSE53544
cromossomo	forward	812824	TSS-846	VNG_RS04150	GSE53544; GSE112133
cromossomo	forward	818018	TSS-853	VNG_RS04165	GSE53544
cromossomo	forward	915135	TSS-929	VNG_RS04750	GSE12923
cromossomo	forward	949916	TSS-954	VNG_RS04945	GSE12923
cromossomo	forward	1021757	TSS-1015	VNG_RS05350	GSE12923
cromossomo	forward	1029436	TSS-1022	VNG_RS05385	GSE12923
cromossomo	forward	1066453	TSS-1056	VNG_RS05590	GSE12923
cromossomo	forward	1085083	TSS-1072	VNG_RS05695	GSE12923
cromossomo	forward	1085114	TSS-1073	VNG_RS05695	GSE12923
cromossomo	forward	1107509	TSS-1098	VNG_RS05805	GSE53544
cromossomo	forward	1272788	TSS-1240	VNG_RS06675	GSE12923
cromossomo	forward	1272826	TSS-1241	VNG_RS06675	GSE12923; GSE53544
cromossomo	forward	1356277	TSS-1317	VNG_RS07115	Oxygen-Shift
cromossomo	forward	1358907	TSS-1324	VNG_RS07130	GSE12923
cromossomo	forward	1464013	TSS-1405	VNG_RS07680	Δlsm
cromossomo	forward	1580440	TSS-1488	VNG_RS08330	Δlsm
cromossomo	forward	1629675	TSS-1534	VNG_RS08575	Δlsm
cromossomo	forward	1691736	TSS-1589	ID=gene1770	Δlsm
cromossomo	forward	1767833	TSS-1648	VNG_RS09235	GSE112133
cromossomo	forward	1902517	TSS-1809	VNG_RS09935	Δlsm
cromossomo	reverse	44400	TSS-2058	VNG_RS00195	GSE53544
cromossomo	reverse	45172	TSS-2059	VNG_RS00195	GSE12923; GSE53544
cromossomo	reverse	45191	TSS-2060	VNG_RS00195	GSE12923; GSE53544
cromossomo	reverse	45203	TSS-2061	VNG_RS00195	GSE12923; GSE53544
cromossomo	reverse	45219	TSS-2062	VNG_RS00195	GSE12923; GSE53544
cromossomo	reverse	45264	TSS-2063	VNG_RS00195	GSE12923; GSE53544
cromossomo	reverse	54856	TSS-2087	VNG_RS00240	GSE53544
cromossomo	reverse	54974	TSS-2088	VNG_RS00240	GSE53544
cromossomo	reverse	55221	TSS-2089	VNG_RS00240	GSE53544
cromossomo	reverse	55304	TSS-2090	VNG_RS00240	GSE53544
cromossomo	reverse	55475	TSS-2091	VNG_RS00240	Δlsm
cromossomo	reverse	117381	TSS-2178	VNG_RS00585	GSE12923; GSE53544
cromossomo	reverse	125373	TSS-2186	VNG_RS00630	Δlsm
cromossomo	reverse	155014	TSS-2209	VNG_RS00740	GSE12923
cromossomo	reverse	251687	TSS-2304	VNG_RS01270	GSE112133
cromossomo	reverse	259670	TSS-2313	VNG_RS01300	GSE12923
cromossomo	reverse	292104	TSS-2333	ID=gene291	GSE112133
cromossomo	reverse	470203	TSS-2490	VNG_RS02445	GSE61975
cromossomo	reverse	624412	TSS-2597	VNG_RS03225	Oxygen-Shift
cromossomo	reverse	657231	TSS-2618	VNG_RS03395	Δlsm; GSE15788; GSE112133
cromossomo	reverse	766017	TSS-2735	VNG_RS03890	GSE12923
cromossomo	reverse	918779	TSS-2894	VNG_RS04770	Δlsm; GSE12923; GSE112133
cromossomo	reverse	936066	TSS-2904	VNG_RS04865	GSE53544
cromossomo	reverse	1006603	TSS-2965	VNG_RS05265	Δlsm
cromossomo	reverse	1016810	TSS-2975	VNG_RS05320	Δlsm
cromossomo	reverse	1086865	TSS-3029	VNG_RS05700	Δlsm

cromossomo	reverse	1126929	TSS-3065	VNG_RS05895	GSE112133
cromossomo	reverse	1183688	TSS-3095	VNG_RS06165	GSE53544; GSE112133
cromossomo	reverse	1184121	TSS-3096	VNG_RS06165	Δlsm
cromossomo	reverse	1184140	TSS-3097	VNG_RS06165	Δlsm
cromossomo	reverse	1184209	TSS-3098	VNG_RS06165	Δlsm
cromossomo	reverse	1206856	TSS-3123	VNG_RS06285	Δlsm; GSE12923
cromossomo	reverse	1252691	TSS-3165	VNG_RS06520	GSE112133
cromossomo	reverse	1420704	TSS-3297	VNG_RS07475	GSE12923
cromossomo	reverse	1443939	TSS-3318	VNG_RS07595	Δlsm
cromossomo	reverse	1457833	TSS-3333	VNG_RS07650	GSE12923
cromossomo	reverse	1534209	TSS-3409	VNG_RS08080	Δlsm
cromossomo	reverse	1625391	TSS-3512	VNG_RS08550	GSE53544
cromossomo	reverse	1627823	TSS-3513	VNG_RS08560	Δlsm
cromossomo	reverse	1825849	TSS-3681	VNG_RS09510	GSE12923
cromossomo	reverse	1840383	TSS-3698	VNG_RS09595	GSE53544
cromossomo	reverse	1863937	TSS-3718	VNG_RS09730	Δlsm; GSE61975
cromossomo	reverse	1971047	TSS-3821	VNG_RS10305	GSE53544; GSE112133
cromossomo	reverse	2006517	TSS-3869	VNG_RS10485	GSE12923
plasmideo_pNRC100	forward	113318	TSS-4361	VNG_RS12740	GSE53544
plasmideo_pNRC100	forward	113903	TSS-4362	VNG_RS12740	GSE53544
plasmideo_pNRC100	forward	114589	TSS-4366	VNG_RS12745	Δlsm; GSE112133
plasmideo_pNRC100	forward	116073	TSS-4377	VNG_RS12755	Δlsm ; GSE61975
plasmideo_pNRC100	forward	121557	TSS-4386	VNG_RS12765	Δlsm
plasmideo_pNRC100	forward	121578	TSS-4387	VNG_RS12765	Δlsm
plasmideo_pNRC100	forward	121930	TSS-4388	VNG_RS12765	GSE112133
plasmideo_pNRC100	forward	126734	TSS-4393	VNG_RS12790	Δlsm
plasmideo_pNRC100	forward	131809	TSS-4412	VNG_RS12820	GSE112133
plasmideo_pNRC100	forward	135283	TSS-4416	VNG_RS12830	Δlsm
plasmideo_pNRC100	forward	135310	TSS-4417	VNG_RS12830	Δlsm
plasmideo_pNRC100	forward	141178	TSS-4427	VNG_RS12865	GSE112133
plasmideo_pNRC100	forward	144176	TSS-4433	VNG_RS12880	GSE61975
plasmideo_pNRC100	forward	147769	TSS-4441	VNG_RS12910	GSE12923; GSE61975
plasmideo_pNRC100	forward	149378	TSS-4447	VNG_RS12920	GSE61975
plasmideo_pNRC100	forward	149401	TSS-4448	VNG_RS12920	GSE61975
plasmideo_pNRC100	reverse	131445	TSS-4499	VNG_RS12815	GSE53544
plasmideo_pNRC100	reverse	137330	TSS-4510	VNG_RS12845	GSE112133
plasmideo_pNRC100	reverse	138145	TSS-4511	VNG_RS12855	GSE61975
plasmideo_pNRC100	reverse	138253	TSS-4512	VNG_RS12855	GSE61975
plasmideo_pNRC100	reverse	138291	TSS-4513	VNG_RS12855	GSE61975
plasmideo_pNRC200	reverse	126561	TSS-4111	VNG_RS11080	GSE112133
plasmideo_pNRC200	reverse	126573	TSS-4112	VNG_RS11080	GSE112133
plasmideo_pNRC200	reverse	126689	TSS-4113	VNG_RS11080	GSE112133
plasmideo_pNRC200	reverse	259442	TSS-4261	VNG_RS11680	Δlsm
plasmideo_pNRC200	reverse	278195	TSS-4299	VNG_RS11750	GSE61975
plasmideo_pNRC200	reverse	282317	TSS-4306	VNG_RS11765	Δlsm
plasmideo_pNRC200	reverse	286332	TSS-4311	VNG_RS11785	Δlsm
plasmideo_pNRC200	reverse	299369	TSS-4324	VNG_RS11850	Δlsm

Anexo H - Domínios proteicos associados a iTSS

Cromossomo	Orientação	TSS Posição	Nome	pFAM	Função
cromossomo	forward	3768	TSS-7	PF02516	N-glycosylation of asparagines
cromossomo	forward	37498	TSS-66	PF07282	DNA-binding domain
cromossomo	forward	71399	TSS-147	PF04892	unknow function
cromossomo	forward	89891	TSS-165	PF01420	DNA binding
cromossomo	forward	90004	TSS-166	PF01420	DNA binding
cromossomo	forward	91380	TSS-173	PF04851	NTPase domain superfamily are characterized by a conserved nucleotide phosphate-binding motif
cromossomo	forward	138170	TSS-222	PF13193	AMP-binding enzyme C-terminal domain
cromossomo	forward	201879	TSS-297	PF10120	thiamine-phosphate synthase
cromossomo	forward	228512	TSS-322	PF00633	DNA-binding motif
cromossomo	forward	497286	TSS-537	PF00571	bind ligands with an adenosyl group
cromossomo	forward	548604	TSS-581	PF01554	multidrug and toxic compound extrusion
cromossomo	forward	556156	TSS-591	PF00989	bind ligand
cromossomo	forward	556586	TSS-592	PF00512	Histidine Kinase A domain
cromossomo	forward	672581	TSS-691	PF00204	Transducer domain
cromossomo	forward	717906	TSS-713	PF13549	ATP-grasp domain
cromossomo	forward	734074	TSS-731	PF13646	intracellular transport
cromossomo	forward	775560	TSS-800	PF13417	Glutathione S-transferase
cromossomo	forward	801364	TSS-824	PF14667	integral membrane region
cromossomo	forward	808321	TSS-842	PF05050	methyltransferase
cromossomo	forward	812824	TSS-846	PF05050	methyltransferase
cromossomo	forward	848489	TSS-875	PF16897	nucleotide phosphate-binding motif
cromossomo	forward	848886	TSS-876	PF02824	suggests a ligand (most likely nucleotide)-binding
cromossomo	forward	894561	TSS-913	PF07731	Copper binding
cromossomo	forward	999039	TSS-1003	PF13193	AMP-binding enzyme C-terminal domain
cromossomo	forward	1096604	TSS-1087	PF07705	Cell adhesion related domain
cromossomo	forward	1209940	TSS-1173	PF03949	NAD binding domain
cromossomo	forward	1218026	TSS-1179	PF00127	Copper binding
cromossomo	forward	1239722	TSS-1208	PF00571	bind ligands with an adenosyl group
cromossomo	forward	1473489	TSS-1410	PF02517	protease
cromossomo	forward	1518374	TSS-1440	PF00535	Glycosyltransferases
cromossomo	forward	1567352	TSS-1483	PF03464	hydrolytic activity at the peptidyl transferase centre – domain 2
cromossomo	forward	1580440	TSS-1488	PF08241	Methyltransferase domain
cromossomo	forward	1670608	TSS-1565	PF00352	TATA box binding protein
cromossomo	forward	1684546	TSS-1582	PF01938	Predicted RNA-binding protein
cromossomo	forward	1729542	TSS-1623	PF13187	bind to iron-sulfur clusters
cromossomo	forward	1902517	TSS-1809	PF06146	unknow function
cromossomo	reverse	181802	TSS-2244	PF13555	ATPase
cromossomo	reverse	181934	TSS-2245	PF13555	ATPase
cromossomo	reverse	182466	TSS-2246	PF08706	unknow function
cromossomo	reverse	208544	TSS-2272	PF00382	Protein binding domain
cromossomo	reverse	209027	TSS-2273	PF00382	Protein binding domain
cromossomo	reverse	251687	TSS-2304	PF13412	DNA binding domains
cromossomo	reverse	256110	TSS-2310	PF01588	tRNA-binding-domain
cromossomo	reverse	259670	TSS-2313	PF01435	metallopeptidases
cromossomo	reverse	271013	TSS-2319	PF06470	Responsible for formation of the DNA interacting dimer
cromossomo	reverse	470203	TSS-2490	PF02080	bind an unidentified ligand
cromossomo	reverse	566851	TSS-2558	PF06798	unknow function
cromossomo	reverse	582732	TSS-2576	PF01163	Protein Kinases catalytic domain
cromossomo	reverse	657231	TSS-2618	PF00382	Protein binding domain
cromossomo	reverse	727720	TSS-2670	PF04659	C-terminus of flaD
cromossomo	reverse	755165	TSS-2715	PF00589	C-terminal catalytic domain
cromossomo	reverse	755196	TSS-2716	PF00589	C-terminal catalytic domain
cromossomo	reverse	838185	TSS-2816	PF01903	Sirohydrochlorin cobalt chelatase (CbiX) and sirohydrochlorin iron chelatase (SirB) N-terminal domain
cromossomo	reverse	918779	TSS-2894	PF00571	bind ligands with an adenosyl group
cromossomo	reverse	961801	TSS-2932	PF02080	bind an unidentified ligand
cromossomo	reverse	1183688	TSS-3095	PF07510	endonucleases
cromossomo	reverse	1320759	TSS-3220	PF13662	nucleotidyl transferase/hydrolase domain
cromossomo	reverse	1510073	TSS-3387	PF00575	RNA binding domain
cromossomo	reverse	1519851	TSS-3393	PF02080	bind an unidentified ligand
cromossomo	reverse	1520160	TSS-3395	PF02254	NAD binding
cromossomo	reverse	1573533	TSS-3459	PF00006	Nucleotide-binding
cromossomo	reverse	1574135	TSS-3460	PF00306	C-terminal Domain
cromossomo	reverse	1574159	TSS-3461	PF00306	C-terminal Domain
cromossomo	reverse	1574312	TSS-3463	PF00306	C-terminal Domain
cromossomo	reverse	1627823	TSS-3513	PF00127	Copper binding

cromossomo	reverse	1642117	TSS-3525	PF02889	required for assembly of functional endoplasmic reticulum translocons.
cromossomo	reverse	1661756	TSS-3544	PF00579	dinucleotide-binding domain
cromossomo	reverse	1678795	TSS-3559	PF02747	dna interaction
cromossomo	reverse	1686546	TSS-3567	PF00902	Sec-independent protein translocase protein
cromossomo	reverse	1702346	TSS-3580	PF01502	Phosphoribosyl-AMP cyclohydrolase
cromossomo	reverse	1721371	TSS-3596	PF00571	bind ligands with an adenosyl group
cromossomo	reverse	1758630	TSS-3621	PF00496	substrate-binding domain
cromossomo	reverse	1854354	TSS-3712	PF13307	nucleotide phosphate-binding motif
cromossomo	reverse	1990209	TSS-3843	PF00679	Elongation factor G C-terminus
cromossomo	reverse	1991171	TSS-3844	PF03144	domain II of elongation factor 2
plasmídeo_pNRC100	forward	121298	TSS-4385	PF00589	C-terminal catalytic domain
plasmídeo_pNRC200	forward	251492	TSS-4012	PF13166	atpase
plasmídeo_pNRC200	forward	287904	TSS-4058	PF01385	transposase domain
plasmídeo_pNRC200	forward	287990	TSS-4059	PF01385	transposase domain
plasmídeo_pNRC200	forward	290031	TSS-4061	PF00136	DNA polymerase type-B B2 subfamily catalytic domain
plasmídeo_pNRC200	reverse	146770	TSS-4136	PF02080	bind an unidentified ligand
plasmídeo_pNRC200	reverse	154920	TSS-4151	PF07282	DNA-binding domain
plasmídeo_pNRC200	reverse	245273	TSS-4242	PF00185	Aspartate/ornithine binding domain
plasmídeo_pNRC200	reverse	253767	TSS-4255	PF00589	C-terminal catalytic domain
plasmídeo_pNRC200	reverse	275237	TSS-4293	PF02150	RNA polymerases M/15 Kd subunit
plasmídeo_pNRC200	reverse	286332	TSS-4311	PF02082	DNA Binding

Anexo I - Domínios proteicos associados a iTSS em outras espécies

pFAM	Função	Organismo	Gene	<i>H. salinarum</i> gene
PF00006	ATP synthase alpha-beta	<i>C. jejuni</i>	CJE_RS00535	VNG_RS08290
PF00006	ATP synthase alpha-beta	<i>N. lactamica</i>	NLA_RS09725	VNG_RS08290
PF00006	ATP synthase alpha-beta	<i>C. jejuni</i>	CJE_RS00525	VNG_RS08290
PF00185	OTCace	<i>C. jejuni</i>	CJE_RS06185	VNG_RS11625
PF00185	OTCace	<i>C. jejuni</i>	CJE_RS05430	VNG_RS11625
PF00204	DNA_gyrase	<i>N. lactamica</i>	NLA_1740	ID=gene692
PF00204	DNA_gyrase	<i>N. lactamica</i>	NLA_RS03165	ID=gene692
PF00571	CBS domain	<i>S. oneidensis</i>	SO_2815	VNG_RS02575; VNG_RS04770; VNG_RS09025
PF00571	CBS domain	<i>T. kodakarensis</i>	TK1626	VNG_RS02575; VNG_RS04770; VNG_RS09025
PF00575	S1 RNA Binding Domain	<i>N. lactamica</i>	NLA_11360	VNG_RS07945
PF00575	S1 RNA Binding Domain	<i>N. lactamica</i>	NLA_0360	VNG_RS07945
PF00679	Elongation factor G C-terminus	<i>C. jejuni</i>	CJE0542	ID=gene2082
PF00679	Elongation factor G C-terminus	<i>S. oneidensis</i>	SO_0228	ID=gene2082
PF00679	Elongation factor G C-terminus	<i>S. oneidensis</i>	SO_0842	ID=gene2082
PF01385	OrfB_IS605	<i>H. pylori</i>	HP0438	VNG_RS11800; VNG_RS11205
PF01385	OrfB_IS605	<i>H. pylori</i>	HP0989	VNG_RS11800; VNG_RS11205
PF01385	OrfB_IS605	<i>H. pylori</i>	HP0997	VNG_RS11800; VNG_RS11205
PF01385	OrfB_IS605	<i>H. pylori</i>	HP1095	VNG_RS11800; VNG_RS11205
PF01385	OrfB_IS605	<i>H. pylori</i>	HP1534	VNG_RS11800; VNG_RS11205
PF01385	OrfB_IS605	<i>T. kodakarensis</i>	TK1842	VNG_RS11800; VNG_RS11205
PF07282	OrfB_Zn_ribbon	<i>H. pylori</i>	HP0438	VNG_RS11800; VNG_RS11205
PF07282	OrfB_Zn_ribbon	<i>H. pylori</i>	HP0989	VNG_RS11800; VNG_RS11205
PF07282	OrfB_Zn_ribbon	<i>H. pylori</i>	HP0997	VNG_RS11800; VNG_RS11205
PF07282	OrfB_Zn_ribbon	<i>H. pylori</i>	HP1095	VNG_RS11800; VNG_RS11205
PF07282	OrfB_Zn_ribbon	<i>H. pylori</i>	HP1534	VNG_RS11800; VNG_RS11205
PF07282	OrfB_Zn_ribbon	<i>T. kodakarensis</i>	TK1842	VNG_RS11800; VNG_RS11205
PF07282	OrfB_Zn_ribbon	<i>T. kodakarensis</i>	TK0298	VNG_RS11800
PF07282	OrfB_Zn_ribbon	<i>T. kodakarensis</i>	TK0931	VNG_RS11800
PF02080	TrkA_C	<i>T. kodakarensis</i>	TK1575	VNG_RS02445; VNG_RS05015; VNG_RS07995; VNG_RS11170
PF02080	TrkA_C	<i>T. kodakarensis</i>	TK1595	VNG_RS02445; VNG_RS05015; VNG_RS07995; VNG_RS11170
PF03144	GTP_EFTU_D2	<i>N. lactamica</i>	NLA_14640	ID=gene2082
PF03144	GTP_EFTU_D2	<i>T. kodakarensis</i>	TK0308	ID=gene2082
PF03144	GTP_EFTU_D2	<i>H. volcanii</i>	HVO_0356	ID=gene2082
PF00589	Phage_integrase	<i>S. oneidensis</i>	SO_A0045	VNG_RS03850; VNG_RS12765; VNG_RS11655
PF00382	TFIIB	<i>T. kodakarensis</i>	TK1280	VNG_RS01035
PF02824	TGS	<i>T. kodakarensis</i>	TK1286	VNG_RS04330
PF03949	Malic_M	<i>T. kodakarensis</i>	TK1963	VNG_RS06300
PF00136	DNA_polymerase_activity	<i>H. volcanii</i>	HVO_0858	VNG_RS11805
PF00579	tryptophan-tRNA_ligase_activity	<i>H. volcanii</i>	HVO_RS05375	VNG_RS08695
PF00902	TatC	<i>H. volcanii</i>	HVO_RS05600	VNG_RS08845
PF01435	metalloendopeptidase_activity	<i>N. lactamic</i>	NLA_14150	VNG_RS01300
PF01554	drug_transmembrane_transporter	<i>N. lactamica</i>	NLA_14180	VNG_RS02850

Anexo J - Quebras do sinal de expressão gênica obtidos através de RNA-seq e associados com terminadores intrínsecos da transcrição do tipo polyT

Cromossomo	Posição	Orientação	Gene	Distância stop codon	PolyT 5/4	PolyT 10/6
chr	65573	forward	VNG_RS00280	102	-	-
chr	66554	forward	VNG_RS00290	21	-	-
chr	95873	forward	VNG_RS00440	21	-27	-
chr	119150	forward	VNG_RS00595	99	-13	-16
chr	158454	forward	VNG_RS00755	147	-15	-19
chr	214469	forward	VNG_RS01065	45	-	-
chr	442491	forward	VNG_RS02260	25	-25/-13	-
chr	579775	forward	VNG_RS03020	34	-26	-27
chr	598835	forward	VNG_RS03115	92	-44/5	6
chr	676623	forward	VNG_RS03475	137	-	-
chr	752951	forward	VNG_RS03845	149	-43/-12	-
chr	760557	forward	VNG_RS03870	5	-	-
chr	774683	forward	VNG_RS03940	138	-	-
chr	775844	forward	VNG_RS03950	40	-	-
chr	835810	forward	VNG_RS04245	8	5	-
chr	852941	forward	VNG_RS04355	63	-	-
chr	853118	forward	VNG_RS04360	91	-	-
chr	893888	forward	VNG_RS04635	98	-	-
chr	950921	forward	VNG_RS04955	62	-	-
chr	953329	forward	VNG_RS04970	63	-	-
chr	997698	forward	VNG_RS05215	33	0	-
chr	1030299	forward	VNG_RS05385	43	-	-
chr	1055333	forward	VNG_RS05520	110	-29/5	-
chr	1060133	forward	VNG_RS05545	148	-	-
chr	1068439	forward	VNG_RS05600	65	-	-
chr	1068490	forward	VNG_RS05600	116	-	-
chr	1101364	forward	VNG_RS05765	14	-	-
chr	1107645	forward	VNG_RS05805	87	-	-
chr	1113022	forward	VNG_RS05840	5	-	-
chr	1219926	forward	VNG_RS06355	38	-	-
chr	1229092	forward	VNG_RS06400	150	-3/5	-3
chr	1245001	forward	VNG_RS06465	105	-	-
chr	1296929	forward	VNG_RS06805	141	-16	-
chr	1317976	forward	VNG_RS06900	17	-	-
chr	1327451	forward	VNG_RS06955	6	-	-
chr	1338177	forward	VNG_RS07020	12	-	-
chr	1356578	forward	VNG_RS07115	127	-5	-8
chr	1405160	forward	VNG_RS07360	81	-	-
chr	1417395	forward	VNG_RS07440	4	-8	-
chr	1458673	forward	VNG_RS07655	114	-	-17
chr	1502243	forward	VNG_RS07885	145	12	-
chr	1505379	forward	VNG_RS07900	47	-	-
chr	1558286	forward	VNG_RS08215	6	-4	-
chr	1671378	forward	VNG_RS08750	37	-14	-
chr	1676522	forward	VNG_RS08780	50	-24	-
chr	1678574	forward	VNG_RS08795	80	-	-
chr	1697454	forward	ID=gene1777	104	-	-
chr	1864993	forward	VNG_RS09735	106	-	-
chr	1866438	forward	VNG_RS09750	56	-	-
chr	1874666	forward	VNG_RS09785	120	-	-
chr	1880708	forward	VNG_RS09805	65	-25/-16/6	-24/2
chr	1880896	forward	VNG_RS09810	67	-	-
chr	1905085	forward	VNG_RS09945	58	-40/8	-
chr	1930329	forward	VNG_RS10060	63	-15	-19
chr	1932508	forward	VNG_RS10065	85	-	-
chr	1935660	forward	VNG_RS10100	38	-	-
chr	1955262	forward	VNG_RS10230	36	-	-
chr	1959098	forward	VNG_RS10240	42	-18	-22
chr	1963705	forward	VNG_RS10255	137	23	-
chr	1980170	forward	VNG_RS10340	25	-	-
chr	1982000	forward	VNG_RS10355	72	-	-
chr	1984195	forward	VNG_RS10370	75	-	-30
chr	20506	reverse	VNG_RS00105	28	-	-
chr	34567	reverse	VNG_RS00160	85	-14	-19/8
chr	154580	reverse	VNG_RS00740	115	-	-
chr	165121	reverse	VNG_RS00790	32	-16	-19
chr	255714	reverse	VNG_RS01290	29	21	17

chr	280465	reverse	VNG_RS01395	23	-	-
chr	283758	reverse	VNG_RS01410	7	-	-
chr	304225	reverse	VNG_RS01540	44	-14	-16
chr	359349	reverse	VNG_RS01815	40	-16	-
chr	394202	reverse	VNG_RS01995	29	-15	-
chr	430979	reverse	VNG_RS02195	56	-28	-30
chr	445211	reverse	VNG_RS02280	53	-	-
chr	445871	reverse	VNG_RS02290	27	-16	-20
chr	462165	reverse	VNG_RS02405	29	-35	-
chr	531912	reverse	VNG_RS02765	127	-	-
chr	551719	reverse	VNG_RS02870	5	-	-
chr	580421	reverse	VNG_RS03030	31	-15	-18
chr	620344	reverse	VNG_RS03210	120	-	-
chr	663624	reverse	VNG_RS03435	25	-21	-
chr	751778	reverse	VNG_RS03840	126	-24	-24
chr	756599	reverse	VNG_RS03855	77	-32	-39
chr	811078	reverse	VNG_RS04145	75	-	-
chr	847285	reverse	VNG_RS04320	20	-13	-
chr	855604	reverse	VNG_RS04385	116	-17	-20
chr	882130	reverse	VNG_RS04560	15	-	-11
chr	895680	reverse	VNG_RS04650	28	-	-
chr	896214	reverse	VNG_RS04655	42	-15	-
chr	916305	reverse	VNG_RS04760	89	-1	-
chr	920273	reverse	VNG_RS04785	4	-	-
chr	950070	reverse	VNG_RS04950	9	-3	-
chr	992740	reverse	VNG_RS05190	12	-	-
chr	1052119	reverse	VNG_RS05515	30	-20	-
chr	1103699	reverse	VNG_RS05785	43	-16	-19
chr	1131361	reverse	VNG_RS05925	84	10	10
chr	1244980	reverse	VNG_RS06470	2	-	-
chr	1281627	reverse	VNG_RS06725	117	-	-
chr	1298839	reverse	VNG_RS06820	95	-	-
chr	1304090	reverse	VNG_RS06830	79	-	-
chr	1378241	reverse	VNG_RS07235	8	-	-
chr	1421180	reverse	VNG_RS07485	32	-	-
chr	1444117	reverse	VNG_RS07600	10	25	-
chr	1473890	reverse	VNG_RS07730	54	-	-
chr	1495325	reverse	VNG_RS07835	44	-	-
chr	1499148	reverse	VNG_RS07860	57	-	-
chr	1538783	reverse	VNG_RS08110	58	-14	-
chr	1549573	reverse	VNG_RS08170	148	-	-
chr	1566023	reverse	VNG_RS08255	79	-	-
chr	1570271	reverse	VNG_RS08280	8	-	-
chr	1619956	reverse	VNG_RS08525	20	-15	-
chr	1655995	reverse	VNG_RS08675	67	-16	-19
chr	1657891	reverse	VNG_RS08680	39	-15	-18
chr	1660776	reverse	VNG_RS08695	43	-	-
chr	1702402	reverse	VNG_RS08920	43	-	-
chr	1783299	reverse	VNG_RS09320	41	-	-
chr	1820901	reverse	VNG_RS09490	55	-	-
chr	1826362	reverse	VNG_RS09515	39	-6	-9
chr	1845550	reverse	VNG_RS09630	70	-15	-16
chr	1881970	reverse	VNG_RS09825	94	-	-
chr	1882274	reverse	VNG_RS09830	105	-	-
chr	1917933	reverse	VNG_RS10010	116	-	-
chr	1977701	reverse	VNG_RS10330	45	-13	-
chr	1978785	reverse	VNG_RS10335	40	-	-
chr	1984163	reverse	VNG_RS10375	79	-	-
chr	1993992	reverse	VNG_RS10430	148	-	-
plasmid_pNRC100	115070	reverse	VNG_RS12750	42	11	7
plasmid_pNRC100	131000	reverse	VNG_RS12815	123	-1	-
plasmid_pNRC200	220513	forward	VNG_RS11535	3	-	-
plasmid_pNRC200	272013	forward	VNG_RS11730	28	-	-
plasmid_pNRC200	136682	reverse	VNG_RS11110	80	-	-
plasmid_pNRC200	217524	reverse	VNG_RS11520	149	-	-
plasmid_pNRC200	227397	reverse	VNG_RS11560	120	11	6
plasmid_pNRC200	279023	reverse	VNG_RS11755	137	-	-
plasmid_pNRC200	319814	reverse	VNG_RS11970	52	-	-

Anexo K - Transcription start site associated RNAs (TSSaRNAs) are ubiquitous in all domains of life

Artigo TSSaRNAs

Transcription Start Site Associated RNAs (TSSaRNAs) Are Ubiquitous in All Domains of Life

Livia S. Zaramela¹*, Ricardo Z. N. Vêncio²*, Felipe ten-Caten¹, Nitin S. Baliga³, Tie Koide¹*

1 Department Biochemistry and Immunology, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil, **2** Department of Computing and Mathematics, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil, **3** Institute for Systems Biology, Seattle, Washington, United States of America

Abstract

A plethora of non-coding RNAs has been discovered using high-resolution transcriptomics tools, indicating that transcriptional and post-transcriptional regulation is much more complex than previously appreciated. Small RNAs associated with transcription start sites of annotated coding regions (TSSaRNAs) are pervasive in both eukaryotes and bacteria. Here, we provide evidence for existence of TSSaRNAs in several archaeal transcriptomes including: *Halobacterium salinarum*, *Pyrococcus furiosus*, *Methanococcus maripaludis*, and *Sulfolobus solfataricus*. We validated TSSaRNAs from the model archaeon *Halobacterium salinarum* NRC-1 by deep sequencing two independent small-RNA enriched (RNA-seq) and a primary-transcript enriched (dRNA-seq) strand-specific libraries. We identified 652 transcripts, of which 179 were shown to be primary transcripts (~7% of the annotated genome). Distinct growth-associated expression patterns between TSSaRNAs and their cognate genes were observed, indicating a possible role in environmental responses that may result from RNA polymerase with varying pausing rhythms. This work shows that TSSaRNAs are ubiquitous across all domains of life.

Citation: Zaramela LS, Vêncio RZN, ten-Caten F, Baliga NS, Koide T (2014) Transcription Start Site Associated RNAs (TSSaRNAs) Are Ubiquitous in All Domains of Life. PLoS ONE 9(9): e107680. doi:10.1371/journal.pone.0107680

Editor: Szabolcs Semsey, Niels Bohr Institute, Denmark

Received: March 9, 2014; **Accepted:** August 18, 2014; **Published:** September 19, 2014

Copyright: © 2014 Zaramela et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction.

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 and 11/14455-4 to FC]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: tkoide@fmrp.usp.br

These authors contributed equally to this work.

Introduction

Molecular mechanisms that are conserved throughout evolution, or arise independently to perform similar tasks are of major interest to biology [1]. Evolutionary conservation and convergence are strong indicators of important biological functions. Understanding commonalities and differences across organisms from all three domains of life have therefore served as powerful means to discover and characterize important molecular mechanisms.

The roles of non-coding RNA (ncRNA) molecules have proven to be especially elusive. Only recently, high-throughput technologies have revealed that ncRNAs have important functions across diverse biological systems and processes [2,3]. Among the newly discovered ncRNAs is an intriguing class of transcription start site associated RNAs (TSSaRNAs) that have thus far been observed in eukaryotes and bacteria [4–7].

Based on their location, TSSaRNAs have been speculated to play a role in transcription initiation [5,6,8]; and based on their tissue-specific regulation they have also been putatively implicated in epigenetic regulation [5,9]. TSSaRNAs have also been reported in bacteria where it is suggested that they could be part of a regulatory mechanism that prevents transcription initiation until a functional RNA polymerase complex has assembled [4]. In both eukaryotes and bacteria, the production of these transcripts seems

to be associated with stalled RNA polymerase [4–6]. The RNA polymerase pausing model is the most accepted TSSaRNA biogenesis hypothesis and its functional implications is still under investigation [10,11].

Regardless, TSSaRNA ubiquity across eukaryotes and bacteria suggests that TSSaRNAs are ancient and must have been present in LUCA. Discovery of TSSaRNAs in archaea would lend credibility to this hypothesis and provide clues into why they are evolutionarily conserved across all organisms.

Results and Discussion

Discovery of TSSaRNAs in the third domain of life

In the present work, we investigated whether TSSaRNAs do indeed exist in archaea and, thus, ubiquitous across all three domains of life. By mining publicly available data, we gathered evidence for TSSaRNAs in 10 archaeal transcriptomes (*H. salinarum*, *M. maripaludis*, *S. solfataricus*, *P. furiosus*, *N. equitans*, *M. kandleri*, *H. volcanii*, *M. psychrophilus*, *M. maei* and *P. abyssi* [12–21], see supplemental material), including compendia of gene expression profiles over growth curves for 4 organisms: *H. salinarum* [14], *M. maripaludis* [12], *S. solfataricus* [12,13] and *P. furiosus* [12] (**Figure 1**). We mined publicly available gene expression datasets from GEO [22] (<http://www.>

Figure 1. Data-mining on archaeal gene expression public datasets. Each panel shows an example of TSSaRNA presence in different archaeal transcriptomes. Yellow arrow represents the cognate gene, blue dashed lines represent TSSaRNA estimated regions and horizontal axis represent organism's genome coordinates. Heatmaps represent gene expression profiles over growth curves and are color-coded according to $\log_{10}$ expression ratios between each time point relative to reference condition. Light blue horizontal bars represent tiling array probe intensities for reference conditions for *H. salinarum*, *P. furiosus* and *M. maripaludis*. Dark blue points (*S. solfataricus* only) represent RNA-seq reads coverage data. Figure S1 show examples for additional archaea. doi:10.1371/journal.pone.0107680.g001

ncbi.nlm.nih.gov/geo/), SRA [23] (<http://www.ncbi.nlm.nih.gov/sra>) and UCSC Archaeal Genome Browser [24] (<http://archaea.ucsc.edu/>). Datasets not available in public databases were obtained directly from publications.

Expression of a putative TSSaRNA, measured either by hybridization intensities or by read coverage, had a distinct signature characterized by a sharp rise in signal that plateaus over a small distance and then decays precipitously. This signature was conserved across most transcriptomes that were analyzed, and across all sequencing (Illumina, SOLiD and Roche 454) and microarray (NimbleGen and Agilent) platforms, and all library construction protocols (strand-specific and non-strand specific) [12–21] (Figure S1). Aiming TSSaRNAs discovery in all archaeal organisms, all datasets were manually inspected.

TSSaRNAs in *H. salinarum* NRC-1

The consistency of TSSaRNAs discovery across all platforms and organisms justified further experimentation for independent validation. *H. salinarum* is a model organism for halophilic archaea and has been extensively studied in the last decade. It became a prime model to study aspects of gene expression regulation, especially due to the establishment of predictive quantitative models with high accuracy [25].

In order to precisely map TSSaRNAs in *H. salinarum* NRC-1, we performed a strand-specific RNA-seq experiment using non-fragmented small RNAs in the 20–230 bp range. Two biological replicates were extracted from cultures under standard growth conditions [26]. From these replicates, 3.4 million reads were aligned to *H. salinarum* NRC-1 genome.

The reads from TSSaRNAs create a surplus in coverage values when taken together with reads from the cognate gene (Figure 2, Figure S2). A given genomic location can have two sets of aligned reads starting exactly there: (i) reads from transcripts greater than 151 nt but truncated at any length, up to the maximum sequencing length limit (151 bp) and (ii) identical full-length reads from transcripts smaller than 151 nt. Although both sets map to the same initial position, the former show repeatedly the same start and end genomic coordinates. We used relative enrichment of the aligned start position as a feature to automatically detect TSSaRNAs (peaks in “start counts” profiles in Figure 2). Using this approach, we discovered 652 TSSaRNAs that were evenly distributed on both strands, and associated with 25% of all annotated protein coding genes.

To distinguish between processed and primary transcripts, we performed a dRNA-seq experiment [27]. Since primary transcripts have their 5' ends intact, a TEX (Terminator 5'-Phosphate-Dependent Exonuclease) enzyme treatment would enrich a sample for them. Comparing sequenced reads from treated (TEX+) and control (TEX-) libraries it is possible to identify primary TSSaRNAs. Using this approach, we refined our observations and defined 179 primary TSSaRNAs that were evenly distributed on both strands, and associated with 7% of all annotated protein coding genes (Table S1). It is important to note that dRNA-seq experiments are prone to false negatives [28–29], thus, it is possible that more than 179 TSSaRNAs do exist. To turn the association of TSSaRNAs to transcription start sites (TSS)

robust, we choose to further investigate only those small RNAs strictly correlated to primary TSS positions validated by dRNA-seq data.

The TSSaRNA sizes in *H. salinarum* ranged from 16 nt to 146 nt with a median size of 27 nt (Figure S3A). The distribution and median size of TSSaRNAs was consistent across many organisms: murine (range: 20 nt to 90 nt, median 20 nt) [6]; human, chicken and fruit fly (range in all three: 13 nt to 28 nt, median 18 nt) [5]. By contrast, the distribution of TSSaRNA sizes in some bacterial organisms was much narrower, e.g., *E. coli* (range: 33 nt to 40 nt) and *M. pneumoniae* (range: 35 to 55 nt, few TSSaRNAs up to ~100 nt) [4]. The proximal locations of TSSaRNAs to translation initiation sites of cognate genes (Figure S3B) are consistent with previous observations that most transcripts in *H. salinarum* are leaderless [14]. As for bacteria and eukaryote, the distribution of TSSaRNAs location shows that there are some TSS internal to annotated CDSs, which may point to structural annotation imprecision or alternative transcripts.

Transcriptome data indicates multiple and time-varying RNA polymerase pausing sites

The current understanding is that the production of TSSaRNA transcripts is associated with stalled RNA polymerase during cognate gene transcription in eukaryote and bacteria [4–6]. This polymerase pausing hypothesis is becoming the prime biogenesis model for TSSaRNA and is bringing key insights into gene expression regulation [10,11], eclipsing alternative hypothesis such as degradative 3' end processing or non-degradative (cleavage) gene processing.

In archaea, the absence of a set of RNA-seq reads starting just before TSSaRNA reads' ends (Figure S4) argue against the cleavage biogenesis hypothesis. Moreover, the observation that TSSaRNA compositional/thermodynamical properties are no different from similar regions in non-cognate gene sequences (Figure S5) argue against the degradative biogenesis hypothesis, following the same rationale put forward by Yus *et al.* [4]. Unsurprisingly, given that the molecular mechanisms involved in RNA polymerase pausing are complex [27] and often involve gene specific structures [28], there were no clear pausing site signatures in the vicinity of all 179 primary TSSaRNA 3' ends, or even considering all 652 putative TSSaRNAs. Altogether, we have no evidence to believe that only archaea would present a different biogenesis process other than RNA polymerase pausing. To explore this hypothesis properties, we created a simple computational model for RNA polymerase pausing biogenesis scenario (File S1, File S2). This model explores only two parameters for RNA polymerase: elapsed time paused at any given genomic location and time between successive transcription initiation events (Figure S6).

Using multiple pausing sites along a gene with different retention times, the model explains a recurrent RNA-seq experimental observation in our datasets: an ensemble of full-length reads aligned at the same starting position, but with different sizes. We validated this model's implication by performing classical northern-blot experiments for two highly expressed genes: one showing signs of multiple pausing sites (VNG0101G)

Figure 2. Transcriptome mapping using small RNA sequencing. Strand-specific RNA-seq experiment was performed using non-fragmented small RNAs (20–230 pb). Panels A–D show examples of TSSaRNAs and their cognate genes. (A) VNG1213C, forward strand (B) VNG0101G, reverse strand (C) VNG0635G, forward strand. (D) VNG2293G, reverse strand. For each panel, the genome position and CDSs location and orientation (grey arrows) are indicated at the horizontal axis. The uppermost graphic show $\log_2$ transformed amount of reads covering a given genomic coordinate. Vertical dashed lines represents the TSSaRNA region. The intermediary graphic represents the amount of aligned reads whose start position maps to a given genomic coordinate, the “start profile”.

doi:10.1371/journal.pone.0107680.g002

and one derived from a single pausing site (VNG1213G). VNG0101G encodes a conserved cold shock protein and was selected for further validation since the signal associated with its TSSaRNA was top ranked in tiling array experiments [14]. Notwithstanding the low sensitivity of detecting low abundance RNAs with northern blot [5], the 26 nt TSSaRNA was observed as a distinct band along with its cognate gene transcript (Figure 3B). Along with the northern-blot band directly corresponding to the most frequent reads aligned at VNG0101G’s TSS position (Figure 3A), it is possible to see other less stronger bands, which sizes also correspond to less abundant RNA-seq reads. The computational model can easily recapitulate these observations by using multiple retention positions and times (File S1). If, on the other hand, only one genomic position stalls a RNA polymerase,

then only one type of small molecule associated with the TSS would be created. This case is also observed experimentally for VNG1213C gene, a probable exonuclease: RNA-seq data shows a population of reads concentrated around 72 nt, which maps directly with the single band found in the northern-blot experiment (Figure 3CD). Therefore, our transcriptome data indicates that it is possible to find multiple RNA polymerase pausing sites along a gene sequence.

Remarkably, it was clear from gene expression profiles that dynamical behavior of a TSSaRNA may be distinct from that of its cognate gene. In some cases, the cognate gene level does not change, but expression of the TSSaRNA has distinct dynamics, with up to 16 fold up-regulation or down-regulation to different degrees (Figure 4AC). We also observed instances when both

Figure 3. TSSaRNAs reads size distribution and northern blot validation. Panel A shows the size histogram of all reads that aligned their 5' ends at VNG0101G's TSS. It is possible to verify an enrichment of sizes of 26 nt, 62 nt, 90 nt and 106 nt, which corresponds to the bands observed in the northern blot experiment. Panel B shows northern blot analysis for TSSaRNA associated with VNG0101G. Panel C shows the size histogram of all reads that aligned their 5' ends at VNG1213C's TSS. Panel D shows northern blot analysis for TSSaRNA associated with VNG1213C.
doi:10.1371/journal.pone.0107680.g003

TSSaRNA and cognate gene were differentially regulated, albeit with different patterns (**Figure 4BD**). Imposing stringent criteria, we identified at least 10 TSSaRNA differentially expressed relative to their cognate genes (**Table S2, Figure S7**). Such differential expression patterns would not be expected if transcription of a TSSaRNA and the full-length transcript of its cognate gene were not regulated by environmental signals, nor could it arise as an experimental artifact of tiling array hybridization and processing. Using pausing sites that can vary their retention time along the growth curve, the RNA polymerase pausing model explains our experimental observation that TSSaRNA can have distinct dynamical behavior relative to their cognate gene. Although counterintuitive, it is possible to generate dynamical profiles such

as the ones where TSSaRNA levels remains constant and its cognate gene varies and *vice versa*, only exploring the two parameters of the model: elapsed time spent paused and time between successive transcripts initiation events (**Figure S8, File S2**).

Therefore, our transcriptome analysis indicates that there is probably RNA polymerase pausing rhythm regulation in response to environmental perturbations. Future experimental work would reveal how this rhythm may be tuned and what are the implication of this regulation.

Figure 4. TSSaRNAs differentially expressed. Panels A, B, C and D are expression profiles of TSSaRNAs differentially expressed over a typical growth-curve relative to the control growth condition. Horizontal axis represent 13 growth curve points from different phases in a standard laboratory batch culture. Solid lines are TSSaRNA expression profiles and dashed lines their cognate gene expression profile.
doi:10.1371/journal.pone.0107680.g004

Conclusions

In this study we demonstrated that TSSaRNAs are also present in archaea. Our findings complement previous discoveries of these ncRNAs in eukaryotes and bacteria, to show that TSSaRNAs are ubiquitous in all domains of life. Furthermore, the northern-blot banding patterns in our experiment were consistent with previous observations in eukaryotes and bacteria [6,30], suggesting that TSSaRNAs may be accompanied by a population of transcripts. The prevalent TSSaRNA biogenesis hypotheses, the RNA polymerase pausing, would easily explain these patterns as well as our observation of TSSaRNA/cognate gene differential

expression. Comparative transcriptome analysis among all domains of life will be critical for elucidating the precise roles played by TSSaRNAs, in order to explain why they are evolutionarily conserved.

Materials and Methods

Data-mining on archaeal gene expression datasets

To verify the presence of TSSaRNAs in archaea, we mined archaeal publicly available gene expression datasets from GEO [22], SRA [23] and UCSC Archaeal Genome Browser [24].

In this study we analyzed the transcriptome of 11 archaea: *Halobacterium salinarum* NRC-1, *Pyrococcus furiosus* DSM 3638, *Methanococcus maripaludis* S2, *Sulfolobus solfataricus* P2, *Nanoarchaeum equitans* Kin4-M, *Methanopyrus kandleri* AV19, *Sulfolobus acidocaldarius* MW001, *Haloferax volcanii* DS2, *Methanobolbus psychrophilus* R15, *Methanosarcina mazei* Gö1 and *Pyrococcus abyssi* [12–21]. Only *S. acidocaldarius* data did not present sufficient coverage to clearly show at least one TSSaRNAs signature. Therefore, our observations were made for 10 organisms. Archaeal transcriptomes for which dynamical information was available were highlighted in this work: *Halobacterium salinarum* NRC-1 [14], *Pyrococcus furiosus* DSM 3638 [12], *Methanococcus maripaludis* S2 [12] and *Sulfolobus solfataricus* P2 [12,13]. Original accession numbers for these datasets are: GSE13150, GSE18630, GSE38821, GSE26782, GSE44979, SRP028191, SRX188664. Datasets not available in public databases were obtained directly from publications. A brief description for each dataset used is provided in the **Table S3**.

The expression signal for putative TSSaRNAs locations is a distinct signature characterized by a sharp rise in signal that plateaus over a relatively small distance and then decays precipitously. Tiling array probe intensities and log ratio data for all growth curve time points were obtained from GEO and processed as described in [14]. Heatmaps for expression profiles over the growth curve were relative to a reference growth condition and visualized in Gaggly Genome Browser [31]. Raw RNA-seq datasets were processed by: i) trimming each library using FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to remove adapters; ii) mapping against appropriate reference genomes using Bowtie [32]; and iii) visualizing non-normalized reads coverage as a proxy for gene expression using the integrative tool Gaggly Genome Browser [31].

Cell cultivation and small RNA isolation

H. salinarum NRC-1 was grown in CM media, in a water bath incubator at 37°C with agitation of 125 r.p.m. Reference samples were cultured under standard growth conditions [26], at mid-log phase (OD₆₀₀≈0.5). Small RNAs for RNA-seq libraries and Total RNAs for dRNA-seq libraries and northern blot experiments were isolated using the MirVana RNA extraction kit (Ambion).

RNA-seq library preparation, sequencing and pre-processing

Two small RNA libraries (biological replicates) from *H. salinarum* NRC-1 were prepared for sequencing. Small RNAs at mid-log phase cultures were extracted. For each sample, 10 µg of small RNAs were treated with RNase-free DnaseI (Fermentas) in a final volume of 30 µL. The reaction was incubated for 45 min at 37°C and the RNA was purified using phenol/chloroform purification. 1 µg of treated small RNA was ligated to RNA 3' Adapter (RA3) using T4 RNA ligase 2 truncated (BioLabs) for 1 hour at 28°C, in the presence of RNase inhibitor. Once RA3 was ligated, we performed the RNA 5' Adapter (RA5) ligation using T4 RNA ligase in the presence of 10 mM ATP. cDNA was synthesized using specific oligos for 5' and 3' adapters using SuperScript III Reverse Transcriptase, according to Illumina Truseq protocol. cDNA libraries were amplified and samples were separated in a Novex 6% PAGE gel. cDNAs from 20 bp up to 230 bp were isolated from the gel and subjected to quantification and quality analysis.

The resulting double stranded cDNA was sequenced on Illumina Miseq v2 platform. Biological replicates were sequenced

in the same flow-cell using different indexes. Strand-specific sequencing was performed in MiSeq set to 151 cycles per manufacturer's instructions.

Reads were trimmed using FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/links.html) to remove adapters. Processed reads were aligned against *H. salinarum* NRC-1 reference genome (chromosome: NC_002607, plasmid pNRC100: NC_001869.1 and plasmid pNRC200: NC_002608.1) using Bowtie [32] with default parameters (except by “m” parameter, since we discarded ambiguous alignment). Overall, 3,489,281 aligned reads from biological replicates combined were considered in subsequent analysis.

RNA-Seq data were submitted to NCBI's SRA website under the accession number **SRP035406**.

TSSaRNA definition in *H. salinarum*

Since *H. salinarum* small RNA libraries were made without fragmentation, we can observe two sets of reads consistently aligned at the same position near the start codon of a gene: (i) reads marking the transcription start site (TSS) of the gene itself, truncated at diverse lengths up to 151 bp; and (ii) reads smaller than 151 bp consistently found with the same 3' end, thus, being full-length reads (**Figure S9**). Type (ii) reads are generally generated by TSSaRNAs.

We used relative enrichment of reads' aligned start coordinates as a parameter to automatically detect TSSaRNAs. We looked for the most frequent start coordinate near the start codon of a CDS. The search was performed in a window starting 50 bp upstream of the translation start site and comprising at maximum 20% of CDS length. To make sure that the TSS is reliable, reads must sum up more than 20 counts. This procedure can detect TSSs, but it is still necessary to split TSSaRNA and cognate gene signals. To isolate the TSSaRNA signal, the most abundant read smaller than 151 bp is defined as the TSSaRNA full-length sequence. All other reads starting at the same position are related to the cognate gene. To be conservative, TSSaRNA reads are only retained if they sum up at least 10 counts.

dRNA-seq library preparation, sequencing and analysis

Total RNAs were treated with TurboTM DNase (Ambion) and incubated with TerminatorTM 5'-phosphate-dependent exonuclease (Epicentre) (TEX+ sample) or only in buffer reaction (TEX-sample) at 30°C for 60 min, at proportion of 1 U TEX per 1 µg total RNA. Reaction products were purified with RNeasy MinElute Cleanup Kit (QIAGEN) and incubated with 1 U of Tobacco Acid Pyrophosphatase (TAP) (Epicentre) at 37°C for 1 hour in order to generate 5'-mono-phosphates RNAs able to bind to sequencing adapters. Reactions were purified again with RNeasy MinElute Cleanup Kit (QIAGEN).

Sequencing libraries were prepared with 1 µg of treated (TEX+) and untreated (TEX-) samples using a similar protocol described above for RNA-seq experiments. To ensure sequencing of a wider range of transcripts we increased the extension time on cDNA amplification step to 1 min and isolated molecules from 20 bp up to ~480 bp on the gel. Paired-end sequencing was performed on Illumina Miseq v2 platform using 300 cycles kit. Forward reads were trimmed and mapped to the reference genome using Bowtie [32] as previous described. 435,339 reads corresponding to TSSaRNAs were used in subsequent analysis. TSSaRNAs presenting at least a 95% reads enrichment in TEX+ library relative to the TEX- library were considered as primary transcripts.

Northern-blot

For Northern-blot analyzes, 30 µg of total RNA treated with RNase-free DNaseI (Fermentas) was separated on polyacrylamide gel (8% acrylamide:bisacrylamide [29:1], 8 M urea, 1xTris–borate–EDTA buffer). RNAs were transferred to Hybond-N+ membranes (GE Healthcare) and hybridized with ³²P-labeled oligonucleotides (5′-AGTGTGCTTGAAGAAGTCAACTTCG-CCTGTGCGCCATTGCAACT-3′ for VNG0101G and 5′-AAA-AGTGGCCGTGGGCAGCGGCCACCCGAT-3′ for VNG1213C) using Rapid-hyb buffer (GE Healthcare). Signals were detected by autoradiography using a M35A X-Omat Processor (Kodak). Genes encoding a conserved cold-shock protein (VNG0101G) (updated annotation: Supplementary Material 2 table from [14] and a probable exonuclease (VNG1213C) (updated annotation: UCSC Archaeal Genome Browser [24] and HaloLex project [33]) were chosen for this analysis.

Promoter and structural analysis of TSSaRNA sequences

DNA sequences of 11 bp around TSSaRNA 3′ ends were analyzed for conserved patterns using MEME with default parameters [34] in order to identify possible RNA polymerase pausing site motifs. Secondary structures of TSSaRNAs were predicted using the GeneRfold Bioconductor package interface for Vienna RNA library [35]. In this analysis, Gibbs Free Energy of predicted structures derived from TSSaRNAs sequences were compared to sequences from non-cognate genes derived from similar regions.

TSSaRNAs differential expression analysis

Differential expression of TSSaRNAs in *H. salinarum* NRC-1 was computed from a published dataset generated by tiling array hybridization of total RNA from 13 time points over a growth curve [14]. Using TSSaRNAs sequence coordinates information defined by single-base resolution RNA-seq, we revisited hybridization data and automatically selected a tiling array probe that best fits each TSSaRNA. The selected probe was required to have the highest TSSaRNA sequence coverage and, at the same time, should not cover any length beyond the TSSaRNA end (**Figure S10**). We compared the TSSaRNA representative probe intensity with the median intensity of the upstream region and also, with the intensity of the cognate gene. To be considered differentially expressed, this probe must have a substantial difference in relative intensity when compared to the other cognate gene probes and its surrounding (**Figure S11**). A TSSaRNA probe must show at least 10-fold difference relative to the overall relative intensity of its cognate gene: $V = M_{\text{TSSaRNA}} - M_{\text{cognate}} \geq 1$, where $M = \log_{10}(t/t_{\text{ref}})$, t_{ref} is taken at the reference time point in [14], t is taken at the growth curve time point when the second most different $|V|$ is seen, M_{cognate} is the median of all cognate gene probes starting beyond TSSaRNA 3′ end. The same procedure is also required for an upstream region to make sure that TSSaRNA probe is not a merely continuum of adjacent transcript signal. Therefore, a differentially expressed probe must also show at least a 2-fold difference relative to the overall relative intensity of an upstream region. This upstream region is 300 bp long, 120 bp away from TSSaRNA start (**Figure S10**). If, there is an annotated gene closer than 200 bp from the TSSaRNA start, the aforementioned region is ignored and the whole adjacent CDS region is considered for probe averaging.

RNA polymerase pausing computational model

A simple RNA polymerase pausing model was created (**Figure S6**) and implemented in R programming language (**File S1, File**

S2). The model attributes a waiting time for each base position along a virtual gene. For simplicity, this waiting time is taken to be 1 arbitrary time unit. A RNA polymerase pausing site is a position where a moving RNA polymerase stalls for more than the default waiting time. This time is called “stalled time” (Δt). There is an “intrinsic transcription initiation time interval” ($\Delta \tau$), which is the time it takes between two successive RNA polymerases to start their trajectory along the gene from the first base pair to the gene’s end. These two time interval parameters are the most important parameters. Other auxiliary parameters are: gene length L , pause position L' and total simulation elapsed time T . A RNA polymerase is not allowed to keep traveling along the gene if there is another one stalled at the next base pair. In this case it releases its transcript and detaches from DNA, terminating the transcription process. Also, the stalled RNA polymerase that blocked the previous one is not affected and only keep moving forward when its waiting time at pausing site is up.

Supporting Information

Figure S1 Data-mining on archaeal gene expression public datasets. For all archaea, the yellow arrow represents the cognate gene, blue dashed lines represent TSSaRNA estimated regions and horizontal axis represent organism’s genome coordinates. Heatmaps for **A, B, C** and **D** represent gene expression profiles over growth curves. Heatmaps are color-coded according to log10 expression ratios between each time point relative to reference growth condition samples. Light blue horizontal bars for **A, B, C** represent tiling array probe intensities for reference conditions. Dark blue points for **D, E, F, G** and **H** represent RNA-seq reads coverage data. Frames **I** and **J** were extracted directly, with minor adjustments, from published figures. Reads in **I** were originally from Jäger et al 2009’s Figure 1. Reads in **J** were originally from Toffano-Nioche et al 2013’s Figure S4. Light blue crosses for **K** and **L** represent Nimblegen tiling array probe intensities for the reference conditions. Red crosses for **K** and **L** represent Nimblegen tiling array probe which best matches the TSSaRNA. (PDF)

Figure S2 Additional examples of transcriptome mapping using small RNA sequencing. Strand-specific RNA-seq experiment was performed using non-fragmented small RNAs (20–230 pb). Panels A–D show examples of TSSaRNAs and their cognate genes. (A) VNG0725H, forward strand (B) VNG1182H, reverse strand (C) VNG2014H, forward strand. (D) VNG2658G, reverse strand. For each panel, the genome position and CDSs location and orientation (grey arrows) are indicated at the horizontal axis. The uppermost graphic show log2 transformed amount of reads covering a given genomic coordinate. Vertical dashed lines represents the TSSaRNA region. The intermediary graphic represents the amount of aligned reads whose start position maps to a given genomic coordinate, the “start profile”. (PDF)

Figure S3 Properties of the 179 TSSaRNAs identified by small RNA-seq and dRNA-seq. **A** – Size distribution. **B** – Distribution of the distances between TSSaRNA start position and cognate gene CDSs start codon position. **C** – Distribution of Pearson correlation between each TSSaRNA and its cognate gene. (PDF)

Figure S4 Schematic illustration of a putative signature if non-degradative processing biogenesis hypotheses would hold. Dark blue points represent RNA-seq reads coverage

data. The yellow arrow represents a gene. Green vertical bars represent mapped reads start positions along genome coordinates and their abundances. Light blue highlight represents TSSaRNA sequence region. The prediction illustrated by the figure **is not** found in *H. salinarum* sequencing experiments. (PDF)

Figure S5 Gibbs Free Energy distribution of secondary structure predictions for TSSaRNAs and regular near-TSS sequences. Histogram considers non-cognate genes sequences with similar localization and same size as the TSSaRNA median sizes. Vertical blue bars represent values for actual TSSaRNA sequences. (PDF)

Figure S6 Schematic illustration of the RNA polymerase pausing computational model. An arbitrary gene of length L bp is considered. Genomic position L' represents the pausing site. Each moment in time is depicted by successive drawings from the upper left panel downwards until the rightmost lower panel. Time passing is not represented in constant flux and downward vertical arrows illustrate the amount of time passed. Every $\Delta\tau$ units of time a new RNA polymerase arrives at position 1 bp and keep transcribing forward at a constant velocity of v bp/unit of time. Arriving at the pausing point, a RNA polymerase receives $v = 0$ and waits there for Δt units of time, leaving then again with the same velocity v until it reaches the end of the gene at position L and releasing the full-length transcript. The parameters $\Delta\tau$ and Δt are the most critical for the model and are called “time spent stalled” and “intrinsic transcription initiation interval”, respectively. If an incoming RNA polymerase encounter another RNA polymerase just a base pair downstream, it cannot go further and releases the DNA sequence freeing the transcript synthesized up to that position/moment. The RNA polymerases released due to downstream blocking are shown with their IDs inside their red circle representation and those still active are show with their IDs below. This illustration depicts several moments between the first RNA polymerase (ID #1) start at position 1 bp until it reaches the last position L bp, along with several RNA polymerases (IDs #2, #3, #5, ..., #n+2) that produced TSSaRNAs due to early transcription termination. (PDF)

Figure S7 Expression profiles of TSSaRNAs differentially expressed over a typical growth-curve relative to the control growth condition. Horizontal axis represent 13 growth curve points from different phases in a standard laboratory batch culture. Solid lines are TSSaRNA expression profiles and dashed lines their cognate gene expression profile. (PDF)

Figure S8 Simple RNA polymerase pausing computational model simulation. Simulated expression profiles of TSSaRNAs differentially expressed during a growth curve relative to the control condition. Tiling microarray output simulation for a 35 bp TSSaRNA in a 2 kb cognate gene. Varying the only two model parameters it is possible to generate situations in which the cognate gene expression level remains constant over time and TSSaRNA levels can vary almost arbitrarily. Panels A to D are build mimicking our experimental setup displayed in manuscript's Figures 4 and Figure S7 exploring the parameter space. Vertical-axis – $\log_2$ ratios between simulated quantity of transcripts in each time point (I_t) and the amount simulated at reference time-point (I_{ref}). Black solid line – TSSaRNA expression profile. Black dashed line – cognate gene expression profile (constant over time and arbitrarily set to the same value of reference condition). Panel E

shows all kinds of $\log_2$ ratios that can be obtained for a TSSaRNA probe and its cognate gene when scanning the parameter space: $\Delta\tau$ and Δt , “intrinsic transcription initiation interval” and “time spent stalled”, respectively. This example scans $\Delta\tau$ from 2 to 700 time units, Δt from 3 to 700 time units and simulates a 2 kb gene with a 35 bp TSSaRNA associated. Highlighted points in Panel E are examples of relatively constant TSSaRNA levels with an appropriately 3-fold difference in cognate gene level (light blue and purple circles, corresponding to $\Delta t = 46$ and $\Delta\tau = 12$ time units and $\Delta t = 14$ and $\Delta\tau = 8$ time units, respectively), and a 32-fold difference in TSSaRNA levels with relatively constant cognate gene levels (red and green circles, corresponding to $\Delta t = 250$ and $\Delta\tau = 4$ time units and $\Delta t = 250$ and $\Delta\tau = 84$ time units, respectively). Qualitatively, almost any complex dynamical behavior can be obtained if the pausing rhythm and the RNA polymerase arriving rate are jointly regulated by environmental clues. (PDF)

Figure S9 Illustration of the TSSaRNA identification procedure. The yellow arrow represents a cognate gene's coding sequence region (CDS). Green vertical bars represent mapped reads' start coordinates and their abundances. The grey circle highlights the most frequent start coordinate. The grey box zoom illustrates the set of reads which mapped to this specific coordinate position, composed by two populations: identical reads from TSSaRNAs and other reads that originate from cognate gene transcripts. Black horizontal bars represent regions, relative to translation initiation site position (start codon position), around which the search for the most frequent start coordinate position was performed. (PDF)

Figure S10 Illustration of the probe selection for TSSaRNA differential expression analysis. Light blue horizontal bars illustrate tiling array probe intensities for the reference condition. The best probe that represent the TSSaRNA is highlighted in red. The yellow arrow represents a cognate gene. Black dashed lines represent the TSSaRNA boundaries defined by RNA-seq. Gray boxes represent the regions used to calculate the neighbourhood expression intensity, which was compared to the TSSaRNA probe. (PDF)

Figure S11 Illustration of the method used to define a differentially expressed TSSaRNA in *H. salinarum*. The three-dimensional tiling microarray data (relative intensities vs growth-curve vs genome loci) is reduced to two dimensions and then to a single representative value. A TSSaRNA expression profile is considered distinct from its cognate gene if there are at least two time-points in which their relative intensities are at least 10-fold apart. Relative intensities are considered between a time-point and the reference growth condition. Relative intensities for TSSaRNAs are provided by the best tiling array probe (red horizontal bar) to fit a RNA-seq based TSSaRNA boundaries definition (see Figure S10). Relative intensities for their cognate genes are provided by the median of all non-overlapping adjacent tiling array probes (horizontal magenta dashed line). From all time-points along the growth-curve, the final differential expression value to be reported is from the data slice (fold-change vs position projection) where the 2nd top difference between TSSaRNA and cognate gene is found (blue vertical dashed line at $t = 5$). (PDF)

Table S1 TSSaRNAs identified in *Halobacterium salinarum* transcriptome. (XLS)

Table S2 Subset of TSSaRNAs identified in *Halobacterium salinarum* transcriptome differentially regulated relative to their cognate gene. (XLS)

Table S3 Detailed information on all 10 organism datasets used to detect TSSaRNA in archaea. (XLS)

File S1 RNAPol pausing computational model with multiple pausing sites (R language script). (R)

File S2 RNAPol pausing computational model with varying pausing rhythm (R language script). (R)

References

- Stern DL (2013) The genetic causes of convergent evolution. *Nat Rev Genet* 14: 751–764. Available: <http://www.ncbi.nlm.nih.gov/pubmed/24105273>. Accessed 11 December 2013.
- Gong H, Vu G-P, Bai Y, Chan E, Wu R, et al. (2011) A Salmonella small non-coding RNA facilitates bacterial invasion and intracellular replication by modulating the expression of virulence factors. *PLoS Pathog* 7: e1002120. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3174252&tool=pmcentrez&rendertype=abstract>. Accessed 31 May 2013.
- Lease RA, Smith D, McDonough K, Belfort M (2004) The Small Noncoding DsrA RNA Is an Acid Resistance Regulator in *Escherichia coli*. *J Bacteriol* 186: doi:10.1128/JB.186.18.6179.
- Yus E, Güell M, Vivancos AP, Chen W-H, Lluch-Senar M, et al. (2012) Transcription start site associated RNAs in bacteria. *Mol Syst Biol* 8: 585. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3377991&tool=pmcentrez&rendertype=abstract>. Accessed 20 January 2014.
- Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, et al. (2009) Tiny RNAs associated with transcription start sites in animals. *Nat Genet* 41: 572–578. Available: <http://www.ncbi.nlm.nih.gov/pubmed/19377478>. Accessed 5 March 2013.
- Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, et al. (2008) Divergent transcription from active promoters. *Science* 322: 1849–1851. Available: <http://www.ncbi.nlm.nih.gov/pubmed/19597342>.
- Cserzo M, Turu G, Varnai P, Hunyady L (2010) Relating underrepresented genomic DNA patterns and tRNAs: the rule behind the observation and beyond. *Biol Direct* 5: 56. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3583238&tool=pmcentrez&rendertype=abstract>.
- Henriques T, Gilchrist DA, Nechaev S, Bern M, Muse GW, et al. (2013) Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals. *Mol Cell*: 1–12. Available: <http://www.ncbi.nlm.nih.gov/pubmed/24184211>. Accessed 6 November 2013.
- Taft RJ, Hawkins PG, Mattick JS, Morris KV (2011) The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization. *Epigenetics Chromatin* 4: 13. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3170176&tool=pmcentrez&rendertype=abstract>. Accessed 16 October 2013.
- Buckley MS, Kwak H, Zipfel WR, Lis JT (2014) Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes Dev* 28: 14–19. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3894409&tool=pmcentrez&rendertype=abstract>. Accessed 11 July 2014.
- Jonkers I, Kwak H, Lis JT (2014) Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons: 1–25. doi:10.7554/eLife.02407.
- Yoon SH, Reiss DJ, Bare JC, Tenenbaum D, Pan M, et al. (2011) Parallel evolution of transcriptome architecture during genome reorganization. *Genome Res* 21: 1892–1904. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3205574&tool=pmcentrez&rendertype=abstract>. Accessed 14 February 2014.
- Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, et al. (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res* 20: 133–141. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2798825&tool=pmcentrez&rendertype=abstract>. Accessed 21 May 2013.
- Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, et al. (2009) Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol* 5: 1–16. Available: <http://www.ncbi.nlm.nih.gov/pubmed/19536208>.
- Chen Z, Yu H, Li L, Hu S, Dong X (2012) The genome and transcriptome of a newly described psychrophilic archaeon, *Methanobus psychrophilus* R15, reveal its cold adaptive characteristics. *Environ Microbiol Rep* 4: 633–641. Available: <http://www.ncbi.nlm.nih.gov/pubmed/23760934>. Accessed 14 February 2014.
- Jäger D, Sharma CM, Thomsen J, Ehlers C, Vogel J, et al. (2009) Deep sequencing analysis of the *Methanosarcina mazei* G61 transcriptome in response to nitrogen availability. *PNAS* 106: 21878–21882.
- Martens B, Amman F, Manoharadas S, Zeichen L, Orell A, et al. (2013) Alterations of the Transcriptome of *Sulfolobus acidocaldarius* by Exoribonuclease aCPSF2. *PLoS One* 8: e76569. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3792030&tool=pmcentrez&rendertype=abstract>. Accessed 25 December 2013.
- Su AAH, Tripp V, Randau L (2013) RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile *Methanopyrus kandleri*. *Nucleic Acids Res* 41: 6250–6258. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3695527&tool=pmcentrez&rendertype=abstract>. Accessed 5 February 2014.
- Toffano-Nicchio C, Ott A, Crozat E, Nguyen AN, Zymnicki M, et al. (2013) RNA at 92°C. *RNA Biol* 10: 1211–1220.
- Randau L (2012) RNA processing in the minimal organism *Nanoarchaeum equitans*. *Genome Biol* 13: R63. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3491384&tool=pmcentrez&rendertype=abstract>. Accessed 4 October 2013.
- Ammar R, Torti D, Tsui K, Gebbia M, Durbic T, et al. (2012) Chromatin is an ancient innovation conserved between Archaea and Eukarya. *Elife* 1: e00078. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3510453&tool=pmcentrez&rendertype=abstract>. Accessed 19 November 2013.
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41: D991–5. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531084&tool=pmcentrez&rendertype=abstract>. Accessed 11 December 2013.
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36: D13–21. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238880&tool=pmcentrez&rendertype=abstract>. Accessed 20 December 2013.
- Chan PP, Holmes AD, Smith AM, Tran D, Lowe TM (2012) The UCSC Archaeal Genome Browser: 2012 update. *Nucleic Acids Res* 40: D646–52. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245099&tool=pmcentrez&rendertype=abstract>. Accessed 23 June 2012.
- Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, et al. (2007) A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131: 1354–1365.
- Baliga NS, DasSarma S (1999) Saturation mutagenesis of the TATA box and upstream activator sequence in the haloarchaeal bop gene promoter. *J Bacteriol* 181: 2513–2518.
- Sharma CM, Hoffmann S, Darfeuille F, Reigier J, Findeiss S, et al. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464: 250–255. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20164839>. Accessed 5 March 2013.
- Jorjani H, Zavolan M (2014) TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. *Bioinformatics* 30: 971–974. Available: <http://www.ncbi.nlm.nih.gov/pubmed/24371151>. Accessed 16 July 2014.
- Amman F, Wolfinger MT, Lorenz R, Hofacker IL, Stadler PF, et al. (2014) TSSAR: TSS annotation regime for dRNA-seq data. *BMC Bioinformatics* 15: 1–11.

Acknowledgments

We thank FMRP Miseq sequencing facility; Carolina Marcano and Guilherme Mendes from Illumina Corp. We thank Prof. Angela K. Cruz laboratory for helping with radioactive experiments. We are specially grateful to Diego M. Salvanha for extremely helpful discussions on Gaggie Genome Browser usage; to all LaBiSiMi (Laboratório de Biologia Sistemática de Microorganismos) members specially Sílvia Epifânio and José Vicente Gomes Filho. We thank the anonymous reviewers for extremely helpful criticism that improved our work substantially.

Author Contributions

Conceived and designed the experiments: TK RZNV. Performed the experiments: LSZ FtC. Analyzed the data: LSZ FtC RZNV. Contributed reagents/materials/analysis tools: TK NSB. Wrote the paper: LSZ RZNV NSB TK. Critical advice: NSB.

30. Hot D, Slupek S, Wulbrecht B, D'Hondt A, Hubans C, et al. (2011) Detection of small RNAs in *Bordetella pertussis* and identification of a novel repeated genetic element. *BMC Genomics* 12: 207. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3110155&tool=pmcentrez&rendertype=abstract>. Accessed 20 December 2013.
31. Bare JC, Koide T, Reiss DJ, Tenenbaum D, Baliga NS (2010) Integration and visualization of systems biology data in context of the genome. *BMC Bioinformatics* 11: 1–8.
32. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2690996&tool=pmcentrez&rendertype=abstract>. Accessed 1 March 2012.
33. Pfeiffer F, Broicher A, Gillich T, Klee K, Mejia J, et al. (2008) Genome information management and integrated data analysis with HaloLex. *Arch Microbiol* 190: 281–299. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2516542&tool=pmcentrez&rendertype=abstract>. Accessed 11 July 2014.
34. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37: W202–8. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2703892&tool=pmcentrez&rendertype=abstract>. Accessed 29 January 2014.
35. Lucas A, Thermes C (2006) GeneRfold: R for genes and sequences, using viennaRNA package. <http://www.tbi.univie.ac.at/~ivo/RNA>.