



UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA
MOLECULAR

**CARACTERIZAÇÃO *IN SILICO* DE lncRNAs E GENES
LINHAGEM-ESPECÍFICOS ASSOCIADOS A
REGENERAÇÃO DE ESTRUTURAS COMPLEXAS EM
PEIXES**

ARTHUR FELIPE VASCONCELOS FERREIRA REIS

**BELÉM – PARÁ
AGOSTO/2023**



UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA
MOLECULAR

**CARACTERIZAÇÃO *IN SILICO* DE lncRNAs E GENES
LINHAGEM-ESPECÍFICOS ASSOCIADOS A
REGENERAÇÃO DE ESTRUTURAS COMPLEXAS EM
PEIXES**

ARTHUR FELIPE VASCONCELOS FERREIRA REIS

Dissertação submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal do Pará como requisito para obtenção do grau de Mestre em Genética e Biologia Molecular com ênfase em Bioinformática.

Orientadora: Prof.^a Dr.^a Josane de Freitas Sousa

**BELÉM – PARÁ
AGOSTO/2023**



UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA
MOLECULAR

**CARACTERIZAÇÃO *IN SILICO* DE lncRNAs E GENES
LINHAGEM-ESPECÍFICOS ASSOCIADOS A
REGENERAÇÃO DE ESTRUTURAS COMPLEXAS EM
PEIXES**

ARTHUR FELIPE VASCONCELOS FERREIRA REIS

BANCA EXAMINADORA:

Prof.^a. Dr.^a Josane de Freitas Sousa (ICB/UFPA - FMRP/USP – Orientadora)

Prof. Dr. Wilson Araújo da Silva Jr (FMRP/USP– Membro Externo)

Prof. Dr. Rafael Azevedo Baraúna (ICB/UFPA – Membro)

Prof.^a. Dr.^a Maria Paula Cruz Schneider (ICB/UFPA – Membro)

**BELÉM – PARÁ
AGOSTO/2023**

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

R375c Reis, Arthur Felipe Vasconcelos Ferreira.
 Caracterização in silico de lncRNAs e genes linhagem- específicos
 associados a regeneração de estruturas complexas em peixes / Arthur
 Felipe Vasconcelos Ferreira Reis. — 2023.
 65 f. : il. color.

 Orientador(a): Prof^ª. Dra. Josane de Freitas Sousa
 Dissertação (Mestrado) - Universidade Federal do Pará,
 Instituto de Ciências Biológicas, Programa de Pós-Graduação em
 Genética e Biologia Molecular, Belém, 2023.

 1. Regeneração. 2. Análises de transcriptomas. 3. RNAs
 não-codificantes. 4. Genes linhagem-específicos. I. Título.

CDD 576.5

AGRADECIMENTOS

No íterim da minha jornada para o desenvolvimento desta dissertação, diversas pessoas e instituições contribuíram, intelectual, moral e financeiramente. Findada esta etapa da minha carreira, não poderia deixar de agradecê-las.

Obrigado, Prof.^a Dr.^a Josane de Freitas Sousa, profissional e pesquisadora excepcional, pela paciência e disponibilidade na orientação desta dissertação. Grato pela sua compreensão durante este processo sendo fundamental para que eu pudesse concluir o Mestrado transpondo as intempéries. Isso, sem dúvida, fez toda a diferença. Meus mais sinceros sentimentos de gratidão.

Obrigado, Prof.^a Dr.^a Paula Schneider, pelo acolhimento durante o meu ingresso no PPGBM, uma vez que sou oriundo da Química, bem como por todo suporte.

Meu muito obrigado ao Laboratório de Evolução e Desenvolvimento pelo fornecimento dos dados para a obtenção dos resultados necessários para a finalização desta dissertação.

Agradeço aos meus pais, Helena e Luiz, por todo o suporte necessário. Afinal, sempre pude contar com vocês quando eu mais precisei. Obrigado!

Agradeço à Universidade Federal do Pará (UFPA) pela infraestrutura fornecida, ao CNPq pela concessão da bolsa de estudo e ao Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM) através da pessoa da Coordenadora Prof.^a Dr.^a Ândrea Kely Campos Ribeiro dos Santos estendo meus agradecimentos aos membros e aos colaboradores do mesmo que direta ou indiretamente contribuíram para a finalização desta etapa.

INSTITUIÇÕES E FONTES FINANCIADORAS



Universidade Federal do Pará – UFPA



Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq



Programa de Pós-Graduação em Genética e Biologia Molecular – PPGBM



Laboratório de Evolução e Desenvolvimento – Evo&DevoLab

SUMÁRIO

AGRADECIMENTOS.....	2
INSTITUIÇÕES E FONTES FINANCIADORAS.....	3
LISTA DE FIGURAS E TABELAS.....	6
RESUMO.....	8
ABSTRACT.....	9
1. INTRODUÇÃO.....	10
1.1. Regeneração de estruturas complexas em vertebrados.....	10
1.2. Perfis Globais de Expressão Gênica e Vias Moleculares Enriquecidas na Regeneração de Estruturas Complexas.....	14
1.3. Genes linhagem-específicos e RNAs não codificantes na regeneração.....	19
2. OBJETIVOS.....	26
2.1. Objetivo Geral.....	26
2.2. Objetivos Específicos	26
3. METODOLOGIA.....	27
3.1. Montagem de novo do transcriptoma de cauda em regeneração do <i>Protopterus annectens</i>	27
3.2. Workflow para seleção de linhagem-específicos com potencial codificante e com expressão alterada durante a regeneração.....	29
3.3. Workflow para seleção de potenciais RNAs longos não codificantes com expressão alterada durante a regeneração.....	30
3.4. Mapeamento dos transcritos no Genoma do <i>Protopterus annectens</i>	31
3.5. Geração de redes de interação por co-expressão.....	32
4. RESULTADOS.....	34
4.1. Análise de expressão gênica diferencial na cauda em regeneração de <i>Protopterus annectens</i>	34
4.2. Genes linhagem específicos com potencial codificante diferencialmente expressos durante a regeneração.....	36
4.3. Potenciais lncRNAs diferencialmente expressos durante a regeneração de cauda.....	39
4.4. Redes de interação de co-expressão entre os genes diferencialmente expressos na cauda em regeneração.....	43
5. DISCUSSÃO.....	46
5.1. Genes linhagem específicos com potencial codificante apresentam expressão alterada durante a regeneração de cauda.....	46
5.2. Identificação de potenciais transcritos candidatos à lncRNAs com expressão alterada durante a regeneração de cauda.....	48
5.3. Módulos de co-expressão ativos durante a regeneração de cauda.....	50
6. CONCLUSÕES.....	52
7. REFERÊNCIAS.....	53

Apêndice A – Estruturas secundárias provenientes da análise do RNAfold web suit.

Apêndice B – Tabela dos resultados para os 126 genes linhagem-específicos nas análises de: Expressão Diferencial, Pfam e os respectivos E-values associados e o CEMiTool com os respectivos módulos associados a cada transcrito.

Apêndice C – Tabela dos resultados para os 46 candidatos a lncRNAs linhagem-específicos nas análises de: Expressão Diferencial, o CEMiTool com os respectivos módulos associados a cada transcrito e o Annocript com as probabilidades de codificação. Os candidatos marcados na coloração verde são os que, após análise no RNAfold web suit apresentaram estrutura secundária termodinamicamente viável.

LISTA DE FIGURAS E TABELAS

- Figura 1** – Distribuição filogenética da capacidade de regenerar apêndices pares entre os vertebrados. Na parte superior, representação esquemática do endoesqueleto e esqueleto dérmico de apêndices de vertebrados. Na parte inferior, uma filogenia de vertebrados calibrada no tempo retrata o estado do conhecimento atual, no momento em que este estudo foi realizado. Linhagens contendo uma ou mais espécies capazes de regeneração de nadadeiras ou membros após amputação do endoesqueleto são indicadas em azul, aquelas incapazes indicadas em laranja e aquelas onde não existe informação em preto (adaptada de DARNET et al., 2019)..... **11**
- Figura 2** – Comparação do blastema e formação da estrutura em diferentes processos de regeneração. (a) Estágios da regeneração da mandíbula em anfíbios. Após a amputação, um epitélio de ferida se forma sobre o tecido remanescente da mandíbula. Células do blastema se acumulam na margem seccionada da mandíbula, originando-se de diversos tecidos maduros. Essas células proliferam, passam por um processo de repadronização e diferenciação, resultando em uma estrutura semelhante à mandíbula original, porém com características notáveis. (b) Estágios da regeneração de membros. A regeneração de membros também envolve a formação de um blastema. Após a amputação, um epitélio de ferida se forma na margem cortada do membro. As células do blastema, também provenientes de vários tecidos maduros, proliferam, passam por um processo de repadrão e rediferenciação, resultando na formação de todas as estruturas do membro ausente. O membro regenerado é idêntico ao original (adaptada de VIEIRA & McCUSKER, 2018)..... **13**
- Figura 3** – Via mTOR é ativada durante regeneração da nadadeira de *Polypterus senegalus*. Cortes histológicos congelados de nadadeiras em regeneração, pelos períodos de tempo indicados após a amputação, foram incubados com anticorpo primário anti-pS6 (Cell Signaling, #2215, que reconhece a forma fosforilada da proteína ribossomal S6) e posteriormente com anticorpo secundário Alexa-488 (Thermo-Fisher), visualizado em verde. Núcleos foram marcados com DAPI; hpa (horas pós-amputação); dpa (dias pós-amputação); barra de escala = 200 μ m. (Laboratório de Evolução e Desenvolvimento - UFPA)..... **18**
- Figura 4** – Via mTOR é necessária para a regeneração de nadadeira em *Polypterus senegalus*. Imagens representativas de nadadeiras de animais cujas nadadeiras foram amputadas e que foram mantidos após a amputação por 35 dias em solução com 0,1 % de DMSO (controle) ou 2,5 mM de rapamicina (inibidor de mTORC1). Registros fotográficos do processo de regeneração foram feitos nos tempos indicados; dpa (dias pós-amputação). No final do período de análise (35dpa), as nadadeiras regeneradas foram novamente amputadas e submetidas a diafanização (imagens à direita) para observação da estrutura esquelética. O experimento foi realizado com triplicatas biológicas em cada grupo e que apresentaram resultado semelhante aos mostrados aqui (Laboratório de Evolução e Desenvolvimento - UFPA)..... **18**
- Figura 5** - Expressão de genes linhagem específicos recém-identificados em um gênero de salamandra, *Notophthalmus viridescens*. Análise RT-PCR em tempo real identificou novos candidatos a genes relacionados a famílias proteicas em tecidos do coração, cérebro, baço, olho total, fígado, cauda, membro e pulmão de adultos (adaptada de LOOSO et al., 2016)..... **20**
- Figura 6** – Esquema resumido de alguns mecanismos de interação dos RNAs longos não codificantes (lncRNA) humanos com biomoléculas celulares (adaptada de KAZIMIERCZYK et al., 2020)..... **21**
- Tabela 1** – Estatísticas da montagem de novo do transcriptoma do *Protopterus annectens* (adaptada de Veríssimo et al., 2020). **28**
- Figura 7** – Fluxograma da metodologia utilizada para seleção de transcritos correspondentes a genes codificadores linhagem específicos e genes de RNAs não-codificantes a partir de dados de RNA-seq da cauda em regeneração do *P. annectens*..... **33**
- Figura 8** – Volcano Plots ilustrando a expressão diferencial dos 75.447 contigs derivados dos dados de RNA-seq durante os estágios pós-amputação de cauda do *Protopterus annectens*. (a) 1 dpa, (b) 14 dpa e (c) 21 dpa. Os pontos em verde, vermelho e cinza correspondem aos contigs com regulação negativa (Down Regulated), regulação positiva (Up Regulated) e ausência de significância estatística (Not

significant), respectivamente.....	35
Figura 9 – Volcano Plots ilustrando a expressão diferencial de alguns genes linhagem-específicos codificadores de proteína sem anotação derivados dos dados de RNA-seq e superexpressos durante os estágios pós-amputação do <i>Protopterus annectens</i> . (a) 1 dpa, (b) 14 dpa e (c) 21 dpa. Os tais genes linhagem-específicos estão denotados em vermelho escuro no presente gráfico. Os demais pontos em verde claro, vermelho claro e cinza claro correspondem aos contigs com regulação negativa (Down Regulated), regulação positiva (Up Regulated) e ausência de significância estatística (Not significant), respectivamente.....	37
Figura 10 – Distribuição de domínios proteicos conhecidos no grupo de 126 transcritos sem ortologia aparente, mas com potencial codificante.....	38
Figura 11 – Distribuição de probabilidade do potencial codificante dos 5964 contigs analisados pelo Annocript v. 2.0.1, com destaque para o valor de cutoff (linha amarela) para adesão aos critérios de candidatura a lncRNA, estabelecido em menos de 5% de probabilidade de potencial codificante.....	39
Figura 12 – Distribuição de probabilidade dos 46 transcritos identificados como possíveis candidatos a lncRNAs após a análise pelo Annocript v. 2.0.1.....	40
Figura 13 – Volcano Plots ilustrando a expressão diferencial de alguns genes linhagem-específicos correspondentes aos candidatos à lncRNAs derivados dos dados de RNA-seq e que estão superexpressos durante os estágios pós-amputação do <i>Protopterus annectens</i> . (a) 1 dpa, (b) 14 dpa e (c) 21 dpa. Tais candidatos a lncRNAs estão denotados em vermelho escuro no presente gráfico. Os demais pontos em verde claro, vermelho claro e cinza claro correspondem aos contigs com regulação negativa (Down Regulated), regulação positiva (Up Regulated) e ausência de significância estatística (Not significant), respectivamente.....	41
Figura 14 – Predição da estruturas secundária pelo programa RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) dos quatro candidatos a lncRNA diferencialmente expresso durante a regeneração de cauda em <i>Protopterus annectens</i> . (a) TRINITY_DN104611_c0_g2_i1 - MFE/CS: - 74.50 kcal/mol e - 74.50 kcal/mol. (b) TRINITY_DN123033_c16_g1_i1 - MFE/CS: - 75.10 kcal/mol e - 68.30 kcal/mol. (c) TRINITY_DN128349_c6_g5_i1 - MFE/CS: - 76.70 kcal/mol e - 76.70 kcal/mol. (d) TRINITY_DN128368_c9_g5_i1 - MFE/CS: - 55.30 kcal/mol e - 55.30 kcal/mol.....	42
Figura 15 – Gráfico estatístico de independência de escala (R2 versus β).....	43
Figura 16 – Gráfico estatístico da conectividade média (Mean connectivity versus β).....	44
Figura 17 – Perfis de Co-expressão para os 8 módulos (M1 à M8) gerados sob as condições amostrais em triplicatas. O eixo y (ordenada) dos gráficos representa valores médios relativos de expressão gênica e o eixo x (abscissas) assinalam as triplicatas para cada uma das condições experimentais. As colorações cinza, vermelho, verde e azul nos gráficos correspondem representativamente as regiões para as triplicatas das amostras de Controle, 1dpa, 14 dpa e 21 dpa, respectivamente).....	45

RESUMO

Peixes, salamandras e rãs são vertebrados que apresentam uma impressionante capacidade de regenerar estruturas complexas, como nadadeiras e membros inteiros. Análises comparativas de perfil de expressão gênica indicam que em torno de 35 % dos genes superexpressos em estruturas em regeneração de peixes pulmonados estão também superexpressos em salamandra, sugerindo a existência de um programa genético conservado de regeneração. Entretanto, dados de RNA-seq de estruturas em regeneração indicam que além dos genes com alta conservação em diferentes grupos de vertebrados, há um número grande de transcritos sem ortologia aparente e que apresentam expressão diferencial durante a regeneração. Esses transcritos podem incluir genes codificantes de proteínas linhagem específicos, ou RNAs longos não-codificantes, que geralmente são pouco conservados e que representariam também uma classe de genes linhagem-específicos. Com o objetivo de ampliar a compreensão dos mecanismos envolvidos na regeneração, investigamos neste trabalho, por análises *in silico*, transcritos sem anotação, através da análise dos transcriptomas de três estágios da regeneração da nadadeira caudal do peixe pulmonado *Protopterus annectens*. Desenvolvemos e aplicamos workflow para a identificação de potenciais genes-linhagem específicos codificadores ou não codificadores (lncRNAs) de proteínas. Estabelecemos redes de interação por co-expressão verificando a interação de tais genes-linhagem específicos com genes e vias conservadas. Nosso resultado revelou 8336 transcritos sem aparente ortologia com genes de outras espécies e que estão alterados durante a regeneração de cauda de *Protopterus annectens*. Desses, 2372 transcritos apresentam potencial codificante (ORF > 100 aminoácidos), com 126 deles apresentando domínios proteicos conhecidos. Entre os domínios proteicos identificados, a grande maioria corresponde a domínios presentes em proteínas derivadas de retrotransposons. Análises com programas especializados na identificação de RNAs não codificantes revelaram que dos 5964 transcritos sem a presença de uma ORF maior de 100 aminoácidos, 46 transcritos têm alto potencial de representarem lncRNAs linhagem-específicos. Uma análise de co-expressão revelou oito módulos de transcritos de expressão coordenada durante a regeneração e indicou a presença de vários transcritos linhagem-específicos como hubs desses módulos de co-expressão. Os resultados contribuem para uma melhor compreensão dos detalhes moleculares das diferentes categorias de genes envolvidos no processo de regeneração e a potencial interação de vias conservadas com genes linhagem específicos.

Palavras-Chaves: Regeneração, Análises de Transcriptomas, RNAs não-codificantes, Genes linhagem-específicos.

ABSTRACT

Fish, salamanders and frogs are vertebrates with an impressive ability to regenerate complex structures, such as fins and entire limbs. Comparative analyzes of the gene expression profiles indicate that around 35% of the genes overexpressed in lungfish regenerating structures are also overexpressed in salamanders, indicating an evolutionary conserved program of regeneration. RNA-seq data analysis show that in addition to the conserved genes, there is a large number of transcripts with no apparent orthology detected as differentially expressed during regeneration. These transcripts may include genes encoding lineage-specific proteins, or long non-coding RNAs, which are generally poorly conserved and also represent a class of lineage-specific genes. In this work, we investigated non-annotated transcripts differentially expressed during regeneration, through *in silico* analysis of transcriptomes from three stages of regenerating caudal fins of the lungfish *Protopterus annectens*. We applied workflows for the identification and characterization of potentially lineage-specific genes (both coding and non-coding). We established co-expression networks and searched for interactions between the lineage-specific genes and conserved genes or gene pathways. Our results revealed 8336 transcripts with no apparent orthology with genes from other species and which are altered during *Protopterus annectens* tail regeneration. Of these, 2372 transcripts presented coding potential (ORF > 100 aminoacids), with 126 of them having known protein domains. Among the identified protein domains, the vast majority correspond to domains present in proteins derived from retrotransposons. Analyzes with programs specialized in the identification of non-coding RNAs revealed that of the 5964 transcripts lacking an ORF greater than 100 aminoacids, 46 transcripts have a high potential to represent novel lncRNAs. A co-expression analysis revealed eight modules of coordinated expressed transcripts during regeneration and indicated the presence of several lineage-specific transcripts as hubs of these co-expression modules. Our results contribute to a better understanding of the molecular details of the different categories of genes involved in the regeneration process and the potential interaction of conserved pathways with lineage-specific genes.

Key words: Regeneration, Transcriptome Analysis, Non-coding RNAs, lineage-specific genes.

1. INTRODUÇÃO

1.1. Regeneração de estruturas complexas em vertebrados

Até onde se tem conhecimento, as salamandras, girinos e lagartos são os únicos tetrápodes que demonstram a capacidade de regenerar membros e caudas. No entanto, apenas as salamandras são capazes de reconstituir completamente essas estruturas, incluindo a medula espinhal, de maneira idêntica aos originais ao longo de toda a vida (VAGLIA, 1997). A regeneração de membros e caudas em salamandras, como *Ambystoma mexicanum*, tem sido amplamente estudada e caracterizada (BROCKES, 2014). Embora a regeneração de estruturas complexas, como membros e caudas, também tenha sido observada em alguns outros vertebrados, os exemplos são limitados. Por exemplo, a regeneração da cauda em rãs e sapos ocorre antes da metamorfose (SIMON, 2013), e a regeneração de nadadeiras foi observada em peixes pulmonados, que são o grupo irmão dos tetrápodes (CONANT, 1970; NOGUEIRA et al., 2016). Acreditava-se anteriormente que a capacidade de regeneração de membros em todas as suas estruturas, incluindo músculos, ossos e nervos, era uma característica exclusiva desses poucos grupos.

No entanto, evidências paleontológicas indicam que a capacidade de regenerar estruturas complexas, como membros e caudas, já existia em vertebrados há aproximadamente 300 milhões de anos (FRÖBISCH, 2014). Além disso, um estudo recente realizado por DARNET et al. 2019, demonstrou que diversas espécies de peixes atuais são capazes de regenerar completamente suas nadadeiras peitorais após a amputação em uma porção comparável à estrutura óssea dos membros dos tetrápodes (DARNET et al., 2019).

A regeneração de apêndices pares, como nadadeiras ou membros, apresenta uma distribuição mais ampla entre os vertebrados do que se acreditava inicialmente. Isso sugere que a capacidade de regenerar apêndices pares seja uma característica ancestral prévia à separação do grupo que originou os tetrápodes, tendo surgido no ancestral dos peixes (FRÖBISCH et al., 2015). Ao longo da evolução, essa capacidade foi parcial ou totalmente perdida em certos grupos, como os mamíferos, Figura 1.

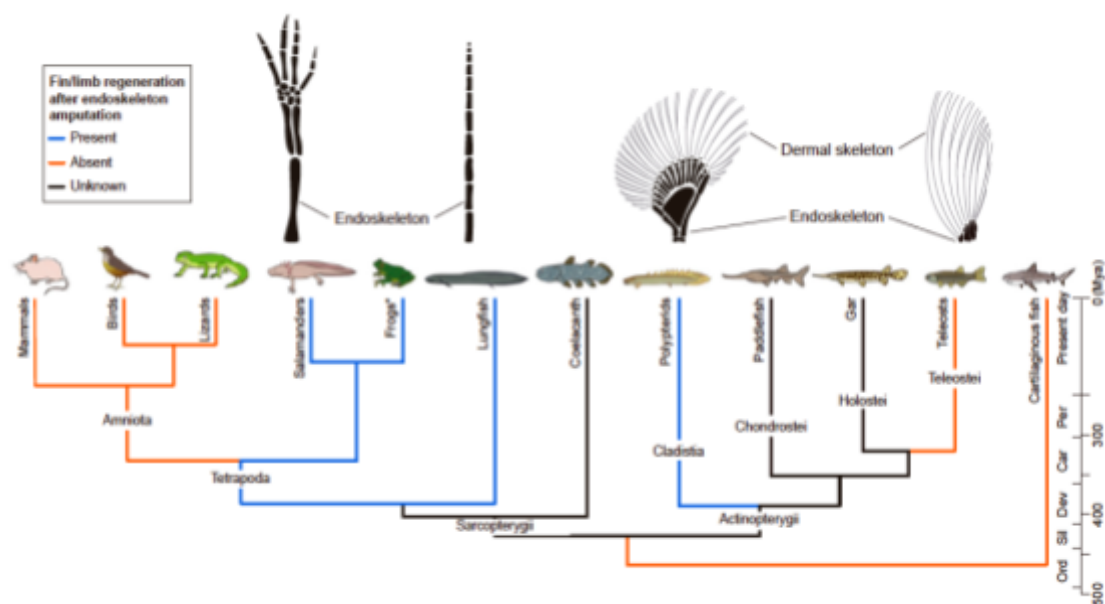


Figura 1 – Distribuição filogenética da capacidade de regenerar apêndices pares entre os vertebrados. Na parte superior, representação esquemática do endoesqueleto e esqueleto dérmico de apêndices de vertebrados. Na parte inferior, uma filogenia de vertebrados calibrada no tempo retrata o estado do conhecimento atual, no momento em que este estudo foi realizado. Linhagens contendo uma ou mais espécies capazes de regeneração de nadadeiras ou membros após amputação do endoesqueleto são indicadas em azul, aquelas incapazes indicadas em laranja e aquelas onde não existe informação em preto (adaptada de DARNET et al., 2019).

Há mais de 150 anos, foram descobertos os peixes pulmonados, pertencentes a um grupo filogeneticamente mais próximo dos tetrápodes (JORGENSEN, 2011; AMEMIYA et al., 2013; AMENIYA, 2013; MEYER, 1992). Essa proximidade taxonômica oferece uma oportunidade para estudar a transição evolutiva da água para a terra, bem como as mudanças moleculares associadas a essa transição (AMARAL; SCHNEIDER, 2018). Esses organismos podem servir como modelos ideais para investigar a origem da regeneração de membros nos tetrápodes. Atualmente, existem três espécies de peixes pulmonados: *Neoceratodus forsteri*, *Protopterus annectens* e *Lepidosiren paradoxa*, encontrados na Austrália, África e Amazônia, respectivamente. Assim como as salamandras, os peixes pulmonados têm a capacidade de regenerar a nadadeira caudal, uma extensão da coluna vertebral e medula espinhal (BRADBURY; BURNSIDE, 2019).

As pesquisas conduzidas por Nogueira et al. 2016, demonstraram que o peixe pulmonado *Lepidosiren paradoxa* é capaz de regenerar suas nadadeiras peitorais, seguindo etapas morfológicas e vias moleculares semelhantes às observadas na regeneração de membros em salamandras. Esses achados indicam que a compreensão dos mecanismos celulares e genéticos que governam a regeneração nos peixes pulmonados pode fornecer

informações valiosas sobre as bases celulares e genéticas envolvidas no processo regenerativo.

Recentes estudos realizados pelo mesmo grupo de pesquisa (VERÍSSIMO et al., 2020) evidenciaram que a regeneração da nadadeira caudal em um outro peixe pulmonado, o *Protopterus annectens*, apresenta características morfológicas e moleculares semelhantes à regeneração da cauda em salamandras. A reconstrução da cauda em *Protopterus annectens* segue um processo semelhante ao das salamandras, envolvendo a ativação de células proliferativas e a restauração das estruturas originais da cauda, tais como músculos, medula espinhal, esqueleto e a padronização dorsoventral (WANG, 2016; SANDOVAL, 2014; ECHEVERRI, 2002).

Um grupo adicional de interesse para investigar a regeneração de apêndices em vertebrados é composto pelos peixes do gênero *Polypterus senegalus*, que são representantes atuais de um dos grupos mais antigos entre os actinopterygii. Um exemplo notável é o *Polypterus senegalus*, que demonstra a capacidade de regenerar completamente suas nadadeiras peitorais, mesmo quando amputadas até o nível do endoesqueleto, uma estrutura homóloga aos membros dos tetrápodes. Análises comparativas revelam que aproximadamente 35% dos genes superexpressos durante a regeneração do *Polypterus* também estão superexpressos durante a regeneração de membros em salamandras, o que indica uma notável conservação do processo regenerativo (DARNET et al., 2019).

O processo regenerativo de apêndices pares ou da cauda, após a amputação, ocorre em etapas distintas. Inicialmente, ocorre a formação de uma estrutura transitória chamada blastema, composta por células indiferenciadas em proliferação (TANAKA, 2003). O blastema (GILBERT, 2008) é um aglomerado de células mesenquimais localizado entre a epiderme da ferida e a superfície da amputação (AKIMENKO et al., 2003). Em seguida, ocorrem processos de rediferenciação e morfogênese, que resultam na reposição das estruturas perdidas (STOCUM, 1999; YOKOYAMA, 2008; VIEIRA & McCUSKER, 2018),
Figura 2.

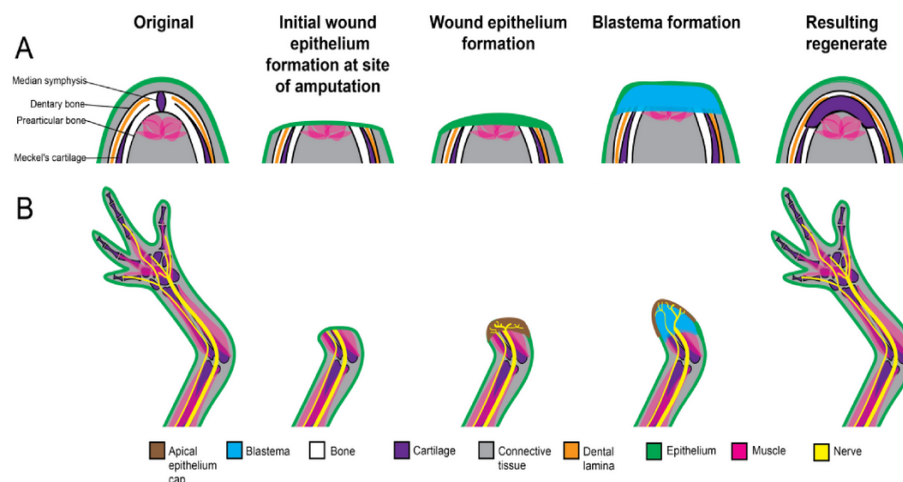


Figura 2 – Comparação do blastema e formação da estrutura em diferentes processos de regeneração. (a) Estágios da regeneração da mandíbula em anfíbios. Após a amputação, um epitélio de ferida se forma sobre o tecido remanescente da mandíbula. Células do blastema se acumulam na margem seccionada da mandíbula, originando-se de diversos tecidos maduros. Essas células proliferam, passam por um processo de repadronização e diferenciação, resultando em uma estrutura semelhante à mandíbula original, porém com características notáveis. (b) Estágios da regeneração de membros. A regeneração de membros também envolve a formação de um blastema. Após a amputação, um epitélio de ferida se forma na margem cortada do membro. As células do blastema, também provenientes de vários tecidos maduros, proliferam, passam por um processo de repadrão e rediferenciação, resultando na formação de todas as estruturas do membro ausente. O membro regenerado é idêntico ao original (adaptada de VIEIRA & McCUSKER, 2018).

As células do blastema apresentam plasticidade, o que lhes permite realizar o último estágio da regeneração: a formação de novos membros, caudas ou outras estruturas complexas. Além disso, possuem uma memória posicional e tecidual que lhes permite reconstruir apenas a porção perdida de forma adequada (SIMON, 2013). Estudos moleculares recentes têm começado a revelar a complexa arquitetura genética envolvida na regeneração de apêndices pares e da cauda dos tetrápodes, o que pode contribuir para a compreensão dos aspectos evolutivos da regeneração e para a identificação de alvos moleculares que possam aumentar o potencial regenerativo de organismos com capacidade limitada ou ausente de regeneração.

1.2. Perfis Globais de Expressão Gênica e Vias Moleculares Enriquecidas na Regeneração de Estruturas Complexas

No que se refere aos mecanismos celulares e genéticos que governam a regeneração em peixes pulmonados, estudos utilizando várias abordagens de análise da expressão gênica têm contribuído para uma melhor compreensão das vias moleculares envolvidas no processo de regeneração (HAAS & WHITED, 2017). Dentre as alterações observadas nas células do blastema até o momento, estão a ativação de oncogenes durante o fechamento da lesão, o aumento da expressão de genes envolvidos no remodelamento da matriz extracelular e no desenvolvimento, bem como a diminuição da expressão de genes relacionados à diferenciação muscular.

Nos últimos anos, uma abordagem amplamente adotada para caracterizar os perfis de expressão global durante a regeneração tecidual é a técnica de sequenciamento do transcriptoma, também conhecida como RNA-seq. Essa metodologia possibilita a identificação dos conjuntos de genes expressos, bem como a detecção de genes com expressão diferencial em tecidos e células, mesmo em organismos cujo genoma ainda não tenha sido sequenciado (LOW & TAMMI, 2017). Essa abordagem foi aplicada com sucesso pelo grupo do Laboratório de Evolução e Desenvolvimento da UFPA no estudo da regeneração de estruturas complexas em peixes como *Lepidosiren paradoxa* (NOGUEIRA et al, 2016), *Polypterus senegalus* (DARNET et al, 2019) e *Protopterus annectens* (VERÍSSIMO et al, 2020).

Em casos de organismos sem genoma disponível é possível realizar a montagem de novo do transcriptoma a partir dos dados de RNA-seq, permitindo a criação de um transcriptoma de referência para o mapeamento e quantificação da expressão gênica (LOW; TAMMI, 2017). Essa abordagem possibilita a condução de estudos de expressão gênica na regeneração de forma mais viável, superando as barreiras técnicas associadas à montagem e anotação de genomas extensos (BISCOTTI et al., 2016), como é o caso dos peixes pulmonados.

Recentemente, os genomas de peixes pulmonados, como *Neoceratodus forsteri* (peixe pulmonado australiano) (MEYER et al., 2021), *Protopterus annectens* (WANG et al., 2021) e *Polypterus senegalus* (BI et al., 2021), foram sequenciados. Os genomas do *Protopterus annectens* e do *Polypterus senegalus* estão disponíveis no National Center for Biotechnology

Information (NCBI) com os códigos de BioProject PRJNA701604 e PRJNA704032, respectivamente. Segundo informações fornecidas pelo NCBI, o genoma do *Protopterus annectens* consiste em 17 cromossomos, com 41834 transcritos associados e 37056 sequências de proteínas associadas. Já o genoma do *Polypterus senegalus* possui 19 cromossomos, com 48165 transcritos associados e 42230 sequências de proteínas associadas.

A análise de expressão gênica diferencial realizada por Nogueira et al.2016, permitiu a montagem de novo do transcriptoma da *Lepidosiren paradoxa* e do representante actinoptérigeo *Polypterus senegalus*. Esses estudos revelaram notáveis semelhanças entre o processo de regeneração desses organismos e a regeneração do apêndice da salamandra, conforme descrito em estudos anteriores (KNAPP et al., 2013; MONAGHAN et al., 2009; STEWART et al., 2013; WU et al., 2013).

Observou-se componentes de outras vias associadas à regeneração nos tetrápodes, incluindo a via Wnt (Wnt5a, Wnt5b, Axin2, Gsk3b, Ctnnb1), Fgf (Fgfr1), Bmp (Bmp1, Bmp4, Smad2), Shh (Ptch2, Gli2), Notch (Notch2), Tgf- β (Inhbb, Tgfb1) acredita-se que Tgfb1 e sua fibronectina alvo (Fn1) forneçam orientação direcional para as células de blastema (STOCUM, 2011), Egf (Vegfa, Megf6, Megf10), componentes e remodeladores de matriz extracelular (C1qtnf3, Col11a1, Fbn2, Mmp11, Adamts14), via de ácido hialurônico (Hyal2, Has2), resposta imune e inflamatória (Il11, Mdk, Nfkbiz) e manutenção de células-tronco (Sox4, Sall4, Chd2) (SCHNAPP et al., 2005; MONAGHAN et al., 2007; HO; WHITMAN 2008; WONG et al., 2011; PONOMAREVA et al., 2015; LI; ZHANG; AMAYA 2016; VITULO et al., 2017; MURAWALA et al., 2018; SINGH et al., 2018).

Observou-se que o gene responsável pela codificação da proteína semelhante a MARCKS (MLP), a qual desempenha uma variedade de funções desde o desenvolvimento embrionário até a resposta inflamatória, é regulado positivamente nos estágios iniciais da regeneração de membros. Essa proteína foi recentemente proposta como um iniciador da regeneração em salamandras (SUGIRA et al., 2016). Além disso, foi constatado que a regulação positiva do gene MLP também ocorre nos estágios iniciais da regeneração da nadadeira do peixe pulmonado.

As evidências sugerem perfis de expressão gênica semelhantes entre os organismos durante o processo de regeneração, o que reforça a hipótese de que os tetrápodes tenham herdado um programa de regeneração de membros durante a transição de nadadeira-membro. Essas análises também respaldam a ideia de um programa genético conservado de regeneração de apêndices em vertebrados, bem como a herança, por parte dos tetrápodes, de

um programa de regeneração de membros dos ancestrais dos peixes sarcopterígeos (AMARAL; SCHNEIDER, 2018).

Anteriormente, devido à ausência de um genoma de referência para o *Protopterus annectens* e o *Polypterus senegalus*, a análise dos dados de RNA-seq era realizada por meio da anotação dos transcritos montados, utilizando o BlastX para compará-los com um banco de proteínas de organismos com genes bem anotados, como o humano ou o zebrafish. A identificação da maioria dos genes diferencialmente expressos e das vias enriquecidas baseava-se, portanto, no conjunto de genes com ortologia em outros grupos de organismos.

Um exemplo de via de sinalização altamente conservada na escala evolutiva e que tem apresentado importante papel durante a regeneração é via de mTOR. O alvo mecanístico de rapamicina (mTOR) é uma serina/treonina quinase pertencente à família de proteínas quinases relacionadas à fosfoinosítídeo quinase (PIKK). Ela atua como subunidade catalítica de dois complexos proteicos distintos: mTORC1 e mTORC2 (mechanistic Target of Rapamycin Complex 1 e 2, respectivamente) (LAPLANTE, 2012). A ação da rapamicina, uma substância isolada em 1975 da bactéria de solo *Streptomyces hygroscopicus*, inibe apenas o mTORC1 e demonstrou atividade antifúngica, imunossupressiva e antitumoral (KIM & GUAN, 2019). A via de sinalização do mTOR é altamente conservada em eucariotos e desempenha um papel fundamental na integração de diversos sinais, promovendo síntese de proteínas e nucleotídeos, crescimento e proliferação celular, sobrevivência e migração (SAXTON & SABATINI, 2017).

O complexo mTOR desempenha um papel crucial na regulação de processos anabólicos e catabólicos em resposta a estímulos ambientais. Ele atua como um integrador de vias ativadas por fatores de crescimento e detecção de nutrientes, coordenando várias funções celulares e ligando a disponibilidade de nutrientes ao controle metabólico (SHIMOBAYASHI, 2014). A presença do mTOR em diferentes compartimentos celulares confere-lhe a capacidade de detectar estresse, sinalizar o crescimento e regular a autofagia. Especificamente, o complexo mTORC1 controla o equilíbrio entre o anabolismo e o catabolismo em resposta a diversos estímulos ambientais (SENGUPTA et al., 2010), fosforilando efetores envolvidos na síntese proteica, como a proteína ribossômica S6 quinase beta-1 (S6K1 ou RPS6KB1), especificamente no resíduo de aminoácido Thr389. Essa quinase fosforila a proteína ribossômica S6, promovendo a tradução de mRNAs envolvidos na biogênese do ribossomo. Portanto, a detecção com anticorpos anti-fosfo-S6-quinase

(pThr/pSer) pode ser utilizada como uma técnica indireta para detectar a atividade do complexo mTORC1 nos processos celulares.

A sinalização de mTOR desempenha um papel essencial no crescimento, proliferação e sobrevivência celular em diversos tecidos do organismo, incluindo o músculo esquelético e cardíaco. No músculo esquelético, o mTOR exerce controle sobre a sinalização anabólica e catabólica, afetando a hipertrofia muscular e a perda de massa muscular (YOON, 2017). Além disso, a sinalização de mTOR também desempenha um papel crítico na regulação da adipogênese, metabolismo lipídico, termogênese e síntese/secreção de adipocinas (CINTI, 2012).

Além de desempenhar um papel conservado na ativação da proliferação celular e controle do crescimento em eucariotos, estudos recentes têm revelado que a via mTOR também regula processos cruciais na determinação do destino celular, como auto-renovação, diferenciação e desdiferenciação celular em metazoários (HE et al., 2019; MIAO et al., 2020). Portanto, essa via tem sido implicada na regeneração em diversos contextos, desde tecidos específicos em mamíferos até estruturas mais complexas, como apêndices, e até mesmo em organismos inteiros (LUND-RICARD et al., 2020). Em modelos de peixe, como o zebrafish, estudos têm investigado os efeitos da inibição da via mTOR na regeneração de nadadeiras amputadas, especificamente no nível dos raios das mesmas (HIROSE et al., 2014; TAKAYAMA et al., 2018).

Em *Polypterus senegalus*, observou-se recentemente que a via mTOR é ativada a partir de 18 horas após a amputação da nadadeira em nível endocondral, com um aumento significativo da ativação nos dias 3 e 5 após a amputação (Figura 3, dados ainda não publicados). Além disso, verificou-se que com o uso de rapamicina, um inibidor de mTOR, a regeneração da nadadeira foi drasticamente inibida (Figura 4, dados ainda não publicados).

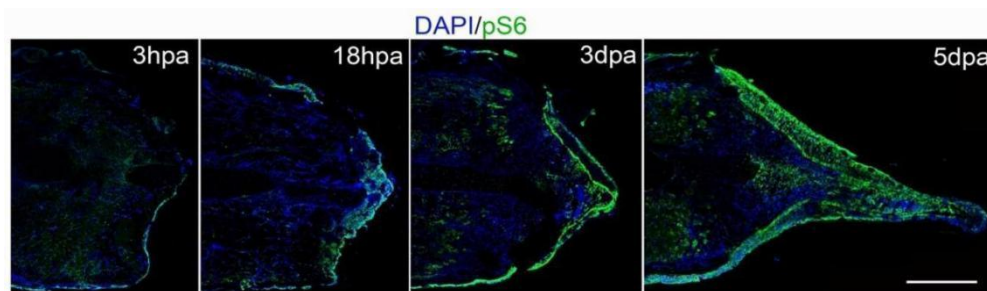


Figura 3 – Via mTOR é ativada durante regeneração da nadadeira de *Polypterus senegalus*. Cortes histológicos congelados de nadadeiras em regeneração, pelos períodos de tempo indicados após a amputação, foram incubados com anticorpo primário anti-pS6 (Cell Signaling, #2215, que reconhece a forma fosforilada da proteína ribossomal S6) e posteriormente com anticorpo secundário Alexa-488 (Thermo-Fisher), visualizado em verde. Núcleos foram marcados com DAPI; hpa (horas pós-amputação); dpa (dias pós-amputação); barra de escala = 200 mm. (Laboratório de Evolução e Desenvolvimento - UFPA).

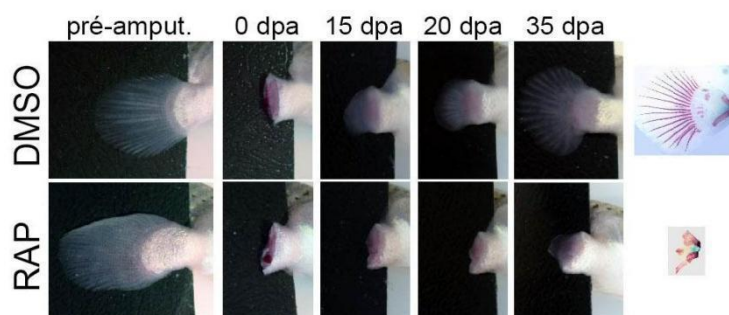


Figura 4 – Via mTOR é necessária para a regeneração de nadadeira em *Polypterus senegalus*. Imagens representativas de nadadeiras de animais cujas nadadeiras foram amputadas e que foram mantidos após a amputação por 35 dias em solução com 0,1 % de DMSO (controle) ou 2,5 mM de rapamicina (inibidor de mTORC1). Registros fotográficos do processo de regeneração foram feitos nos tempos indicados; dpa (dias pós-amputação). No final do período de análise (35dpa), as nadadeiras regeneradas foram novamente amputadas e submetidas a diafanização (imagens à direita) para observação da estrutura esquelética. O experimento foi realizado com triplicatas biológicas em cada grupo e que apresentaram resultado semelhante aos mostrados aqui (Laboratório de Evolução e Desenvolvimento - UFPA).

Embora a via mTOR seja altamente conservada em eucariotos, sabe-se que a capacidade regenerativa varia consideravelmente entre os organismos. Um fator que pode contribuir para atuação diferencial da via mTOR em organismos altamente regenerativos em comparação com organismos com menor capacidade regenerativa é a interação dessa via com o produto de genes linhagem específicos. Em mamíferos, foi observado que um peptídeo curto, codificado por um RNA anotado como um longo não codificante (lncRNA), tem a capacidade de modular negativamente a via mTOR, reduzindo sua atividade durante a regeneração muscular (MATSUMOTO et al., 2017). Desse modo, identificar potenciais genes

linhagem específicos alterados durante a regeneração pode contribuir para a caracterização de circuitos moleculares que modulam a atuação de vias de sinalização conservadas durante a regeneração.

1.3. Genes linhagem-específicos e RNAs não codificantes na regeneração

Os dados obtidos por meio da técnica de RNA-seq revelam que, além dos genes com ortologia conservada ao longo da evolução, há um conjunto significativo de transcritos que não apresentam similaridade aparente, mas apresentam variação de expressão durante o processo de regeneração. Esses transcritos podem incluir genes codificadores de proteínas específicos de determinada linhagem, que estão restritos a um táxon particular, ou podem corresponder a RNAs longos não codificantes (lncRNA). Estes últimos, em geral, possuem baixa conservação e podem representar uma subclasse de genes linhagem específicos. Por exemplo, estudos demonstraram que cerca de 70% dos lincRNAs (RNA não codificantes intergênicos longos), amplamente encontrados no genoma de mamíferos, não possuem homólogos em espécies separadas por mais de 50 milhões de anos (HEZRONI et al., 2015).

Estudos anteriores destacaram a contribuição de genes linhagem específicos no processo de regeneração de membros em salamandras a exemplo do *Notophthalmus viridescens*, Figura 5, onde houve a identificação de novos genes linhagem específicos (LOOSO et al., 2013; KUMAR et al., 2015). Em *Drosophila*, foi demonstrado que a deleção desses genes linhagem específicos pode ser letal, evidenciando sua importância essencial para esses organismos (CHEN et al., 2010). Além disso, uma pesquisa recente identificou uma nova família de proteínas exclusiva das planárias, que desempenha um papel no controle da proliferação e morte celular por meio da interação com a via mTOR, a qual é amplamente conservada em eucariotos (PASCUAL-CARRERAS et al., 2020).

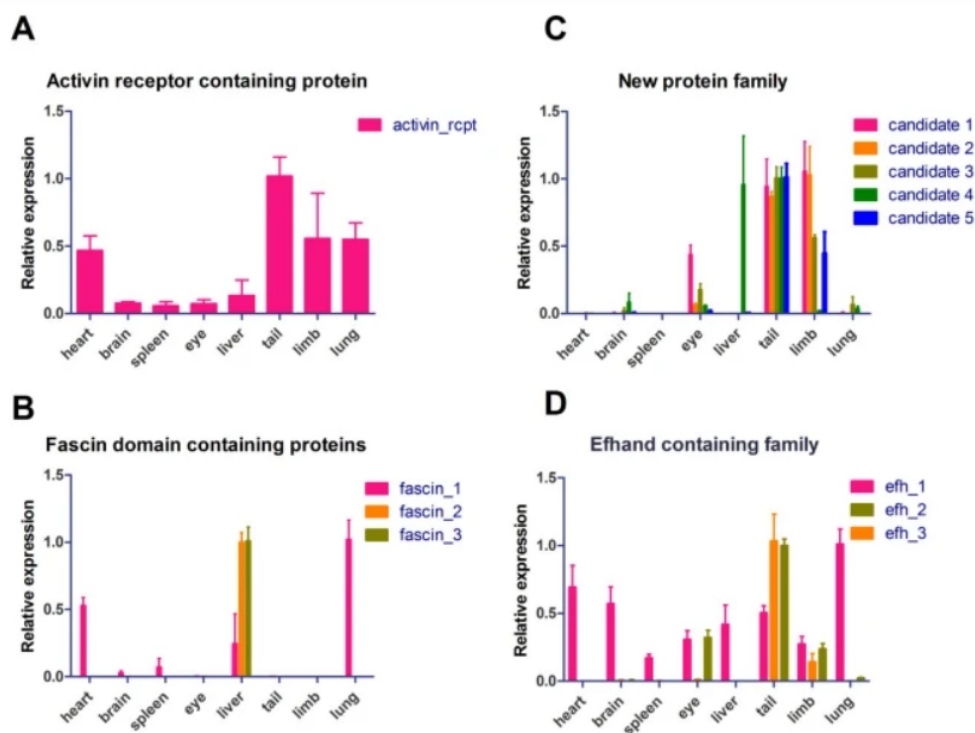


Figura 5 - Expressão de genes linhagem específicos recém-identificados em um gênero de salamandra, *Notophthalmus viridescens*. Análise RT-PCR em tempo real identificou novos candidatos a genes relacionados a famílias proteicas em tecidos do coração, cérebro, baço, olho total, fígado, cauda, membro e pulmão de adultos (adaptada de LOOSO et al., 2016)

Portanto, é plausível que genes linhagem específicos exerçam modulações em vias de sinalização conservadas, resultando em variações de regulação características em um determinado organismo. No contexto da regeneração de apêndices, a hipótese sugere que esses genes linhagem específicos interajam com um programa genético conservado, promovendo modulações distintas nesse programa e contribuindo para as diferentes capacidades regenerativas observadas entre os vertebrados. Apesar dos estudos mencionados anteriormente, o papel dos genes linhagem específicos codificadores de proteínas na regeneração, em geral, tem sido pouco explorado. Quanto aos genes linhagem específicos correspondentes aos RNA longos não codificantes (lncRNAs), a escassez de estudos ainda é notável.

Conforme mencionado anteriormente, é relevante ressaltar que os lncRNAs são uma classe de RNAs não codificantes que desempenham diversos papéis em processos biológicos, representando uma variedade de subclasses funcionais com mecanismos e funções biológicas distintas (CHEN, 2016). No passado, os RNAs não codificantes eram considerados regiões do DNA com pouca ou nenhuma relevância funcional. No entanto, pesquisas recentes revelaram

que eles possuem um papel crucial na regulação do genoma. Os lncRNAs em particular, representando aproximadamente 80% do transcriptoma não codificante humano, desempenham um papel significativo nesse contexto (HUI JIA et al., 2010; MATTICK & MAKUNIN et al., 2006).

Os RNAs não codificantes desempenham um papel na regulação de diversas vias biológicas, Figura 6. Sua atuação ocorre por meio da interação com outras biomoléculas durante diferentes etapas como a transcrição, a pós-transcrição, a tradução e a pós-tradução, influenciando a atividade dos genes (GUTTMAN et al., 2011; KHALIL et al., 2009; ØROM et al., 2010, STRUHL, 2007; van BAKEL et al., 2010; KAZIMIERCZYK et al., 2020).

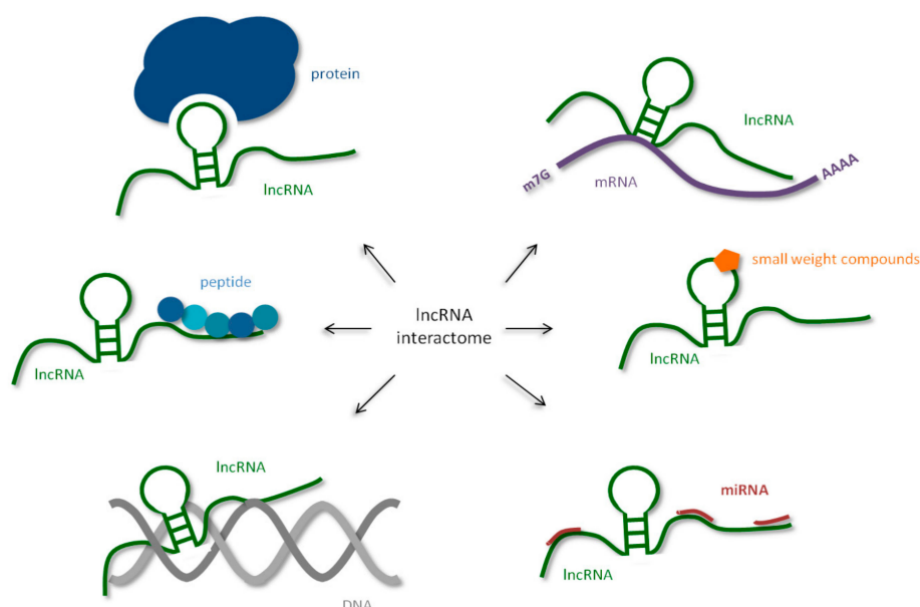


Figura 6 – Esquema resumido de alguns mecanismos de interação dos RNAs longos não codificantes (lncRNA) humanos com biomoléculas celulares (adaptada de KAZIMIERCZYK et al., 2020).

Além disso, alguns lncRNAs possuem open reading frames (ORFs, sigla em inglês) curtas e desempenham um papel duplo na regulação gênica. Eles podem sintetizar micropeptídeos que estabelecem interações intermoleculares e também interagem com estruturas moleculares por meio de seus nucleotídeos (KUNG et al., 2013). Além disso, muitos lncRNAs possuem funções conservadas entre diferentes espécies (CLARK et al., 2012; MARQUES & PONTING, 2009; ULITSKY et al., 2011). Devido à sua natureza instável e alta taxa de renovação no núcleo celular, os lncRNAs podem responder rapidamente

a estímulos externos (MERCER et al., 2009). Os lncRNAs exibem uma alta especificidade em relação a células e tecidos específicos. Eles desempenham um papel importante nas regulações epigenéticas específicas e podem ocorrer sem alterações na “maquinaria” transcricional (IWAKIRI et al., 2017).

Os lncRNAs são definidos principalmente pelo seu comprimento, sendo considerados transcritos com mais de 200 pares de base (200 bp) e com baixo ou nenhum potencial de codificação. Esses transcritos podem apresentar estruturas como Cap 5', cauda poli-A e serem compostos por múltiplos exons. Os lncRNAs podem ser classificados como de caráter antisenso, intergênicos, intercalados ou sobrepostos com genes codificadores de proteínas (MATTICK, 2009). Essa classificação está baseada nos locais de biogênese dos lncRNAs. Além disso, existem outras classificações que levam em consideração sua função, conservação, papel na regulação ou influência na estrutura da cromatina (MARCHESE et al., 2017; LAURENT et al., 2015).

lncRNAs são menos abundantes do que os mRNAs dentro da célula, embora sejam mais numerosos. Sua existência é breve e seus níveis de expressão variam entre diferentes tipos e subtipos de células. Eles exibem maior especificidade tecidual, bem como especificidade em relação ao estado de desenvolvimento e subtipo celular (YUNUSOV et al., 2016). O número total de lncRNAs no genoma humano ainda não é precisamente conhecido, pois novos transcritos são constantemente identificados e as bases de dados são atualizadas regularmente. Esses números podem variar consideravelmente entre bancos de dados, devido às diferentes definições do que constitui um lncRNA. Por exemplo, o GENCODE v34 (HARROW et al., 2012) lista 17960 lncRNAs, enquanto o LncBook (MA et al., 2019) relata mais de 270000 transcritos. O projeto FANTOM-CAT identificou 27919 lncRNAs, dos quais 19175 foram associados a alguma funcionalidade (HON et al., 2017). Apesar disso, apenas algumas centenas de loci de lncRNAs foram caracterizados funcionalmente até o momento. É provável que muitos outros lncRNAs sejam descobertos, e a maioria dos novos lncRNAs identificados por meio de vários métodos ainda precisa ser funcionalmente anotados.

lncRNAs desempenham papéis importantes no envelhecimento, doenças neurológicas e uma variedade de tipos de câncer (FERNANDES et al., 2019; CHEN et al., 2016; CHEN et al., 2018), tornando-os candidatos ideais para o estudo de vias moleculares envolvidas nessas condições. No entanto, em comparação com a predição computacional sistemática de sequências, funções e vias de biogênese de pequenos RNAs não codificantes, como os

microRNAs ou miRNAs, os lncRNAs ainda precisam ser compreendidos de maneira mais eficaz.

Isso ocorre devido à variação no comprimento das sequências, origem genômica e funções dos lncRNAs. Existem desafios conhecidos nos projetos de anotação de lncRNAs, incluindo: **i)** catalogação incompleta das características dos lncRNAs, **ii)** a diversidade de métodos experimentais utilizados para gerar informações sobre os transcritos, **iii)** baixa expressão dos lncRNAs, **iv)** especificidade celular e **v)** dinâmica da função dos lncRNAs em resposta a sinalizações epigenéticas. Os avanços recentes no processamento de Big Data e na aprendizagem de máquina estão auxiliando na superação desses desafios, embora ainda de forma limitada. No entanto, estratégias mais eficientes são essenciais para a predição e caracterização funcional dos lncRNAs.

No que se refere ao seu papel funcional, os lncRNAs desempenham diversas funções e estão principalmente envolvidos na regulação de genes codificadores de proteínas, em um mecanismo conhecido como regulação *cis*. No entanto, em alguns casos, eles também podem regular genes *em trans*, que estão localizados mais distantes no mesmo cromossomo ou em cromossomos diferentes. Além disso, os lncRNAs podem ser encontrados em diferentes regiões celulares, como núcleo, nucléolo, citoplasma e até mesmo nas mitocôndrias, o que geralmente se relaciona com seus modos de ação específicos.

Os lncRNAs não atuam de forma isolada, mas interagem com outros genes e proteínas por meio de complexas vias de interação (WANG & CHANG, 2011; ZHANG et al., 2019). Esses lncRNAs possuem várias funções bem estabelecidas, tais como: remodelação da cromatina (BERNSTEIN & ALLIS, 2005; GUPTA et al., 2010; LEE, 2009), formação de complexos por meio da interação com proteínas (JEON & LEE, 2011; LEE, 2011), fornecimento de suporte estrutural para interações moleculares (CAM et al., 2009), repressão da atividade genética através da interação com complexos de fatores de transcrição (PARASKEVOPOULOU & HATZIGEORGIOU, 2016; HU et al., 2017), co-regulação de genes codificadores de proteínas (FAN et al., 2019), inibição da RNA polimerase II (ESPINOZA et al., 2004), processamento de mRNAs (GONZALEZ et al., 2015), estabilização de processos pós-transcricionais (HE et al., 2019; MUNSCHAUER et al., 2018), afetação do processo de tradução (HU et al., 2014), empacotamento em vesículas extracelulares (CHEN et al., 2019; QUESENBERRY et al., 2015), atuando como precursores de miRNAs (CAI & CULLEN, 2007; GAI et al., 2018) e até mesmo codificação de micropeptídeos (ANDERSON et al., 2015; MATSUMOTO et al., 2017).

Os lncRNAs têm sido associados a uma ampla variedade de doenças. Devido às suas múltiplas funções e mecanismos de ação, esses transcritos estão envolvidos em diversos processos e vias intracelulares do organismo humano, o que sugere sua participação ativa no desenvolvimento de várias doenças. Aproximadamente 80% dos lncRNAs validados experimentalmente listados no banco de dados LncBook (MA et al., 2019) estão associados a mais de 400 doenças. Além disso, o LncRNADisease 2.0 (BAO et al., 2019) contém mais de 200 000 associações previstas entre lncRNAs e doenças.

Dentre as condições patológicas relacionadas aos lncRNAs, destacam-se o câncer de próstata, câncer de cólon, câncer de pulmão, doença de Alzheimer, doenças cardiovasculares, leucemia, diabetes, AIDS e doenças neurodegenerativas, nas quais mutações ou desregulação dos lncRNAs desempenham um papel significativo (CHEN et al., 2016). Por exemplo, o gene maternally expressed 3 (MEG3) transcreve um lncRNA de aproximadamente 1700 nucleotídeos que atua como componente antitumoral, demonstrando a capacidade de inibir a proliferação de células cancerígenas (ZHANG et al., 2003). Outro exemplo é o lncRNA HOTAIR, que apresenta superexpressão em tecidos cancerosos, comparado aos tecidos não cancerosos adjacentes, em pacientes com carcinoma hepatocelular (HCC) submetidos a transplante de fígado, tornando-se um biomarcador para prever a recorrência do tumor (YANG et al., 2011).

O entendimento dos mecanismos pelos quais os lncRNAs contribuem para a etiologia das doenças é crucial para o diagnóstico, tratamento, prognóstico e desenvolvimento de terapias eficazes. Inicialmente considerados como transcritos semelhantes aos mRNAs, porém não codificadores de proteínas, estudos recentes revelaram características distintas dos lncRNAs em relação aos mRNAs (QUINN, 2016). A capacidade dos lncRNAs de regular genes codificadores de proteínas é uma característica especialmente relevante.

Recentes avanços no sequenciamento do transcriptoma têm possibilitado a descoberta de milhares de lncRNAs em diversas espécies (HEZRONI et al., 2015). Embora alguns lncRNAs tenham demonstrado desempenhar papéis importantes em diversos processos biológicos, as funções e mecanismos da maioria desses transcritos ainda permanecem em grande parte desconhecidos. Existem dois desafios significativos que surgem entre o sequenciamento do transcriptoma e a caracterização funcional dos lncRNAs: a identificação de genes genuinamente não codificantes a partir da montagem de novo dos transcriptomas e a seleção dos centenas de candidatos a lncRNAs resultantes para análises experimentais futuras.

Alguns estudos têm destacado a participação dos lncRNAs em processos de regeneração tecidual, como a regeneração de nervos (YAO, 2019) e a regulação de funções e mecanismos no desenvolvimento, degeneração e regeneração da cartilagem, especialmente em doenças como osteoartrite e degeneração do disco intervertebral (ZHU et al., 2019). Além disso, uma pesquisa recente identificou um lncRNA como um regulador específico do desenvolvimento de membros em humanos e camundongos (ALLOU et al., 2021). No entanto, apesar das evidências do papel importante dos lncRNAs nos processos biológicos, pouco se sabe sobre o papel dessas moléculas em organismos não modelo (SALABI, 2021), e há uma escassez de estudos que caracterizem o papel dos lncRNAs no processo de regeneração de apêndices em vertebrados.

Portanto, a utilização de conjuntos de dados contendo transcriptomas de modelos de regeneração proporciona uma oportunidade para identificar genes linhagem específicos, tanto codificantes quanto não codificantes, durante o processo regenerativo de estruturas complexas. Além disso, é possível prever suas interações com vias de sinalização conservadas, como a via mTOR, o que pode fornecer insights importantes sobre os genes e vias de sinalização ativados durante a regeneração em estruturas complexas. Essas informações ajudam a elucidar as diferenças na capacidade regenerativa entre os vertebrados (NOGUEIRA et al., 2016; DARNET et al., 2019; VERÍSSIMO et al., 2020). A análise desses dados permite também a comparação dos programas genéticos empregados durante a regeneração em diferentes grupos de organismos, fornecendo detalhes moleculares para uma análise evolutiva do processo de regeneração. Isso pode revelar pistas sobre como ampliar a capacidade regenerativa, que é relativamente baixa em mamíferos.

Quando considerados os genes com ortologia entre os vertebrados, o programa de expressão gênica de regeneração de estruturas complexas em peixes pulmonados e de alguns representantes dos actinoptérígios é bastante semelhante ao empregado por salamandras. Isso indica a existência de um programa genético conservado de regeneração. Entretanto, um grande número de transcritos sem ortologia aparente é geralmente detectado nos perfis de expressão gênica de tais estruturas em regeneração. Embora pequena parte desses transcritos possam representar estruturas não caracterizadas, muitos correspondem a genes linhagem-específicos (codificantes e não codificantes) que também contribuem para o processo de regeneração.

Assim, a caracterização *in silico* desses transcritos sem ortologia contribuirá para identificar potenciais genes linhagem-específicos e suas possíveis interações com vias

moleculares conservadas. Isso fornecerá uma estimativa do impacto de genes linhagem-específicos no programa genético do processo de regeneração de estruturas complexas em vertebrados.

2. OBJETIVOS

2.1. Objetivo Geral

Identificação de transcritos correspondentes a genes linhagem-específicos codificantes e não codificantes potencialmente envolvidos na regeneração de nadadeiras do peixe pulmonado Sarcopterygii *Protopterus annectens*.

2.2. Objetivos Específicos

- Aplicação de workflow para identificar e selecionar transcritos sem anotação em banco de dados que estejam diferencialmente expressos durante a regeneração através da análise comparativa do transcriptoma em três estágios da regeneração do *Protopterus annectens*;
- Determinação de transcritos, diferencialmente expressos, que apresentem ou não potencial codificante;
- Identificação, considerando os que possuem potencial codificante, os transcritos que estejam associados a famílias proteicas específicas;
- Utilização de ferramenta apropriada para os transcritos sem potencial codificante, a fim de identificar possíveis candidatos a lncRNAs;
- Estabelecimento de redes de interação por co-expressão dos genes diferencialmente expressos na regeneração para identificar possíveis interações entre os genes linhagem-específicos e os demais.

3. METODOLOGIA

3.1. Montagem *de novo* do transcriptoma de cauda em regeneração do *Protopterus annectens*

A montagem do transcriptoma que utilizamos no presente estudo já foi realizada e devidamente publicada (VERÍSSIMO et al, 2020) pelo Laboratório de Evolução e Desenvolvimento da Universidade Federal do Pará. A primeira etapa consistiu na remoção de sequências adaptadoras e as leituras (reads) de baixa qualidade utilizando o programa Trimmomatic (BOLGER; LOHSE; USADEL, 2014) a partir dos seguintes parâmetros: HEADCROP: 21, ILLUMINACLIP: primer_bio.fa: 2:30:10, SLIDINGWINDOW: 4:20, MINLEN: 50.

Após esta etapa, a qualidade dos reads foi verificada utilizando o programa fastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc). Posteriormente as duplicatas técnicas de leituras únicas ou emparelhadas foram removidas, utilizando o programa CD-HIT-DUP, e então avaliamos novamente a qualidade dos reads utilizando fastQC. O transcriptoma de referência foi montado de novo através da ferramenta Trinity (GRABHERR et al., 2011), utilizando os parâmetros a seguir: min_contig_length 200 e min_kmer_cov 3. Após a montagem, foi gerada uma análise estatística utilizando a ferramenta assembly-stats e então foi realizada a busca por genes ortólogos usando o programa BUSCO (SIMÃO et al., 2015). Os transcritos redundantes que o programa Trinity gerou foram removidos com auxílio da ferramenta EvidentialGene (<http://arthropods.eugenics.org/EvidentialGene/trassembly.html>) utilizando o pipeline do programa. Então novamente foram realizadas análises com os programas assembly-stats e BUSCO.

Os dados brutos e montados do transcriptoma foram importados para o programa CLC Genomics Workbench, onde foi utilizado o plugin Molecular Index para mapear as leituras de sequência (reads) cada condição de regeneração e cauda intacta contra o transcriptoma montado. Para cada execução, todos os conjuntos de leituras foram mapeados para o transcriptoma de referência usando o CLC com parâmetros padrão (Qiagen). Em seguida, realizou-se a anotação dos genes bem como com a análise de suas respectivas expressões diferenciais. Os dados de expressão por transcrito foram somados por agrupamento de genes homólogos de humanos usando um script bash (HHGC). Como descrito anteriormente (DARNET et al., 2019), os HHGCs foram definidos agrupando transcrições com E-value de 10^{-3} quando comparados por BlastX com o banco de dados Humano NCBI RefSeq (11/2016).

Para cada HHGC, a expressão foi calculada em TPM e a comparação foi baseada no teste t considerando quatro condições (estrutura complexa intacta e estrutura complexa com 1 dpa, 14 dpa e 21 dpa, onde dpa são os dias pós amputação) com três repetições biológicas independentes. A matriz de similaridade entre as execuções foi calculada usando valores de TPM para cada HHGC, usando a correlação de classificação de Spearman no software Morpheus (<https://software.broadinstitute.org/morpheus>). Os genes foram considerados diferencialmente expressos, negativamente ou positivamente expressos, com significância estatística quando os valores de Fold Change (FC) são maiores que 2 ou menores que -2 e quando o false discovery rate (FDR) ajustado pelo p-value são menores que 0,05 (FDR < 0.05), respectivamente. A Tabela 1 mostra as estatísticas da montagem de novo do transcriptoma do *P. annectens* já devidamente descrita com sua qualidade atestada por Veríssimo et al, 2020.

Number of reads	279,685,470
Number of mapped reads	221,312,168
Number of contigs:	75,447
Number of contigs >=1kb:	21,637
Number of contigs in N50:	10,769
Min contig length:	201
Max contig length:	17,116
Mean contig length:	997.51
Standard deviation of contig length:	1250.86
Median contig length:	442
N50 contig length:	1,987
Number of bases in all contigs:	75,258,961
Number of bases in contigs >=1kb:	53,191,743
%GC	39.95

Tabela 1 – Estatísticas da montagem de novo do transcriptoma do *Protopterus annectens* (adaptada de Veríssimo et al., 2020).

3.2. Workflow para seleção de linhagem-específicos com potencial codificante e com expressão alterada durante a regeneração

Fez-se uma análise sistemática do transcriptoma da nadadeira caudal em regeneração, gerados por nosso grupo (VERÍSSIMO et al, 2020), para catalogar os potenciais genes linhagens-específicos codificadores de proteínas com expressão alterada durante a regeneração que não tenham anotação nos bancos de dados após busca por similaridade utilizando-se BLAST2GO e BlastX contra o UniProtKB/Swiss-Prot.

Para isso, criou-se um pipeline de análise que possui os seguintes critérios para a identificação de genes linhagem-específicos a partir das tabelas de expressão diferencial de RNA-seq do *P. annectens*:

- (i) seleção dos contigs sem anotação nos bancos de dados e com no mínimo 50 reads e apresentando superexpressão (FC maior que 2 no tecido em regeneração em comparação com o controle e com FDR menor que 0,05). Tais dados estatísticos foram obtidos anteriormente com o auxílio do software CLC Genomics Workbench;
- (ii) Após a filtragem dos contigs superepressos na etapa anterior (i), submeteu-se os mesmos para a análise no Transdecoder v3.5.0 (<http://transdecoder.github.io>) com a utilização dos parâmetros default do software, gerando assim um proteoma utilizado na etapa seguinte com aqueles que possuírem uma fase aberta de leitura (ORFs, sigla em inglês) maior que 100 aminoácidos (ORF > 100 aa).
- (iii) Para a identificação de domínios proteicos, utilizou-se o proteoma deduzido gerado pela etapa (ii), compostos por contigs com ORF preditas > 100 aa e consequentemente sequências proteicas deduzidas que sequencialmente foram submetidos para a identificação dos domínios proteicos no HMMER v3.2.1 (<http://hmmer.org/>) com auxílio do banco de dados proteicos Pfam (<https://pfam.xfam.org/>) (MISTRY et al.,2020), utilizando os parâmetros default e E-value = 1 para uma busca menos rigorosa no sentido de detectar uma maior similaridade com domínios proteicos e com posterior avaliação da significância dos achados através da seleção daqueles que apresentaram menores valores de E-value para cada hit.

3.3. Workflow para seleção de potenciais RNAs longos não codificantes com expressão alterada durante a regeneração

Para a identificação de potenciais candidatos a lncRNAs no nosso transcriptoma da cauda em regeneração, empregou-se o pipeline Annocript, descrito por Musacchia et al., 2015. Esse protocolo tem sido amplamente utilizado em estudos recentes, como o de Ceschin et al. 2020, para a identificação de lncRNAs. O principal módulo do pipeline Annocript é o CPC2 - Coding Potential Calculator 2 (KANG et al., 2017). O CPC2 utiliza a Support Vector Machine (SVM), que é um algoritmo de aprendizado de máquina capaz de analisar e reconhecer padrões moleculares. Esses padrões são avaliados com base em quatro características principais: comprimento das ORFs, integridade das ORFs e valores de pontos Isoelétricos (pI) (KANG et al., 2017).

O Annocript v. 2.0.1 é capaz de identificar candidatos potenciais a lncRNAs com base em critérios específicos. Esses critérios incluem o tamanho do transcrito, que deve ser maior que 200 nucleotídeos, a ausência de similaridade com proteínas ou domínios proteicos, a presença mínima ou ausência de ORFs com menos de 100 aminoácidos, a ausência de similaridade com outros ncRNAs presentes no banco de dados Rfam (<https://rfam.xfam.org>), e um potencial codificante menor que 5% ou potencial não-codificante maior que 95%. No presente estudo, foi dada ênfase à ausência total de ORFs nos transcritos, considerando a possibilidade de que lncRNAs possam apresentar micropeptídeos em suas estruturas.

Por último, realizou-se uma análise adicional com os 46 candidatos à lncRNAs linhagem-específicos no RNAfold web server (GRUBER et al., 2008) para predição de suas estruturas secundárias. Tal RNAfold é um websuit onde a análise deu-se através do fornecimento das sequências de interesse com a utilização dos parâmetros *default*. Tais parâmetros consideram as interações dos nucleotídeos das sequências fornecidas e, através da termodinâmica estatística das interações intermoleculares de tais nucleotídeos, gera uma população de estruturas através da distribuição de Poisson–Boltzmann considerando os aspectos termodinâmicos de interação intermolecular, assim obtendo a estrutura secundária centroide (CS, centroid structure) a qual é a representante ou é uma tendência de toda a população gerada, bem como a o fornecimento da estrutura de menos energia livre de Gibbs (MFE, minimum free energy).

3.4. Mapeamento dos transcritos no Genoma do *Protopterus annectens*

Como o genoma do *Protopterus annectens* tornou-se disponível durante a execução deste projeto, incluímos uma etapa de mapeamento dos transcritos de cauda em regeneração em nossa análise, como forma de garantir que os transcritos selecionados por nossas pipelines não representam contaminantes ou artefatos de sequenciamento. As metodologias de mapeamento de contigs são geralmente utilizadas em análises de expressão gênica com dados de RNA-Seq. O mapeamento é um processo essencial para a quantificação dos mesmos bem como para a predição de onde os contigs são originários em um genoma.

Bowtie 2 v. 2.5.1 (<http://bowtie-bio.sourceforge.net/index.shtml>) é um alinhador de sequências rápido e apresenta melhor alinhamento utilizando contigs acima de 50pb (LANGMEAD & SALZBERG, 2012). Ainda, o mapeador Bowtie2 utiliza a abordagem sem junção de exons (unspliced read aligner), utilizando a aplicação da transformada de Burrows-Wheeler para indexar o genoma de referência. Os parâmetros utilizados durante a simulação do Bowtie 2 foram: (i)- Indexação do genoma através da execução do comando `bowtie2-build`; (ii)- Execução do mapeamento com o termo `-p 8`, relacionado aos 8 processadores utilizados; (iii)- Utilização do preset `--very-sensitive-local` (`-D 20 -R 3 -N 0 -L 20 -i S, 1, 0.50`) para a execução do alinhamento local dos contigs com a possibilidade de omitir as bases das pontas dos mesmos ("soft clipped") para maximizar a pontuação do alinhamento, tal característica deste preset torna a simulação mais lenta tendo como contrapartida tendo uma maior sensibilidade e precisão do processo.

A etapa de quantificação é baseada na contagem dos contigs (CONESA et al., 2016). A contagem dos alinhamentos foi feita com auxílio da ferramenta SAMTools (<http://samtools.sourceforge.net/>), por meio de comandos básicos do Linux e utilizando o pacote Bioperl (STAJICH et al., 2002). O SAMTools, auxilia na manipulação de arquivos de saída de alinhamentos, formato SAM e BAM, os quais possuem um formato de texto onde os campos são separados por tab ("t"). Juntamente com comandos como `grep`, `cut`, `wc` e outros foi possível quantificar os contigs sem utilização custosa da memória. Todas as análises foram executadas em um computador com processador Intel Core i7, 32GB de memória RAM e 8 núcleos, utilizando Ubuntu 21.10.

3.5. Geração de redes de interação por co-expressão

Utilizou-se todos os dados de expressão dos transcriptomas gerados de cauda e nadadeiras em regeneração do *Protopterus annectens* (VERÍSSIMO et al, 2020) para analisar padrões de co-expressão. Conexões funcionais entre genes, de uma mesma via ou via que interagem, geralmente são refletidas em perfis de co-expressão.

Para identificar genes com padrões de co-expressão e gerar as redes de interação utilizou-se a ferramenta CEMiTool v. 3.17 (RUSSO et al., 2018). Os arquivos e parâmetros fornecidos para o pacote CEMiTool v. 3.17 são a matriz de expressão normalizada dos genes com o metadata de seus respectivos níveis de expressão para cada uma das condições amostrais, arquivo de anotação das condições amostrais onde há uma coluna denominada "SampleNames" que é a coluna que possui o nome das amostras e outra coluna denominada "Class" que é a coluna onde há a possibilidade de agrupar as amostras de desejo em grupos maiores e específicos, a exemplo deste estudo onde há triplicatas de cada condição amostral como “controle_1”, “controle_2” e “controle_3” agrupando-os assim em uma classe de Controle; `cor_method = c("pearson")` com a utilização da correlação de Pearson para os cálculos realizados, um `filter_pval < 0.05` (default sugerido pelos autores) para a filtragem dos genes que foram utilizados na análise e `ora_pval < 0.05` para um valor-P para análise de super-representação. Além disso o pacote seleciona um valor adequado para o parâmetro do pacote WGCNA (Weighted Wene Co-expression Network Analysis) a fim de que a análise tenha uma aderência ao modelo scale-free de Barabási-Albert.

Ainda, na análise módulos foram gerados com o destacamento dos principais genes que possuem maior conectividade durante a co-expressão, denominados hubs. O CEMiTool v. 3.17 gerou gráficos e tabelas como resultado da análise de co-expressão, tais como: Perfis de co-expressão de cada módulo, R^2 versus β e o mean connectivity versus β . Onde tal parâmetro β é um valor definido pelo CEMiTool v. 3.17, considerando os dados fornecidos, e que reflete, juntamente com outros parâmetros estatísticos mencionados anteriormente, a adesão dos mesmos ao modelo scale-free de Barabási-Albert.

Na Figura 7 pode-se observar um fluxograma, mostrando de forma ilustrativa, com os passos envolvidos na seleção do conjunto de potenciais transcritos correspondentes a genes codificantes de proteínas e, ainda, candidatos a lncRNAs linhagem-específicos alterados na regeneração da cauda do *Protopterus annectens*.

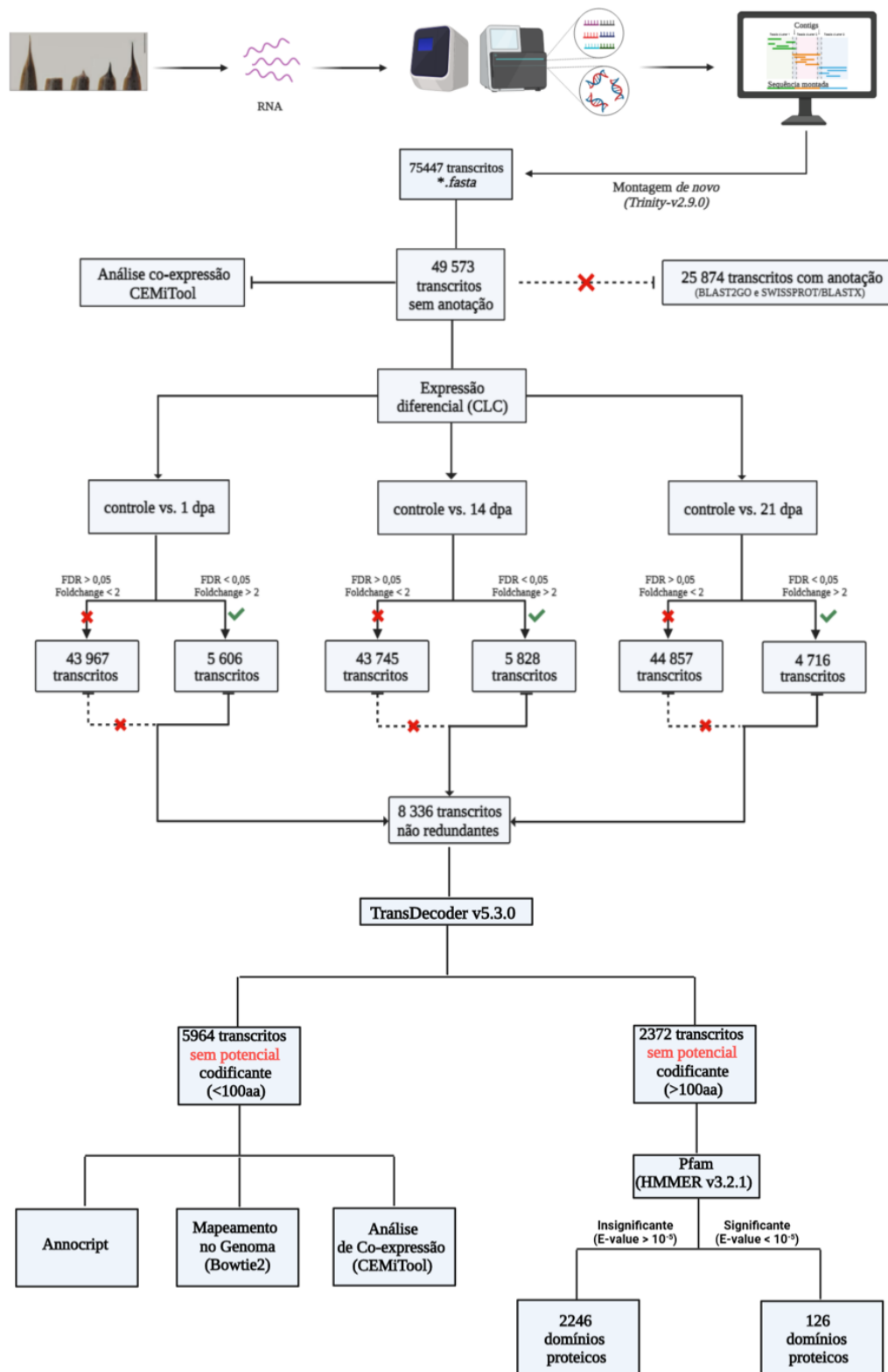


Figura 7 – Fluxograma da metodologia utilizada para seleção de transcritos correspondentes a genes codificadores linhagem específicos e genes de RNAs não-codificantes a partir de dados de RNA-seq da cauda em regeneração do *P. annectens*.

4. RESULTADOS

4.1. Análise de expressão gênica diferencial na cauda em regeneração de *Protopterus annectens*

Com base nos dados de transcriptoma produzidos em um estudo anterior (VERISSIMO et al, 2021), foram identificados transcritos diferencialmente expressos nos estágios pós-amputação da regeneração da cauda do *Protopterus annectens* (1 dpa, 14 dpa e 21 dpa) em comparação com a cauda intacta (controle). Essa análise pode ser visualizada por meio da construção de Volcano Plots, conforme mostrado na Figura 8.

Dos 75.447 contigs totais, após filtragem com base nos parâmetros utilizados nas etapas iniciais de nosso pipeline ($FDR < 0.05$ e $FC > 2$ e sem anotação após Blastx contra bancos de proteínas), foram identificados 8.336 transcritos superexpressos nos estágios de regeneração e que não apresentaram ortologia aparente com genes presentes em bancos de dados de humanos ou de outros organismos.

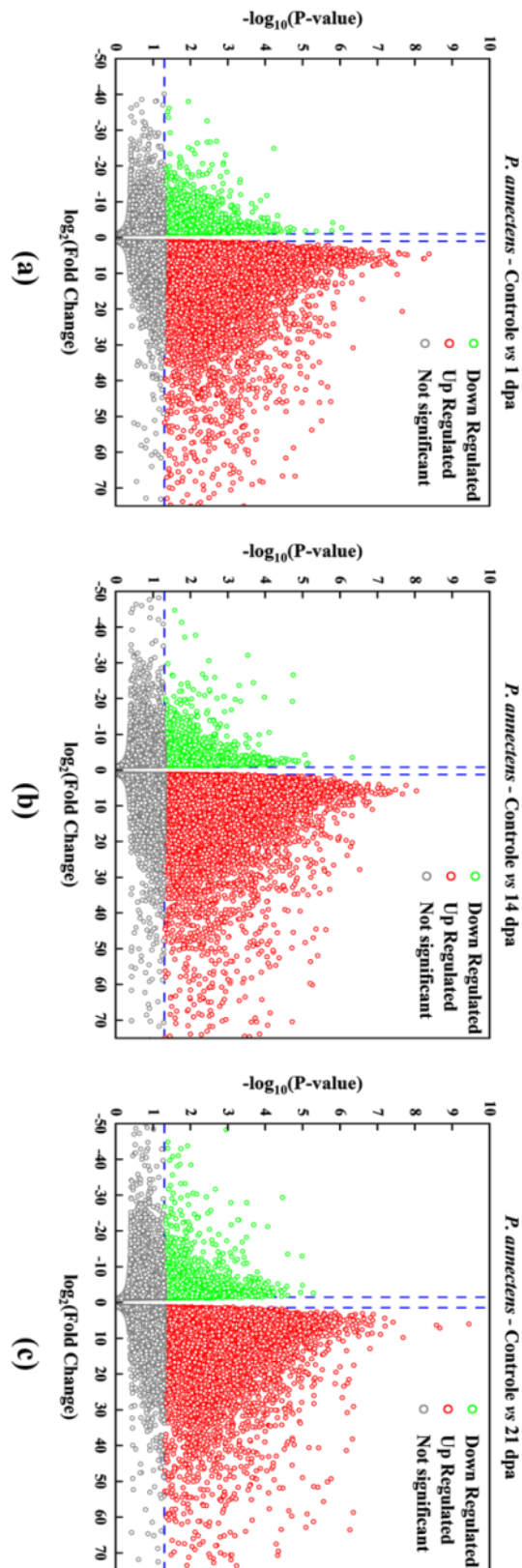


Figura 8 – Volcano Plots ilustrando a expressão diferencial dos 75.447 contigs derivados dos dados de RNA-seq durante os estágios pós-amputação de cauda do *Protopterus annectens*. (a) 1 dpa, (b) 14 dpa e (c) 21 dpa. Os pontos em verde, vermelho e cinza correspondem aos contigs com regulação negativa (Down Regulated), regulação positiva (Up Regulated) e ausência de significância estatística (Not significant), respectivamente.

4.2. Genes linhagem específicos com potencial codificante diferencialmente expressos durante a regeneração

Posteriormente, procedeu-se à análise dos 8336 contigs sem anotação e apresentadno superexpressão em um ou mais estágios de regeneração da cauda do *P. annectens* por meio da utilização do TransDecoder v. 5.3.0, com o objetivo de avaliar o potencial codificante de cada um deles. Nesse contexto, foi constatado que um total de 2372 transcritos superexpressos apresentaram evidências de potencial codificante, com fase de leitura aberta ou open reading frame (ORFs, sigla em inglês) com extensão superior a 100 aminoácidos (aa). Por outro lado, foram identificados 5964 transcritos que não revelaram indícios de potencial codificante, apresentando ORFs inferiores a 100 aa ou ausência das mesmas.

Em seguida, os 2372 transcritos diferencialmente expressos que demonstraram potencial codificante, com ORFs superiores a 100 aminoácidos (aa), foram submetidos a uma análise no banco de dados Pfam (HMMER v. 3.2.1), com o objetivo de identificar domínios proteicos. Entre esses transcritos, 126 apresentaram "hits" significativos para domínios proteicos, com um valor médio de E-value de 6.6×10^{-6} durante a busca no banco de dados. A Figura 9 mostra Volcano Plots com o perfil de expressão de todos os transcritos sem anotação nas comparações da cauda intacta com cada um dos estágios de regeneração analisados. Nesses gráficos, estão em destaque parte dos 126 genes linhagem-específicos codificadores de proteínas com as maiores variações de expressão (Fold Change > 5) ao longo das etapas do processo de regeneração. É possível portanto observar que vários genes desse grupo apresentaram grande alteração de expressão durante a regeneração.

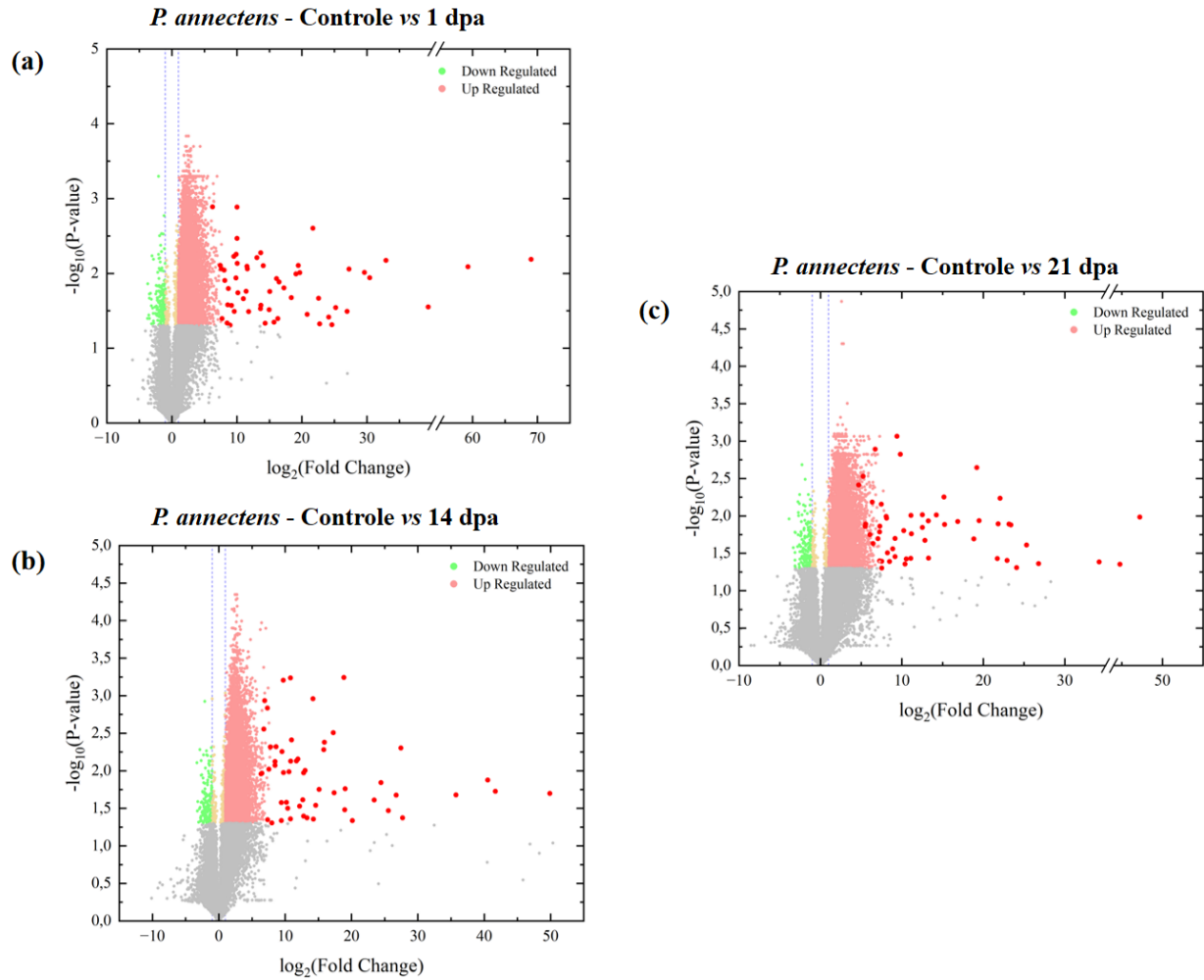


Figura 9 – Volcano Plots ilustrando a expressão diferencial de alguns genes linhagem-específicos codificadores de proteína sem anotação derivados dos dados de RNA-seq e superexpressos durante os estágios pós-amputação do *Protopterus annectens*. (a) 1 dpa, (b) 14 dpa e (c) 21 dpa. Os tais genes linhagem-específicos estão denotados em vermelho escuro no presente gráfico. Os demais pontos em verde claro, vermelho claro e cinza claro correspondem aos contigs com regulação negativa (Down Regulated), regulação positiva (Up Regulated) e ausência de significância estatística (Not significant), respectivamente.

Com relação aos tipos de domínios protéicos identificados nos 126 transcritos, 104 deles foram associados a domínios proteicos de Endonuclease-reverse transcriptase, Reverse transcriptase, GIY-YIG catalytic domain e zinc-binding reverse transcriptase. Além disso, foram encontrados 7 transcritos associados ao domínio Lamina-associated polypeptide 2 alpha, 7 transcritos ao domínio Phage integrase, N-terminal SAM-like domain 1 e 4, e 6 transcritos ao domínio Myb/SANT-like DNA-binding domain. Por fim, 2 transcritos foram associados ao domínio Retroviral aspartyl protease. A distribuição dos domínios encontrados pode ser observada na Figura 10.

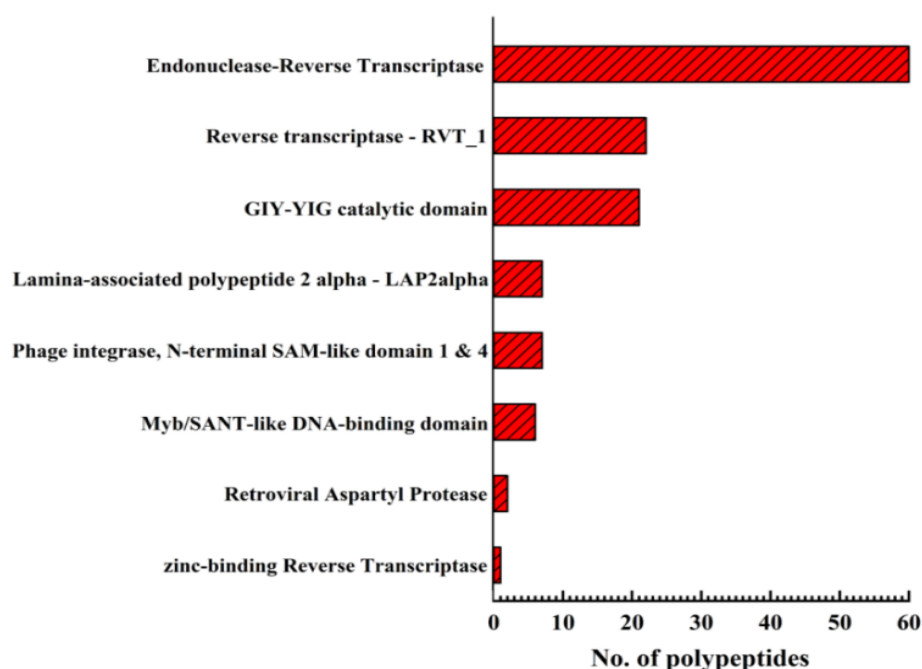


Figura 10 – Distribuição de domínios proteicos conhecidos no grupo de 126 transcritos sem ortologia aparente, mas com potencial codificante.

4.3. Potenciais lncRNAs diferencialmente expressos durante a regeneração de cauda

Foram identificados como possíveis candidatos a lncRNAs um total de 46 transcritos, dentre os 5964 sem potencial codificante apontados pela análise do TransDecoder e que foram submetidos à análise no Annocript v. 2.0.1. Esses 46 transcritos apresentaram valores de potencial codificante significativamente inferiores ao cutoff de 5% estabelecido como default do Annocript v. 2.0.1, Figura 11, ou mesmo como ausência de ORF. É importante ressaltar que os 5964 transcritos analisados nesta etapa tem um tamanho máximo de 6993 nucleotídeos, mínimo de 201 nucleotídeos e uma média de 587 nucleotídeos e os 46 transcritos candidatos têm um tamanho máximo de 479 nucleotídeos, mínimo de 201 nucleotídeos e uma média de 273 nucleotídeos, respectivamente.

A média de probabilidade codificante, dos 46 transcritos, foi de 0,596% o que é o mesmo afirmar que corresponde a uma média de probabilidade não codificante de 99,40%. Esses valores estão consideravelmente abaixo do cutoff estabelecido pelo Annocript, que é inferior a 5% para o potencial codificante e/ou superior a 95% para o potencial não codificante, Figura 12.

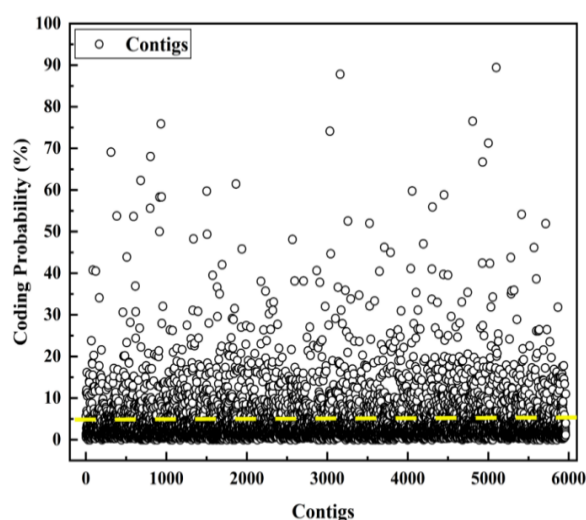


Figura 11 – Distribuição de probabilidade do potencial codificante dos 5964 contigs analisados pelo Annocript v. 2.0.1, com destaque para o valor de *cutoff* (linha amarela) para adesão aos critérios de candidatura a lncRNA, estabelecido em menos de 5% de probabilidade de potencial codificante.

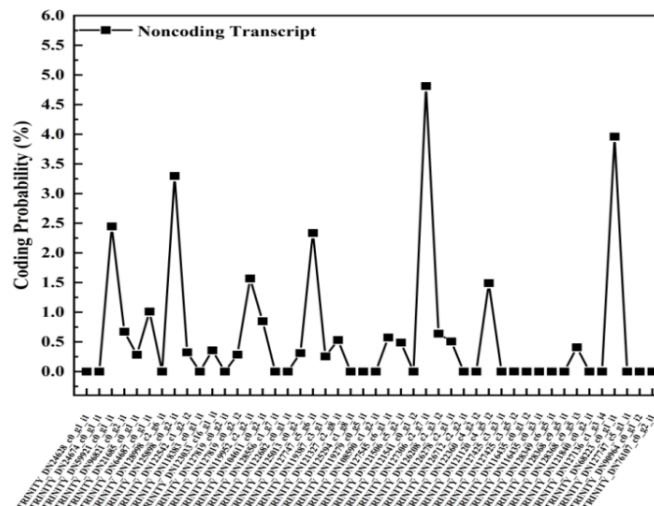


Figura 12 – Distribuição de probabilidade dos 46 transcritos identificados como possíveis candidatos a lncRNAs após a análise pelo Annocript v. 2.0.1.

Adicionalmente, a Figura 13 mostra Volcano Plots dos transcritos sem anotação colocando em destaque agora alguns destes 46 candidatos a lncRNAs com as maiores variações de variações de expressão (Fold Change > 5) ao longo do processo de regeneração.

Por último, como os lncRNAs tendem a apresentar estruturas secundárias complexas, os 46 transcritos apontados como potenciais lncRNAs foram analisados pelo programa RNAfold web server para predição de suas estruturas secundárias. Dos 46 submetidos para a análise 4 transcritos, TRINITY_DN104611_c0_g2_i1, TRINITY_DN123033_c16_g1_i1, TRINITY_DN128349_c6_g5_i1, TRINITY_DN128368_c9_g5_i1, apresentaram alta similaridade entre as predições para as estruturas secundárias de menor energia (MFE) e a estrutura secundária centroide (CS) oriunda da população de estruturas termodinamicamente geradas através da distribuição de Poisson–Boltzmann. A Figura 12, abaixo, mostra uma representação dessas estruturas com as respectivas energias livres de Gibbs associadas a MFE e CS, com o seguintes valores de variação de energia livre (ΔG), respectivamente:

- 74.50 kcal/mol e – 74.50 kcal/mol; – 75.10 kcal/mol e – 68.30 kcal/mol; – 76.70 kcal/mol e – 76.70 kcal/mol; – 55.30 kcal/mol e – 55.30 kcal/mol.

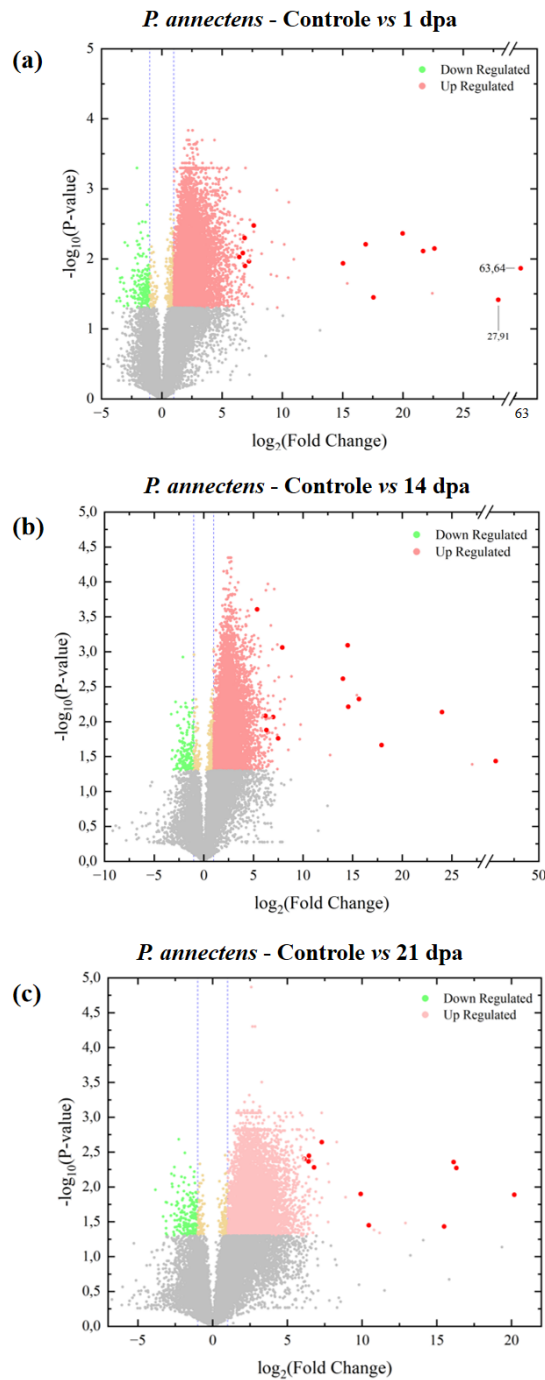


Figura 13 – Volcano Plots ilustrando a expressão diferencial de alguns genes linhagem-específicos correspondentes aos candidatos a lncRNAs derivados dos dados de RNA-seq e que estão superexpressos durante os estágios pós-amputação do *Protopterus annectens*. (a) 1 dpa, (b) 14 dpa e (c) 21 dpa. Tais candidatos a lncRNAs estão denotados em vermelho escuro no presente gráfico. Os demais pontos em verde claro, vermelho claro e cinza claro correspondem aos contigs com regulação negativa (Down Regulated), regulação positiva (Up Regulated) e ausência de significância estatística (Not significant), respectivamente.

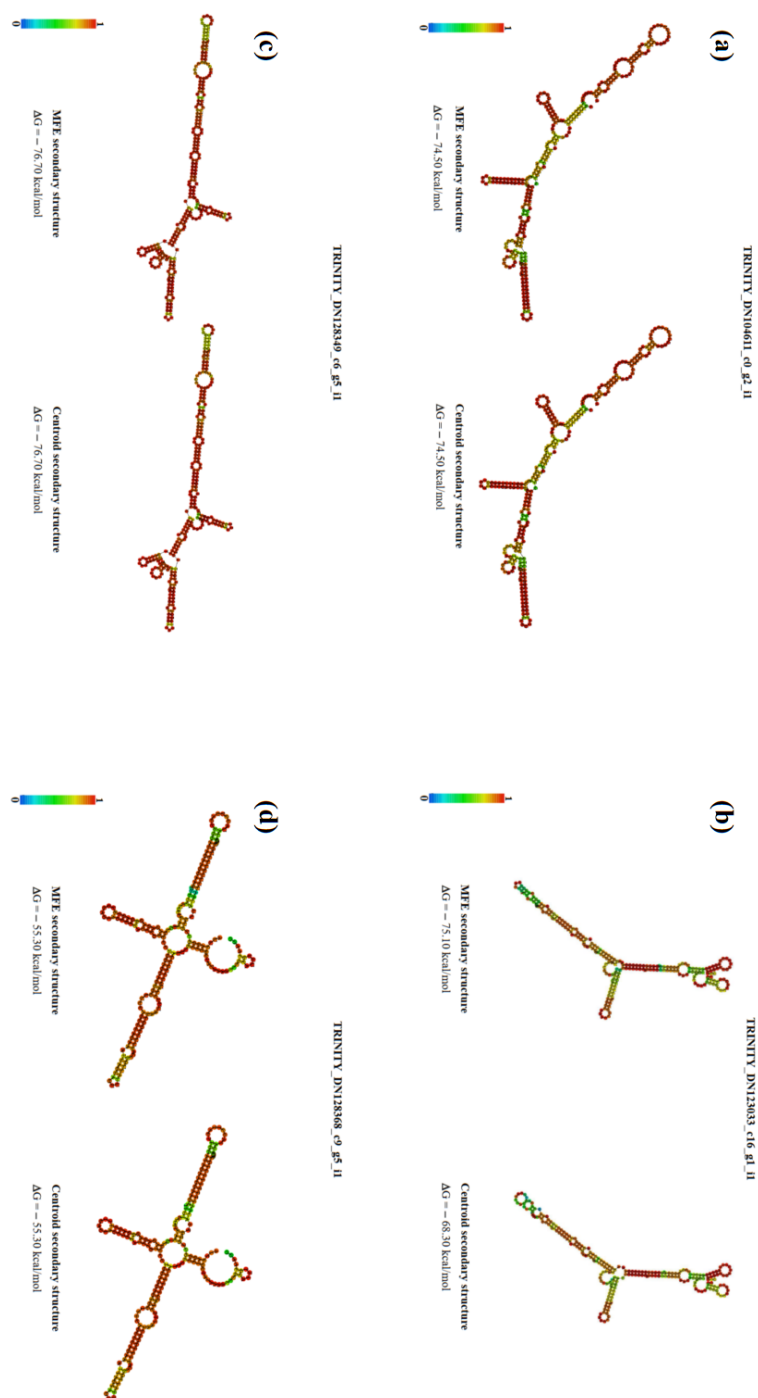


Figura 14 – Predição da estruturas secundária pelo programa RNAfold (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>) dos quatro candidatos a lncRNA diferencialmente expresso durante a regeneração de cauda em *Protopterus annectens*. (a) TRINITY_DN104611_c0_g2_i1 - MFE/CS: - 74.50 kcal/mol e - 74.50 kcal/mol. (b) TRINITY_DN123033_c16_g1_i1 - MFE/CS: - 75.10 kcal/mol e - 68.30 kcal/mol. (c) TRINITY_DN128349_c6_g5_i1 - MFE/CS: - 76.70 kcal/mol e - 76.70 kcal/mol. (d) TRINITY_DN128368_c9_g5_i1 - MFE/CS: - 55.30 kcal/mol e - 55.30 kcal/mol.

4.4. Redes de interação de co-expressão entre os genes diferencialmente expressos na cauda em regeneração

Foram analisados no CEMiTool v. 3.17 um total de 75447 transcritos provenientes das tabelas de expressão diferencial do *Protopterus annectens*. Após a filtragem de acordo com os parâmetros do CEMiTool v. 3.17, 6380 transcritos foram selecionados com base em sua maior probabilidade de estabelecer interações durante a co-expressão, utilizando um valor de p igual a 0,05.

A análise de co-expressão desses 6380 transcritos resultou na geração de gráficos estatísticos onde avaliou-se a adesão dos dados de co-expressão ao modelo de escala livre de Barabási (BARABÁSI; OLTVAI, 2004). A Figura 15 mostra a relação entre R^2 (correlação linear) versus β (parâmetro β) para o gráfico de independência de escala, enquanto a Figura 16 apresenta a conectividade média em relação a β (Mean connectivity versus β).

O CEMiTool v. 3.17 selecionou um valor de β igual a 10 como apropriado para a análise dos dados de RNA-seq fornecidos. Esse valor foi escolhido de forma a maximizar a conectividade média entre os transcritos analisados, mantendo uma relação com o valor de correlação linear (R^2) dos dados, que foi de 0,63.

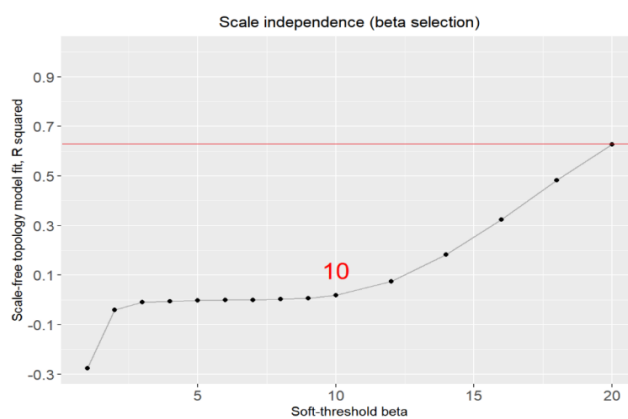


Figura 15 – Gráfico estatístico de independência de escala (R^2 versus β).

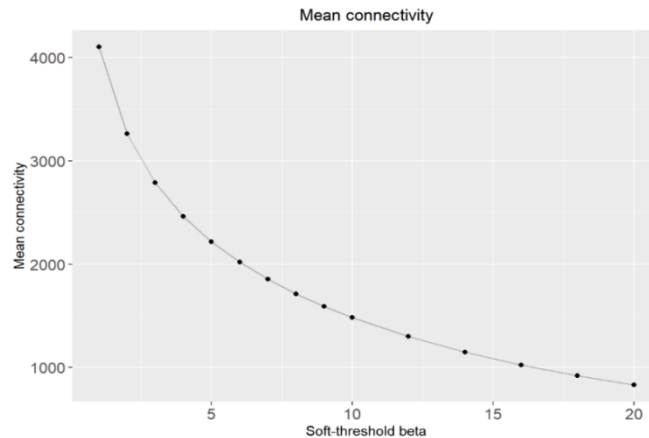


Figura 16 – Gráfico estatístico da conectividade média (Mean connectivity *versus* β).

A análise de co-expressão do *Protopterus annectens* resultou na identificação de 8 módulos de perfis de co-expressão, nos quais cada módulo é composto por um conjunto de transcritos/genes com perfil de expressão similar, conforme ilustrado na Figura 17. A distribuição dos transcritos/genes em cada módulo é apresentada em termos percentuais e numéricos, conforme segue:

M1: 80.5% – 5136; M2: 4.22% – 269; M3: 3.84% – 245; M4: 2.91% – 186; M5: 2.43% – 155; M6: 2.26% – 144; M7: 1.77% – 113; M8: 1.54% – 98.

Como mencionado, 80.5% dos 6380 considerados estatisticamente relevantes, pelo CEMiTool v. 3.17, para a análise de co-expressão estavam no módulo M1 evidenciando, assim, uma importância acentuada durante esse processo regenerativo. Na Figura 17, podemos observar que o módulo M1 corresponde a transcritos/genes superexpressos durante os três estágios de regeneração. Ainda, os genes hubs, aqueles poucos genes os estabeleceram mais conexões com os demais, para o módulo 1 (M1) foram transcritos linhagem-específicos: TRINITY_DN127308_c5_g2_i1, TRINITY_DN126408_c3_g9_i1, TRINITY_DN123033_c17_g11_i1, TRINITY_DN129867_c113_g2_i1, TRINITY_DN122010_c4_g5_i2.

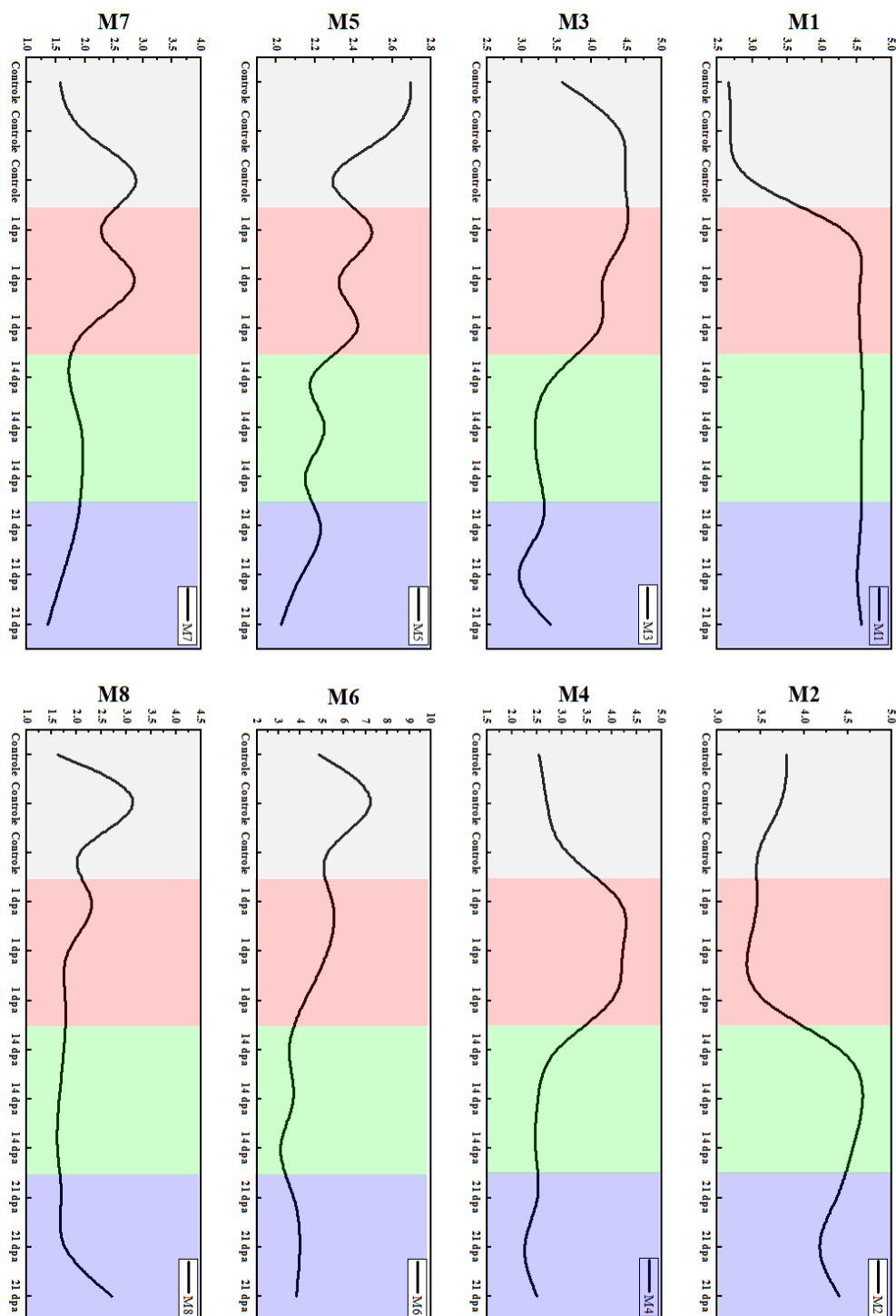


Figura 17 – Perfis de Co-expressão para os 8 módulos (M1 à M8) gerados sob as condições amostrais em triplicatas. O eixo y (ordenada) dos gráficos representa valores médios relativos de expressão gênica e o eixo x (abscissas) assinalam as triplicatas para cada uma das condições experimentais. As colorações cinza, vermelho, verde e azul nos gráficos correspondem representativamente as regiões para as triplicatas das amostras de Controle, 1 dpa, 14 dpa e 21 dpa, respectivamente).

5. DISCUSSÃO

5.1. Genes linhagem específicos com potencial codificante apresentam expressão alterada durante a regeneração de cauda

A partir da análise de expressão diferencial dos dados obtidos por RNA-seq, foi identificado um número de contigs diferencialmente expressos positivamente com significância estatística ($FC > 2$ e $FDR < 0,05$) durante o processo de regeneração nos três estágios pós-amputação (1 dpa, 14 dpa e 21 dpa), em comparação com a cauda não lesionada (controle), conforme mostrado na Figura 8. O número de contigs sem anotação com expressão positiva durante os três estágios de regeneração (1 dpa, 14 dpa e 21 dpa) foi de 5606, 5828 e 4716, respectivamente.

Nossas análises determinaram que 2372 transcritos sem ortologia aparente e com potencial codificante são superexpressos durante a regeneração de cauda e portanto, representam potenciais genes linhagem específicos com papel durante a regeneração. A análise do Pfam revelou um enriquecimento de domínios específicos nos transcritos/polipeptídeos associados aos estágios pós-amputação do *Protopterus annectens*. Observou-se o enriquecimento dos domínios Endonuclease-reverse transcriptase, Reverse transcriptase (RNA-dependent DNA polymerase - RVT_1), GIY-YIG catalytic domain e zinc-binding reverse transcriptase (zf-RVT), totalizando 104 associações, Figura 9. Esses domínios desempenham um papel essencial no processo de clivagem do DNA, bem como em processos transcricionais e de expansão do genoma, em conjunto com componentes genéticos como transposons e retrotransposons. É relevante destacar que um grande número de transcritos relacionados a elementos transponíveis conservados foi detectado como superexpressos na nadadeira caudal em regeneração do *Protopterus annectens*, conforme relatado anteriormente pelo nosso grupo (VERÍSSIMO et al., 2020). É interessante notar, que um considerável número de genes associados a elementos transponíveis linhagem-específicos também possam atuar na regeneração de cauda em *Protopterus annectens*.

Um estudo recente de Wang et al. (2021), que aborda a montagem do genoma do *P. annectens*, identificou a presença de uma quantidade significativa de elementos transponíveis ativos, como elementos nucleares intercalados longos, repetições terminais longas e transposons de DNA. Nossos resultados revelaram um número substancial de domínios

RVT_1 que codificam Transcriptase reversa (RNA-dependent DNA polymerase), corroborando o trabalho de Wang et al. (2021), no qual esse domínio é transportado por retrotransposons amplamente distribuídos pelo genoma do *P. annectens*. É interessante notar que esses retrotransposons apresentam uma correlação linear com o comprimento dos íntrons dos genes do peixe pulmonado. Nossos dados sugerem que, além dos elementos transponíveis conservados, o genoma do *P. annectens* contém transposons ativos específicos desse táxon, os quais podem desempenhar um papel na regeneração da cauda. Vale salientar que os elementos transponíveis têm sido propostos como importantes fatores na geração de diversidade genômica e inovações linhagem-específicas em vertebrados (WARREN et al., 2015).

Em relação ao domínio Lamina-associated polypeptide 2 alpha - LAP2alpha, foram identificados 7 transcritos/poli-peptídios associados a tal domínio durante a busca. Esse domínio está associado às proteínas Lamina A e Lamina C, que são os principais produtos de splicing do gene LMNA e são encontradas no nucleoplasma. Essas proteínas estão envolvidas na estruturação da cromatina, incluindo a LAP2 alfa (VIDAK et al., 2018) e o fator de barreira para auto-integração (BAF) (LOI et al., 2016). A lamina A é transcrita e traduzida a partir de uma proteína precursora conhecida como prelamina A, que sofre processamento pós-traducional com clivagem pela metalopeptidase de zinco STE24 (ZMPSTE24), resultando na formação de um peptídeo curto e da lamina A madura (CAMOZZI et al., 2014). As laminas têm sido descritas como estruturas que desempenham um papel na sinalização celular em modelos de diferenciação muscular e laminopatias musculares (MARALDI et al., 2011).

Estudos têm mostrado a relação entre mTOR e lamina A, destacando a regulação do complexo mTORC1, hiperativação do mTORC2 em tecido adiposo laminopático, hiperativação do mTOR dependente de lamina A mutada em mioblastos, e a redução dos níveis de prelamina A por meio do tratamento com Rapamicina, entre outros (CHIARINI et al., 2019). Portanto, há fortes evidências de que os mecanismos de resposta para a manutenção adequada da lamina A (BERTACCHINI et al., 2013) possam envolver a ativação do mTOR tanto em condições normais quanto patológicas ou de estresse. Esses resultados fornecem evidências adicionais do papel do complexo mTORC1 no processo de regeneração de membros no presente estudo.

O domínio Phage integrase, N-terminal SAM-like domain 1 & 4 foi detectado em 7 transcritos/poli-peptídios em nossa análise. Esse domínio está relacionado a enzimas que medeiam a recombinação unidirecional específica entre duas sequências de DNA. O domínio

Myb/SANT, por sua vez, apresentou 6 transcritos/polipeptídeos associados ao final da busca no Pfam. Essas proteínas estabelecem interações com o DNA e outras proteínas, auxiliando em processos transcricionais, como o splicing do mRNA. Por fim, o domínio Retroviral aspartyl protease (RVP) foi identificado em 2 transcritos/polipeptídios associados e está relacionado às enzimas responsáveis pela clivagem de peptídeos. O Apêndice B traz informações dos resultados de expressão diferencial bem como dos domínios proteicos identificados no Pfam para os 126 transcritos linhagem-específicos com a respectiva significância estatística em termos de E-value.

5.2. Identificação de potenciais transcritos candidatos à lncRNAs com expressão alterada durante a regeneração de cauda

No genoma de *P. annectens* (<https://www.ncbi.nlm.nih.gov/data-hub/taxonomy/7888/>) (WANG et al., 2021), foi identificado que 7,7% (2350) do seu genoma consiste em ncRNAs, tanto anotados quanto não anotados. Nós identificamos 5964 transcritos/contigs sem anotação e que não apresentaram potencial codificante, após a análise do TransDecoder. Após análise com o pacote Annocript, identificamos apenas 46 com alto potencial de representarem lncRNA. O principal módulo de análise utilizado, o CPC2 do Annocript, é amplamente reconhecido e recomendado para análise em organismos não-modelo. Sua alta especificidade de 98,65% e sensibilidade de 57,09% foram atestadas em estudos anteriores com peixes, como a identificação de novos lncRNAs em zebrafish (MISHRA et al., 2021; LI et al., 2020; SALABI et al., 2021). Assim, o método de análise do Annocript é bem restritivo e por apresentar alta especificidade e menor sensibilidade, é possível que o número de lncRNAs dentro do grupo que selecionamos seja maior.

É válido observar algumas características avaliadas pelo CPC2 durante a análise do Annocript que em conjunto são decisivas para a sugestão de um candidato a lncRNA, são elas: O comprimento da ORF é uma medida importante para distinguir transcritos com potencial codificante. A integridade da ORF que é crucial para diferenciar entre RNAs codificantes e não codificantes, verificando se as ORFs estão completas ou incompletas com base nos start e stop códons bem como com a ausência da própria ORF nas sequências fornecidas. Além disso, outro fator importante são os valores de ponto isoelétrico (pI). Os valores de pI foram obtidos traduzindo a ORF mais longa em sequências de aminoácidos e calculando suas propriedades físico-químicas (BJELLQVIST et al., 1994). Assim, fez-se a

presunção de que peptídeos identificados artificialmente em transcritos não codificantes devem ter propriedades físico-químicas diferentes quando comparados com os peptídeos reais oriundos de transcritos codificantes.

Finalmente, um total de 46 transcritos atenderam aos critérios estabelecidos pelo Annocript para serem considerados candidatos a lncRNAs. Esses transcritos apresentaram um tamanho máximo de 479, mínimo de 201 e uma média de 273 nucleotídeos, respectivamente. Além disso, os mesmos não apresentaram ORFs em sua estrutura, indicando a ausência de potencial codificante, e possuem um valor de pI igual a 0. Essas duas últimas características confirmam, como mencionado anteriormente, a inexistência de micropeptídeos codificados nesses transcritos, pois um valor de pI igual a 0 indica a ausência de assinaturas isoeletricas características de estruturas moleculares dos aminoácidos. Vale ressaltar que, de acordo com a metodologia adotada neste estudo, foram selecionados apenas os candidatos a lncRNAs que não apresentavam a possibilidade de traduzir micropeptídeos em sua estrutura de acordo com os critérios de ORFs e pI mencionados. Assim, esses 46 transcritos são considerados como os possíveis candidatos a lncRNAs.

Em dois estudos recentes Xing et al., 2021 bem como Pan et al., 2022 abordam o importante papel desempenhado por micropeptídeos codificados por pequenas ORFs em lncRNAs na patogênese em muitas doenças bem como o seu papel em vias regulatórias que influenciam na função muscular, estabilidade de RNAs, expressão gênica etc (XING et al., 2021; PAN et al., 2022).

Interessante destacar, ainda, o papel dos lncRNAs em contextos fisiopatológicos como as disordens neurais atuando em atividades regulatórias em diversos cenários como diferenciação celular, crescimento celular e a diversas respostas ao estresse, dentre outros. Dentre esses cenários podemos destacar o processo de regeneração neural mediado por um lncRNA *Silc1* juntamente com um fator de transcrição SOX11 que são altamente expressos após dano neural em camundongos (PERRY et al., 2018; HEZRONI et al., 2019).

Sequencialmente, os resultados da análise utilizando a ferramenta RNAfold permitiram uma avaliação adicional dos 46 candidatos a lncRNA, a fim de determinar quais têm potencial viabilidade termodinâmica para a formação de estrutura secundária. Essa predição é baseada em informações sobre a termodinâmica estatística, distribuição de Boltzmann e energia de perturbação das moléculas de nucleotídeos em uma determinada sequência. Portanto, os lncRNAs com menores variações de energia livre de Gibbs (ΔG) e maior concordância estrutural entre os modelos de estrutura secundária MEF e estrutura

secundária centróide (CS) indica uma maior probabilidade de que a predição estrutural dos lncRNAs seja viável. A estrutura secundária centróide como o próprio nome diz é uma estrutura que representa a tendência central ou média das estruturas que foram geradas pelo RNAfold, considerando as diversas perturbações energéticas bem como os aspectos termodinâmicos das mesmas. Ainda, ressaltando que as estruturas MEF são correspondentes a aquelas estruturas de menor energia livre de Gibbs (ΔG). Finalmente, verificou-se os 4 genes linhagem-específicos oriundos dos 46 candidatos a lncRNAs que possuem possíveis estruturas secundárias definidas teoricamente uma vez que tanto MFE quanto CS convergiram, como mostrado na Figura 12. Os Apêndices A e C trazem informações no que tange às possíveis estruturas secundárias para os 46 candidatos a lncRNA quanto informações relativas a expressão diferencial dos mesmos e as probabilidades de codificação de proteínas obtidas pelo Annocript.

5.3. Módulos de co-expressão ativos durante a regeneração de cauda

O procedimento foi realizado utilizando o software CEMiTool para analisar os perfis de co-expressão dos transcritos em relação às diferentes condições experimentais às quais o *P. annectens* foi submetido: controle, 1 dpa, 14 dpa e 21 dpa. Utilizando os respectivos valores de expressão gênica para cada uma das triplicatas das condições. O CEMiTool aplicou um algoritmo de filtragem para selecionar os genes a serem analisados, levando em consideração os valores de expressão e um valor de p igual a 0,05 (filter_pval = 0,05). Isso permitiu utilizar apenas os genes com maior significância estatística na análise de co-expressão.

O modelo de escala livre ou scale-free de Barabási (BARABÁSI; OLTVAI, 2004) é um algoritmo que gera redes sem escala de forma aleatória. Esse modelo sugere que existem poucos genes com muitas conexões e muitos genes com poucas conexões. Portanto, é importante considerar o parâmetro β para avaliar a co-expressão de acordo com o modelo de Barabási. No caso desta análise, o valor selecionado para β foi 10.

Ao estabelecer essa relação com o valor de correlação linear R^2 , foi encontrado um valor de aproximadamente 0,63, que difere do valor padrão de 0,8. No entanto, essa diferença não indica resultados incorretos. O CEMiTool procurou valores de β menores, no caso 10, para que a co-expressão se ajustasse ao modelo scale-free, conforme ilustrado na Figura 13. Assim, os módulos foram capazes de evidenciar os genes com o maior estabelecimento de conexões, conforme visto na Figura 14.

Os perfis de expressão permitiram avaliar a co-expressão dos genes agrupados em cada módulo ao longo das condições de controle, 1 dpa, 14 dpa e 21 dpa. Na Figura 15, o Módulo 1 (M1) é destacado, mostrando um recrutamento aumentado de genes que estão co-expressos em todas os estágios temporais após a amputação em comparação com a condição de controle. Isso sugere que, em situações de estresse, como a amputação, os genes pertencentes a esse módulo tendem a ser recrutados e estabelecendo conexões, evidenciando sua relevância biológica no processo de regeneração.

Cinco hub genes do módulo M1, módulo este que teve 80.5% dos transcritos apresentando co-expressão positiva, são genes sem anotação que apresentaram expressão diferencial de acordo com análises anteriores. Esses genes hubs estão associados a vários processos biológicos, como organização do citoesqueleto, migração celular, diferenciação celular, regulação positiva em resposta ao estresse oxidativo, proliferação muscular de músculos esqueléticos, regulação da sinalização TOR (mTORC1 e mTORC2), desenvolvimento de estruturas anatômicas, regulação da proliferação de fibroblastos, proliferação celular, organização das fibras musculares através da actomiosina e processos de fosforilação, entre outros. Isso foi possível ser verificado ao realizar uma busca em bancos de dados com as sequências dos mesmos. Os genes hubs são elementos que estabelecem mais conexões com outros genes no respectivo módulo. Esses genes desempenham um papel essencial nos sistemas biológicos (BARABÁSI; OLTVAI, 2004).

Além disso, foram identificados 45 transcritos linhagem-específicos relacionados aos domínios proteicos mencionados anteriormente que apresentaram co-expressão aumentada neste mesmo módulo M1 e, portanto, apresentando conexões com os genes hubs, sendo eles: 20 transcritos Endonuclease-reverse transcriptase, 10 transcritos Reverse transcriptase (RNA-dependent DNA polymerase - RVT_1), 6 transcritos Lamina-associated polypeptide 2 alpha - LAP2alpha, 5 transcritos Phage integrase e 4 transcritos Myb/SANT. Adicionalmente, 31 transcritos candidatos a lncRNA também mostraram co-expressão aumentada e conexões com os hubs no Módulo 1 (M1). Portanto, pode-se estabelecer uma relação entre esses genes linhagem-específicos, codificantes e não codificantes, e os processos biológicos representados pelos hubs, com co-expressão constante durante os 3 estágios temporais da regeneração (1 dpa, 14 dpa e 21 dpa) apresentando um papel fundamental para o processo regenerativo do *P. annectens*. Os Apêndices B e C trazem informações dos 126 genes linhagem específicos codificadores de proteínas bem como os 46 candidatos a lncRNA no que tange aos módulos que cada um exibiu co-expressão ou não.

6. CONCLUSÕES

O trabalho realizado inclui a construção de uma pipeline para análise *in silico* de dados de RNA-seq que permitiu identificar potenciais genes linhagem específicos (codificantes e não codificantes) com expressão alterada na cauda em regeneração do peixe pulmonado *Protopterus annectens*.

Foram detectados 8336 transcritos sem aparente ortologia com genes de outras espécies e que estão alterados durante a regeneração de cauda de *P. annectens*. Desses, 2372 transcritos apresentam potencial codificante (ORF > 100 aa), com 126 deles apresentando domínios proteicos conhecidos. Entre os domínios proteicos identificados, a grande maioria corresponde a domínios presentes em proteínas derivadas de retrotransposons. Isso indica a participação de elementos transponíveis linhagem específicos no processo de regeneração.

Uma análise bem restrigente para identificação de RNAs não codificantes revelou que dos 5964 transcritos sem a presença de uma ORF maior de 100 aa, 46 transcritos têm alto potencial de representarem lncRNAs linhagem-específicos com expressão alterada na cauda em regeneração e 4 transcritos possuem estruturas secundárias teóricas viáveis.

Foram identificados oito módulos de transcritos com expressão coordenada durante a regeneração. No principal módulo, o M1, que representa genes superexpressos durante os três estágios analisados de regeneração há cinco hub genes representados por transcritos linhagem específicos, indicando tais genes podem estar envolvidos em múltiplas interações moleculares durante a regeneração.

Os resultados apresentados e devido à representatividade de aproximadamente 80.5% dos genes por parte do módulo 1 (M1), onde no mesmo há 45 genes linhagem-específicos codificantes e 31 genes linhagem-específicos não codificantes, e a co-expressão positiva em todos os estágios da regeneração, é imperativo enfatizar a necessidade de investigações posteriores mais aprofundadas em relação aos componentes genéticos deste módulo específico. A co-expressão positiva pode indicar que esses genes estão trabalhando em conjunto para desencadear respostas específicas de regeneração, talvez ativando vias de sinalização que levam à proliferação e diferenciação celular. Esses genes podem estar desempenhando papéis essenciais em tal processo, possivelmente coordenando diversas funções celulares. Além disso, a interação entre os genes linhagem-específicos e os genes já anotados no genoma é fundamental para compreender como as vias moleculares conservadas

são reguladas durante a regeneração podendo haver conexões diretas ou indiretas entre esses genes, e entender essas interações pode ajudar a identificar os reguladores-chave e os pontos de controle nos processos regenerativos de estruturas complexas em vertebrados.

7. REFERÊNCIAS

AKIMENKO, M. A.; MARÍ-BECCA, M.; BECERRA, J.; GÉRAUDIE, J. Old questions, new tools, and some answers to the mystery of fin regeneration. *Developmental Dynamics*, v. 226, p. 190–201, 2003.

ALLOUL, L.; BALZANO, S.; MAGG, A.; et al. Non-coding deletions identify Maenli lncRNA as a limb-specific En1 regulator. *Nature*, 592(7852):93-98, 2021.

AMARAL, D. B.; SCHNEIDER, I. Fins into limbs: Recent insights from sarcopterygian fish. *Genesis*, v. 56, n. 1, p. e23052, 2018.

AMEMIYA, C. T.; ALFÖLDI, J.; Lee, A.; et al. The African Coelacanth Genome Provides Insights into Tetrapod Evolution. *Nature*, v. 496, p. 311–316, 2013.

ANDERSON, D. M. et al. A micro peptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, p. 595–606, 2015.

BAO, Z.; et al. LncRNADisease 2.0: an updated database of long non-coding RNA associated diseases. *Nucleic Acids Res.*, p. D1034–D1037, 2019.

BARABÁSI, A. -L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, v. 5, n. 2, p. 101-113, 2004.

BERNSTEIN, E.; ALLIS, C. D. RNA meets chromatin. *Genes Dev.*, p. 1635–1655, 2005.

BERTACCHINI J.; BERETTI F.; CENNI V.; et al. The protein kinase Akt/PKB regulates both prelamin A degradation and Lmna gene expression. *FASEB J*, v. 27, p. 2145–2155, 2013.

BI, X.; WANG, K.; YANG, L.; et al. Tracing the genetic footprints of vertebrates landing in non-teleost ray-finned fishes. *Cell*, 184(5):1377-1391.e14, 2021.

BISCOTTI, M. A.; GERDOL, M.; CANAPA, A.; FORCONI, M.; OLMO, E.; PALLAVICINI, A.; BARUCCA, M.; SCHARTL, M. The lungfish transcriptome: A glimpse into molecular evolution event at the transition from water to land. *Scientific Reports*, v. 6, p. 21571, 2016.

BJELLQVIST, B.; BASSE, B.; OLSEN, E.; CELIS, J.E. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis*, pp. 529-539, v. 15. 1994.

- BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, v. 30, n. 15, p. 2114–2120, 2014.
- BRADBURY, E. J.; BURNSIDE, E. R. Moving beyond the glial scar for spinal cord repair. *Nature Communications*, v. 10, n. 1, p. 3879, 2019.
- BROCKES, J. P.; GATES, P. B. Mechanisms underlying vertebrate limb regeneration: Lessons from the salamander. *Biochemical Society Transactions*, v. 42, p. 625–630, 2014.
- CAM, P. H.; CHEN, E. S.; GREWAL, S. I. S. Transcriptional scaffolds for heterochromatin assembly. *Cell*, p. 610–614, 2009.
- CAMOZZI D.; CAPANNI C.; et al. Diverse lamin-dependent mechanisms interact to control chromatin dynamics. Focus on laminopathies. *Nucleus*, v. 5, p. 427–440, 2014.
- CHEN, F.; et al. Extracellular vesicle-packaged HIF-1 α -stabilizing lncRNA from tumour-associated macrophages regulates aerobic glycolysis of breast cancer cells. *Nat. Cell Biol.*, p. 498–510, 2019.
- CHEN, J.; SHISHKIN, A. A.; ZHU, X.; et al. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biology*, v. 17, p. 19, 2016.
- CHEN, S.D.; ZHANG, Y.E.; and LONG, M.Y. New Genes in *Drosophila* Quickly Become Essential. *Science*, p. 1682–1685, v. 330(6011), 2010.
- CHEN, X.; et al. Computational models for lncRNA function prediction and functional similarity calculation. *Briefings in Functional Genomics*, p. 58–82, 2018.
- CHEN, X.; et al. Long non-coding RNAs and complex diseases: from experimental results to computational models. *Briefings in Bioinformatics*, p. 558–576, 2016.
- CHIARINI F.; EVANGELISTI C.; CENNI V.; FAZIO A.; PAGANELLI F.; et al. The Cutting Edge: The Role of mTOR Signaling in Laminopathies. *Int J Mol Sci*, v. 20, p. 847, 2019.
- CINTI S. The adipose organ at a glance. *Dis. Model. Mech*, v. 5, p. 588–594, 2012.
- CLARK, M. B.; et al. Genome-wide analysis of long noncoding RNA stability. *Genome Res.*, p. 885–898, 2012.
- CONANT, E. B. Regeneration in the African Lungfish, *Protopterus* I. GROSS ASPECTS. *J. Exp. Zool.*, v. 174, p. 15–32, 1970.
- CONESA, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*, v. 17, n. 1, p. 13, 2016.
- DARNET, S.; DRAGALZEW, A. C.; AMARAL, D. B.; SOUSA, J. F.; THOMPSON, A. W.; CASS, A. N.; LORENA J.; PIRES, E. S.; COSTA, C. M.; SOUSA, M. P.; FRÖBISCH, N. B.; OLIVEIRA, G.; SCHNEIDER, P. N.; DAVIS, M. C.; BRAASCH, I.; SCHNEIDER, I. Deep evolutionary origin of limb and fin regeneration. *Proceedings of the National Academy of Sciences*, v. 116, p. 15106–15115, 2019.

ECHEVERRI, K.; TANAKA, E. M. Mechanisms of Muscle Dedifferentiation during Regeneration. *Seminars in Cell & Developmental Biology*, v. 13, p. 353–360, 2002.

ESPINOZA, C. A.; et al. B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat. Struct. Mol. Biol.*, p. 822–829, 2004.

FERNANDES, J. C. R.; et al. Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease. *Noncoding RNA*, 2019.

FRÖBISCH, N. B.; BICKELMANN, C.; OLORI, J. C.; WITZMANN, F. Deep-Time Evolution of Regeneration and Preaxial Polarity in Tetrapod Limb Development. *Nature*, v. 527, p. 231–234, 2015.

FRÖBISCH, N. B.; BICKELMANN, C.; WITZMANN, F. Early evolution of limb regeneration in tetrapods: Evidence from a 300-million-year-old amphibian. *Proceedings of the Royal Society B: Biological Sciences*, v. 281, p. 20141550, 2014.

GAI, Y. et al. A Novel lncRNA, Associated With Environmental Stress in Mulberry. *Front. Plant Sci.*, p. 669, 2018.

GILBERT, S. E. *Biologia do Desenvolvimento*. Fundação Calouste Gulbenkian, p. 363, 2008.

GONZALEZ, I.; et al. A lncRNA regulates alternative splicing via establishment of a splicing specific chromatin signature. *Nat. Struct. Mol. Biol.*, p. 370–376, 2015.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnology*, v. 29, p. 644–652, 2011.

GRUBER, A.R.; LORENZ, R.; BERNHANTERNHART, S.H.; NEUBÖCK, R.; HOFACKER, I.L. The Vienna RNA websuite. *Nucleic Acids Res.* 2008, 36, W70–W74

GUANGZHEN H., ZHENKUN L., and MAMTA G.. The long non-coding RNA GAS5 cooperates with the eukaryotic translation initiation factor 4E to regulate c-Myc translation. *PLoS One*, p. e107016, 2014.

GUPTA, R. A.; et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, p. 1071–1076, 2010.

GUTTMAN, M.; et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, p. 295–300, 2011.

HAAS, B. J.; et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocol.*, v. 8, p. 1494–1512, 2013.

HAAS, B. J.; WHITED, J. L. Advances in Decoding Axolotl Limb Regeneration. *Trends in Genetics*, v. 33, p. 553–565, 2017.

HARROW, J.; et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, p. 1760–1774, 2012.

HE, R.; LUO, D.; MO, Y. Emerging roles of lncRNAs in the posttranscriptional regulation in cancer. *Genes Dis.*, p. 6–15, 2019.

HEZRONI, H., PERRY, R. B. T., ULITSKY, I. Long noncoding RNAs in development and regeneration of the neural lineage. *Cold Spring Harb. Symp. Quant. Biol.* 84, 165–177, 2019.

HEZRONI H., KOPPSTEIN D., SCHWARTZ MG., et al. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11(7):1110-22, 2015.

HO, D. M.; WHITMAN, M. TGF-beta signaling is required for multiple processes during *Xenopus* tail regeneration. *Dev Biol.*, v. 315, p. 203-16, 2008.

HON, C.; et al. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, p. 199–204, 2017.

HU, K.; ZHANG, J.; LIANG, M. LncRNA AK015322 promotes proliferation of spermatogonial stem cell C18-4 by acting as a decoy for microRNA-19b-3p. *In Vitro Cell. Dev. Biol. Anim.*, p. 277–284, 2017.

IWAKIRI J.; TERA, G.; HAMADA, M. Computational prediction of lncRNA-mRNA interactions by integrating tissue specificity in human transcriptome. *Biol. Direct*, p. 15, 2017.

JEON, Y.; LEE, J. T. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*, p. 119–133, 2011.

JIA, H.; et al. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA*, pp. 1478–1487, 2010.

JORGENSEN, J. M.; JOSS, J. (Ed.) *The biology of lungfishes*. New Hampshire: Science Publishers, Cap. 5, p. 129-147, 2011.

KANG, Y.J.; YANG D.C.; KONG L.; HOU M.; MENG Y.Q.; WEI L.; et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, v. 45, pp. W12-W16, 2017.

KAZIMIERCZYK, M.; KASPROWICZ, M.K.; KASPRZYK, M.E.; WRZESINSKI, J. Human Long Noncoding RNA Interactome: Detection, Characterization and Function. *Int. J. Mol. Sci.* 21, 1027, 2020.

KHALIL, A. M.; et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, p. 11667–11672, 2009.

KIM, J., GUAN, KL. mTOR as a central hub of nutrient signalling and cell growth. *Nat Cell Biol* 21, 63–71, 2019.

- KNAPP, D. et al. Comparative transcriptional profiling of the axolotl limb identifies a tripartite regeneration-specific gene program. *PLoS ONE*, v. 8, p. e61352, 2013.
- KUMAR, A.; GATES, P.B.; CZARKWIANI, A.; and BROCKES, J.P. An orphan gene is necessary for preaxial digit formation during salamander limb development. *Nat Commun*, 6, 8684, 2015.
- KUNG, J. T. Y.; COLOGNORI D.; LEE, J. T. Long noncoding RNAs: past, present, and future. *Genetics*, p. 651–669, 2013.
- LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth*, v. 9, n. 4, p. 357-359, 2012.
- LAPLANTE M.; SABATINI D.M. mTOR signaling in growth control and disease. *Cell*, v. 149, p. 274–293, 2012.
- LAURENT, G. St.; WAHLESTEDT, C.; KAPRANOV, P. The Landscape of long non-coding RNA classification. *Trends Genet.*, p. 239–251, 2015.
- LEE, J. T. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat. Rev. Mol. Cell Biol.*, p. 815–826, 2011.
- LEE, J. T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev.*, p. 1831–1842, 2009.
- LI, J.; ZHANG, S.; AMAYA, E. The cellular and molecular mechanisms of tissue repair and regeneration as revealed by studies in *Xenopus*. *Regeneration (Oxf)*, v. 3, p. 198-208, 2016.
- LI, J.; ZHANG, X.; LIU, C. The computational approaches of lncRNA identification based on coding potential: Status quo and challenges. *Computational and Structural Biotechnology Journal*, v. 18, p. 3666-3677, 2020.
- LOI M.; CENNI V.; et al. Barrier-to-autointegration factor (BAF) involvement in prelamin A-related chromatin organization changes. *Oncotarget*, v. 7, p. 15662–15677, 2016.
- LOOSO, M., PREUSSNER, J., SOUSOUNIS, K., BRUCKSKOTTEN, M., MICHEL, C.S., LIGNELLI, E., et al. A de novo assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. *Genome Biol*, 14(2), R16, 2013.
- LOW, L.; TAMMI, M. *Bioinformatics A Practical Handbook of Next Generation Sequencing and Its Applications*. World Scientific Publishing Co. Pte. Ltd., 2017.
- MA, L.; et al. LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res.*, p. D128–D134, 2019.
- MARALDI N. M.; CAPANNI C.; CENNI V.; et al. Laminopathies and lamin-associated signaling pathways. *J. Cell. Biochem*, v. 112, p. 979–992, 2011.
- MARCER, T. R.; DINGER, M. E.; MATTICK, J. S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, p. 155–159, 2009.

- MARCHESE, F. P.; RAIMONDI, I.; HUARTE, M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.*, p. 206, 2017.
- MARQUES, A. C. and PONTING, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.*, p. R124, 2009.
- MATSUMOTO, A. et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*, p. 228–232, 2017.
- MATTICK, J. S. and MAKUNIN I. V. Non-coding RNA. *Hum. Mol. Genet.*, p. R17–R29, 2006.
- MATTICK, J. S. The genetic signatures of noncoding RNAs. *PLoS Genet.*, p. e1000459, 2009.
- MEYER A., SCHLOISSNIG S., FRANCHINI P., et al. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature*, v. 590, p. (7845):284-289, 2021.
- MEYER, A. & DOLVEN, S. I. Molecules, Fossils, and the Origin of Tetrapods. *Journal of Molecular Evolution*, v. 35, p. 102–113, 1992.
- MISHRA, S. K. and WANG, H. Computational Analysis Predicts Hundreds of Coding lncRNAs in Zebrafish. *Biology*, v. 10(5), p. 371, 2021.
- MONAGHAN, J. R. et al. Microarray and cDNA sequence analysis of transcription during nerve-dependent limb regeneration. *BMC Biology*, v. 7, p. 1, 2009.
- MONAGHAN, J. R.; WALKER, J. A.; PAGE, R. B.; PUTTA, S.; BEACHY, C. K.; VOSS, S. R. Early gene expression during natural spinal cord regeneration in the salamander *Ambystoma mexicanum*. *J Neurochem.*, v. 101, p. 27-40, 2007.
- MUNSCHAUER, M.; et al. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature*, p. 132–136, 2018.
- MURAWALA, H.; RANADIVE, I.; PATEL, S.; DESAI, I.; BALAKRISHNAN, S. Protein expression pattern and analysis of differentially expressed peptides during various stages of tail regeneration in *Hemidactylus flaviviridis*. *Mech Dev.*, v. 150, p. 1-9, 2018.
- MUSACCHIA, F.; et. al. Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. *Bioinformatics*, v. 31, p. 2199–2201, 2015.
- NOGUEIRA, A. F.; COSTA, C. M.; LORENA, J.; MOREIRA, R. N.; FROTA-LIMA, G. N.; FURTADO, C.; ROBINSON, M.; AMEMIYA, C. T.; DARNET, S.; SCHNEIDER, I. Tetrapod Limb and Sarcopterygian Fin Regeneration Share a Core Genetic Programme. *Nature Communications*, v. 7, p. 13364, 2016.
- OriginPro, Version 2021b. OriginLab Corporation, Northampton, MA, USA.
- ØROM, U. A.; et al. Long noncoding RNAs as enhancers of gene expression”. *Cold Spring Harb. Symp. Quant. Biol.*, p. 325–331, 2010.

- PAN J., WANG R., SHANG F., MA R., RONG Y., ZHANG Y. Functional Micropeptides Encoded by Long Non-Coding RNAs: A Comprehensive Review. *Front Mol Biosci*, 13;9:817517, 2022.
- PARASKEVOPOULOU, M. D. and HATZIGEORGIOU, A. G. Analyzing MiRNA–LncRNA Interactions. *Long Non-Coding RNAs: Methods and Protocols*, p. 271–286, 2016.
- PASCUAL-CARRERAS E., MARIN-BARBA M., HERRERA-ÚBEDA C., et al. Planarian cell number depends on blitzschnell, a novel gene family that balances cell proliferation and cell death. *Development*, 147(7): dev184044, 2020.
- PERRY, R. B., HEZRONI, H., GOLDRICH, M. J., ULITSKY, I. Regulation of neuroregeneration by long noncoding RNAs. *Mol. Cell* 72, 553–567.e5 (2018).
- PONOMAREVA, L. V.; ATHIPPOZHY, A.; THORSON, J. S.; VOSS, S. R. Using *Ambystoma mexicanum* (Mexican axolotl) embryos, chemical genetics, and microarray analysis to identify signaling pathways associated with tissue regeneration. *Comp Biochem Physiol C Toxicol Pharmacol.*, v. 178, p. 128-135, 2015.
- QUESENBERRY, P. J.; et al. Role of extracellular RNA-carrying vesicles in cell differentiation and reprogramming. *Stem Cell Res. Ther.*, p. 153, 2015.
- QUINN, J.; CHANG, H. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, v. 17, p. 47–62, 2016.
- RUSSO, P.; et al. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics*, v. 19(56), p. 1–13, 2018.
- SALABI, F.; JAFARI, H.; NAVIDPOUR, S.; et al. Systematic and computational identification of *Androctonus crassicauda* long non-coding RNAs. *Scientific Reports*, v. 11, p. 4720, 2021.
- SALABI, F.; JAFARI, H.; NAVIDPOUR, S.; SADR, A. S. Systematic and computational identification of *Androctonus crassicauda* long non-coding RNAs. *Scientific Reports*, v. 11, p. 4720, 2021.
- SANDOVAL-GUZMÁN, T.; et al. Fundamental Differences in Dedifferentiation and Stem Cell Recruitment during Skeletal Muscle Regeneration in Two Salamander Species. *Cell Stem Cell*, v. 14, p. 174–187, 2014.
- SENGUPTA S.; PETERSON T.R.; SABATINI D.M. Regulation of the mTOR complex 1 pathway by nutrients, growth factors, and stress. *Mol. Cell*, v. 40, p. 310–322. 2010.
- SCHNAPP, E.; KRAGL, M.; RUBIN, L.; TANAKA, E. M. Hedgehog signaling controls dorsoventral patterning, blastema cell proliferation and cartilage induction during axolotl tail regeneration. *Development*, v. 132, p. 3243–3253, 2005.
- SHIMOBAYASHI M.; HALL M.N. Making new contacts: The mTOR network in metabolism and signaling crosstalk. *Nat. Rev. Mol. Cell. Biol.*, v. 15, p. 155–162, 2014.

- SIMÃO, F. A.; WATERHOUSE, R. M.; IOANNIDIS, P.; KRIVENTSEVA, E. V.; ZDOBNOV, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, v. 31, n. 19, p. 3210–3212, 2015.
- SIMON, A.; TANAKA, E. M. *Limb Regeneration: Limb Regeneration*. Wiley Interdisciplinary Reviews: Developmental Biology, v. 2, p. 291–300, 2013.
- SINGH, B. N.; WEAVER, C. V.; GARRY, M. G.; GARRY, D. J. Hedgehog and Wnt signaling pathways regulate tail regeneration. *Stem Cells Dev.*, v. 27, p. 1426–1437, 2018.
- STAJICH, J. E. et al. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, v. 12, n. 10, p. 1611–1618, 2002.
- STEWART, R. et al. Comparative RNA-seq analysis in the unsequenced axolotl: the oncogene burst highlights early gene expression in the blastema. *PLoS Computational Biology*, v. 9, p. e1002936, 2013.
- STOCUM, D. L. Limb regeneration: Re-entering the cell cycle. *Current Biology*, v. 9, n. 17, p. R644–R646, 1999.
- STOCUM, D. L.; CAMERON, J. A. Looking proximally and distally: 100 years of limb regeneration and beyond. *Developmental Dynamics*, v. 240, p. 943–968, 2011.
- STRUHL, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, pp. 103–105, 2007.
- SUGIURA, T.; WANG, H.; BARSACCHI, R.; SIMON, A.; TANAKA, E. M. MARCKS-like protein is an initiating molecule in axolotl appendage regeneration. *Nature*, v. 531, p. 237–240, 2016.
- TRAQUAIR, R. H. 1871 On the restoration of the tail in *Protopterus annectens*. *Rep. Brit. Ass. Adv. Sci.*, 143, 1871.
- ULITSKY, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet.* 17(10):601–14, 2016.
- ULITSKY, I.; et al. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, p. 1537–1550, 2011.
- VAGLIA, J.L.; BABCOCK, S.K.; HARRIS, R.N. Tail development and regeneration throughout the life cycle of the four-toed salamander *Hemidactylium scutatum*. *Journal of morphology*, v. 233, p.15–29, 1997.
- van BAKEL, H. et al. Most “dark matter” transcripts are associated with known genes. *PLoS Biol.*, p. e1000371, 2010.
- VERISSIMO, K. M.; PEREZ, L. N.; DRAGALZEW, A. C.; SENEVIRATHNE, G.; DARNET, S.; BARROSO, M. W. R.; ARIEL, S. N. C.; MONTEIRO, S. E.; NAZARE, S. M. C.; ELEWA, A.; SHUBIN, N.; FRÖBISCH, N. B.; FREITAS, S. J.; SCHNEIDER, I. Salamander-like tail regeneration in the West African lungfish. *Proc. R. Soc.*, v. 287, p. 20192939, 2020.

VIDAK S.; GEORGIOU K.; et al. Nucleoplasmic lamins define growth-regulating functions of lamina-associated polypeptide 2alpha in progeria cells. *J. Cell Sci*, p. 131, 2018.

VIEIRA, W.A.; McCUSKER, C.D. Regenerative Models for the Integration and Regeneration of Head Skeletal Tissues. *Int. J. Mol. Sci.* 19, 3752, 2018.

VITULO, N.; DALLA VALLE, L.; SKOBO, T.; VALLE, G.; ALIBARDI, L. Transcriptome analysis of the regenerating tail vs. the scarring limb in lizard reveals pathways leading to successful vs. unsuccessful organ regeneration in amniotes. *Dev Dyn.*, v. 246, p. 116-134, 2017.

WANG, F.; et al. Coregulatory long non-coding RNA and protein-coding genes in serum starved cells. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, p. 84–95, 2019.

WANG, H.; SIMON, A. Skeletal Muscle Dedifferentiation during Salamander Limb Regeneration. *Current Opinion in Genetics & Development*, v. 40, p. 108–112, 2016.

WANG, K. C.; CHANG, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, p. 904–914, 2011.

WANG, K.; WANG, J.; ZHU, C.; et al. African lungfish genome sheds light on the vertebrate water-to-land transition. *Cell*, 184(5):1362-1376.e18, 2021.

WONG, R. C.; IBRAHIM, A.; FONG, H.; THOMPSON, N.; LOCK, L. F.; DONOVAN, P. J. L1TD1 is a marker for undifferentiated human embryonic stem cells. *PLoS One*, v. 6, p. e19355, 2011.

WU, C. H.; TSAI, M. H.; HO, C. C.; CHEN, C. Y.; LEE, H. S. De novo transcriptome sequencing of axolotl blastema for identification of differentially expressed genes during limb regeneration. *BMC Genomics*, v. 14, p. 434, 2013.

WUCHER, V.; LEGEAI, F.; HÉDAN, B.; RIZK, G.; LAGOUTTE, L.; LEEB, T.; JAGANNATHAN, V.; CADIEU, E.; DAVID, A.; LOHI, H.; CIRERA, S.; FREDHOLM, M.; et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res*, 5;45(8):e57, 2017.

XING J., LIU H., JIANG W., WANG L. LncRNA-Encoded Peptide: Functions and Predicting Methods. *Frontiers in Oncology*, v.10, 2021.

XUEZHONG C. & BRYAN R. C. The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA*, p. 313–316, 2007.

YANG, Z.; et al. Overexpression of Long Non-coding RNA HOTAIR Predicts Tumor Recurrence in Hepatocellular Carcinoma Patients Following Liver Transplantation. *Annals of Surgical Oncology*, p. 1243–1250, 2011.

YAO C.; YU B. Role of Long Noncoding RNAs and Circular RNAs in Nerve Regeneration. *Front Mol Neurosci*, v. 12, p. 165, 2019.

YOKOYAMA, H. Initiation of limb regeneration: the critical steps for regenerative capacity. *Dev Growth Differ.*, v. 50, p. 13-22. 2008.

YOON M.S. mTOR as a Key Regulator in Maintaining Skeletal Muscle Mass. *Front. Physiol.*, v. 8, p. 788, 2017.

YUNUSOV, D.; et al. HIPSTR and thousands of lncRNAs are heterogeneously expressed in human embryos, primordial germ cells and stable cell lines. *Sci. Rep.*, p. 32753, 2016.

ZHANG, X. et al. Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *Int. J. Mol. Sci.*, 2019.

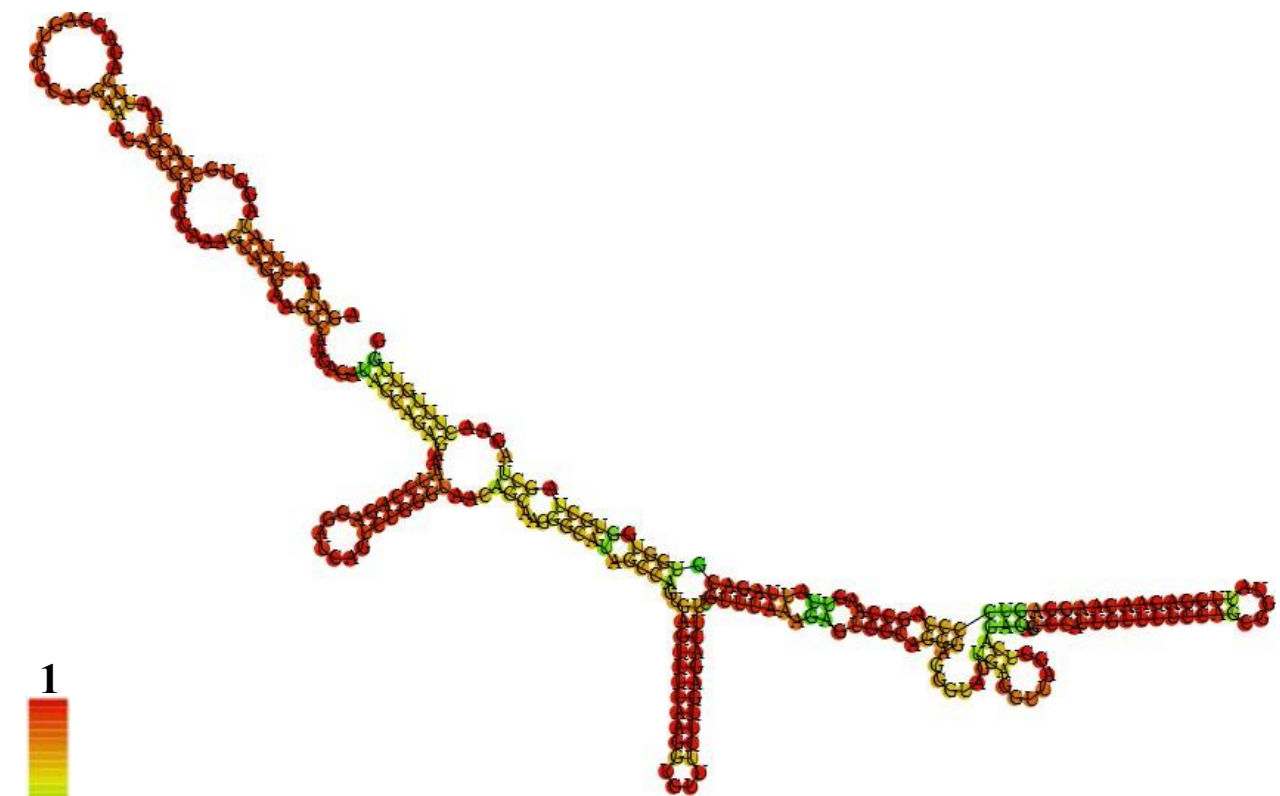
ZHANG, X.; et al. A Pituitary-Derived MEG3 Isoform Functions as a Growth Suppressor in Tumor Cells. *The Journal of Clinical Endocrinology & Metabolism*, p. 5119–5126, 2003.

ZHU, J.; YU, W.; WANG, Y. et al. LncRNAs: function and mechanism in cartilage development, degeneration, and regeneration. *Stem Cell Research & Therapy*, v. 10, p. 344, 2019.

Apêndices

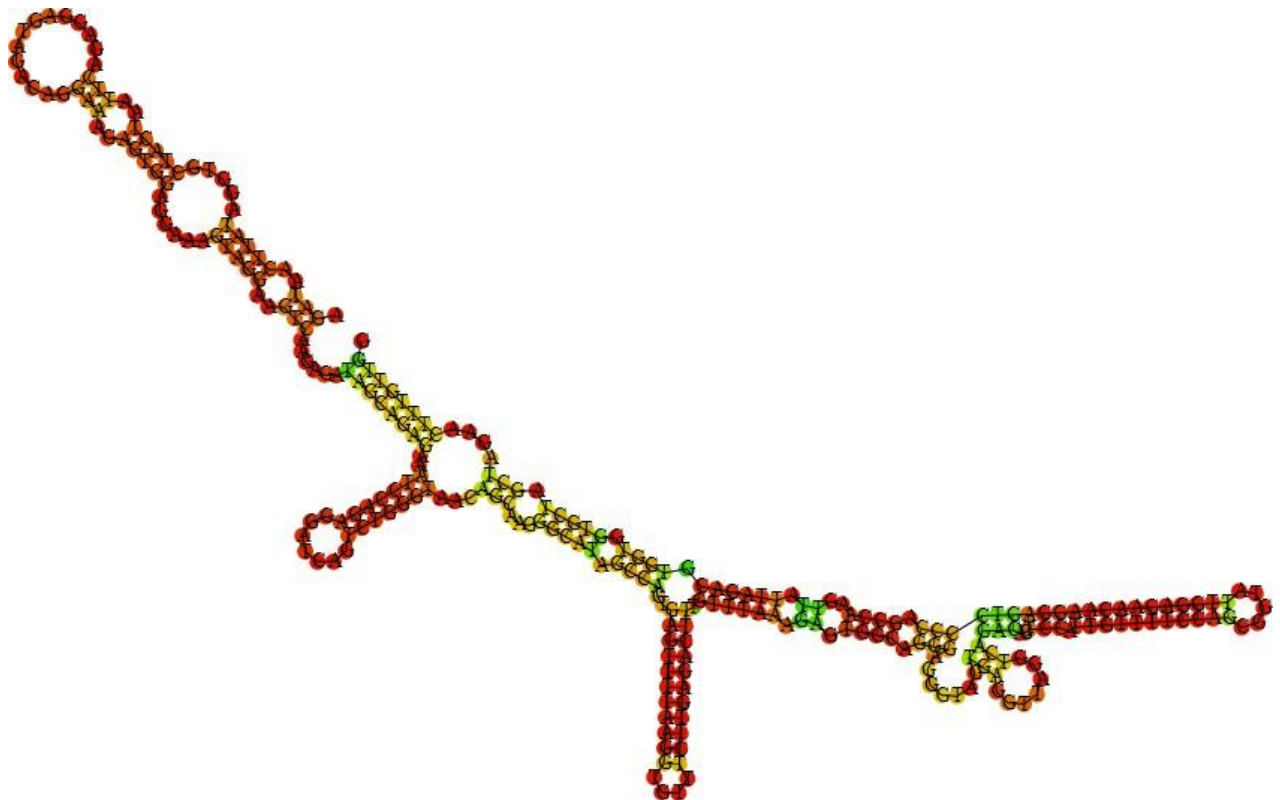
Apêndice A - Estruturas secundárias provenientes da análise do RNAfold web suit.

TRINITY_DN104611_c0_g2_i1



MFE secondary structure

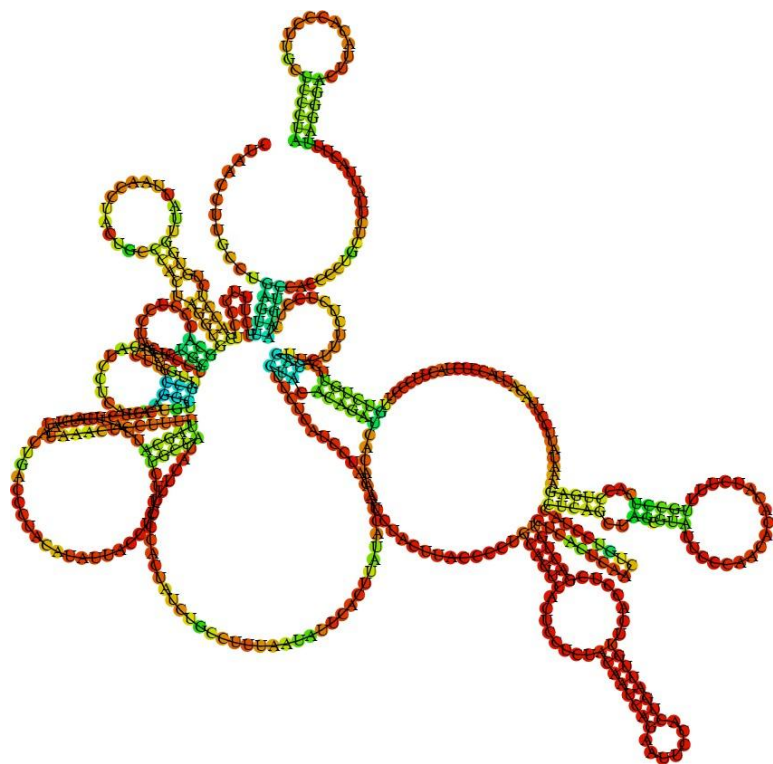
$\Delta G = -74.50$ kcal/mol



Centroid secondary structure

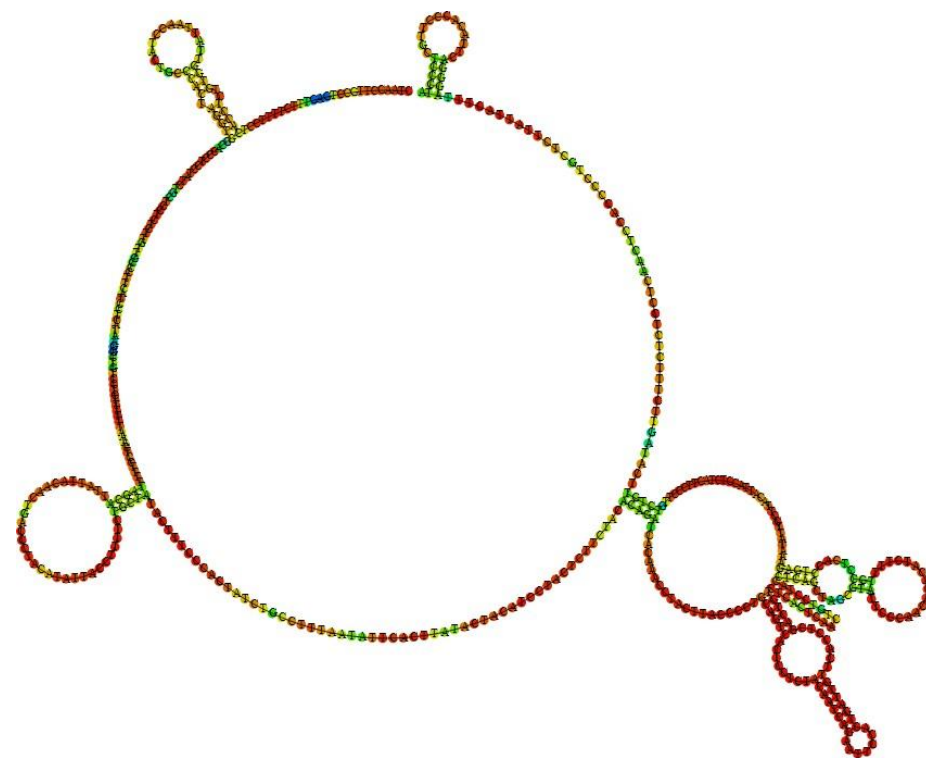
$\Delta G = -74.50$ kcal/mol

TRINITY_DN108590_c1_g2_i1



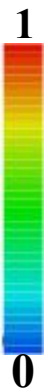
MFE secondary structure

$\Delta G = -50.54$ kcal/mol

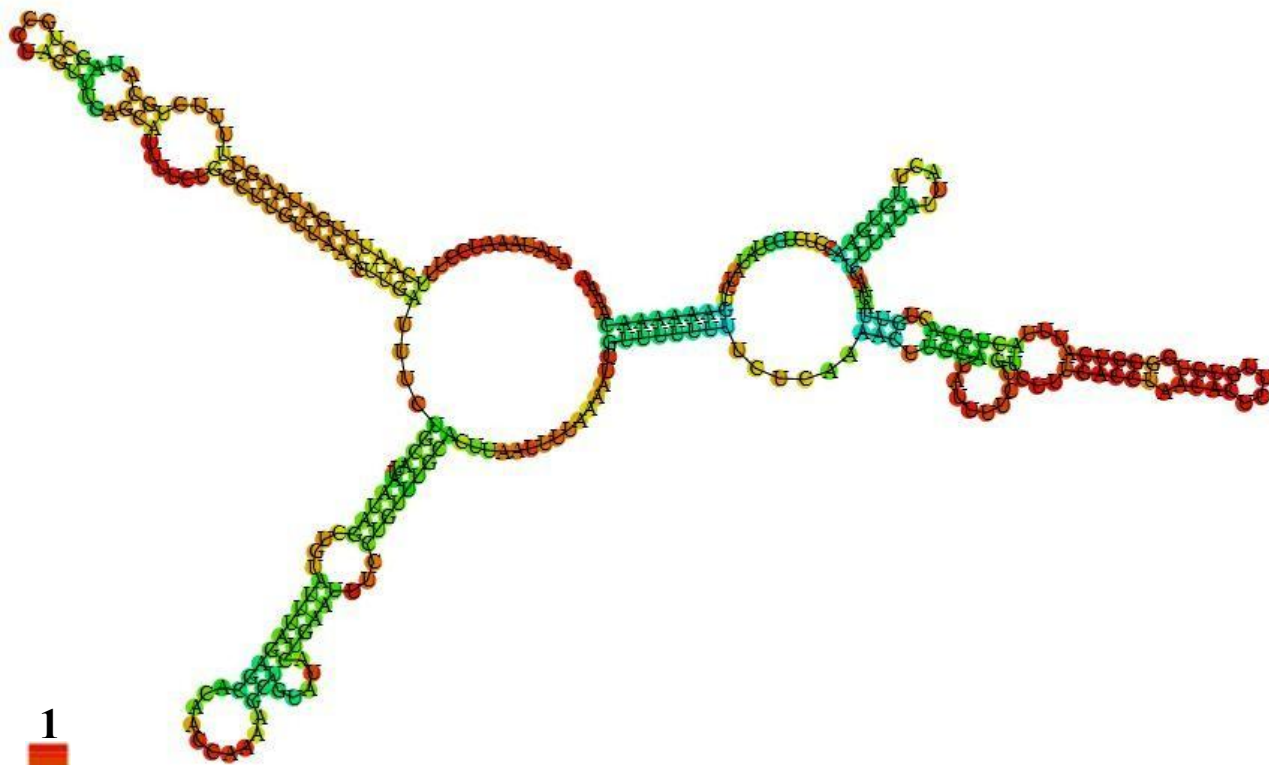


Centroid secondary structure

$\Delta G = -40.64$ kcal/mol



TRINITY_DN116435_c0_g1_i2



MFE secondary structure

$\Delta G = -43.00$ kcal/mol

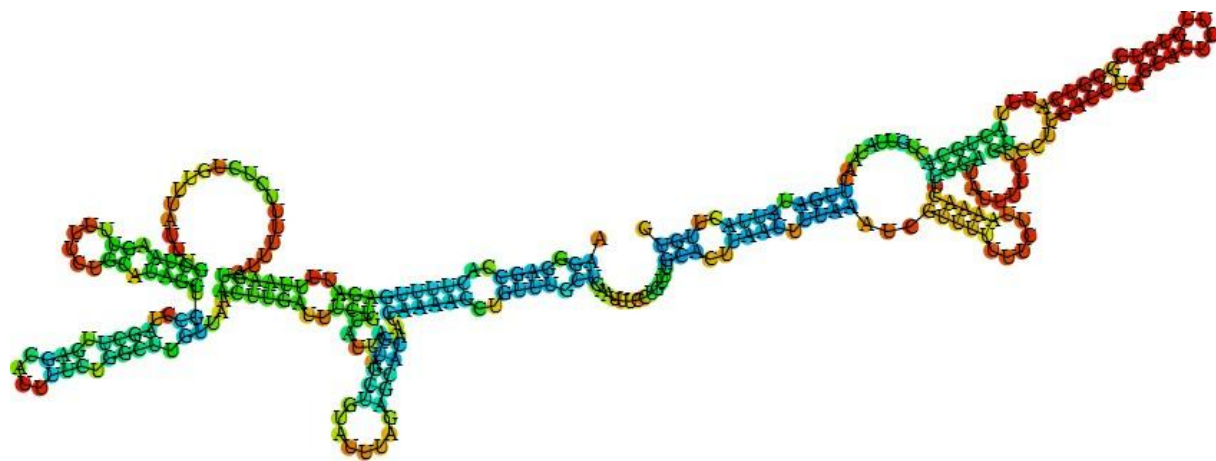


Centroid secondary structure

$\Delta G = -31.40$ kcal/mol

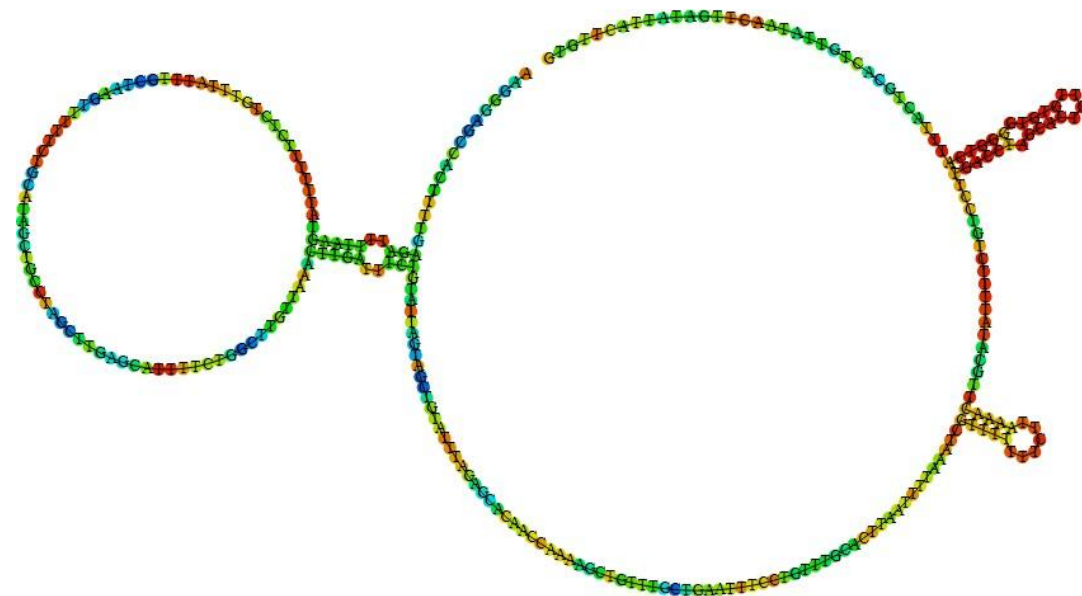


TRINITY_DN116435_c0_g3_i1



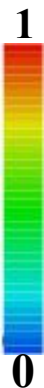
MFE secondary structure

$\Delta G = -46.50$ kcal/mol

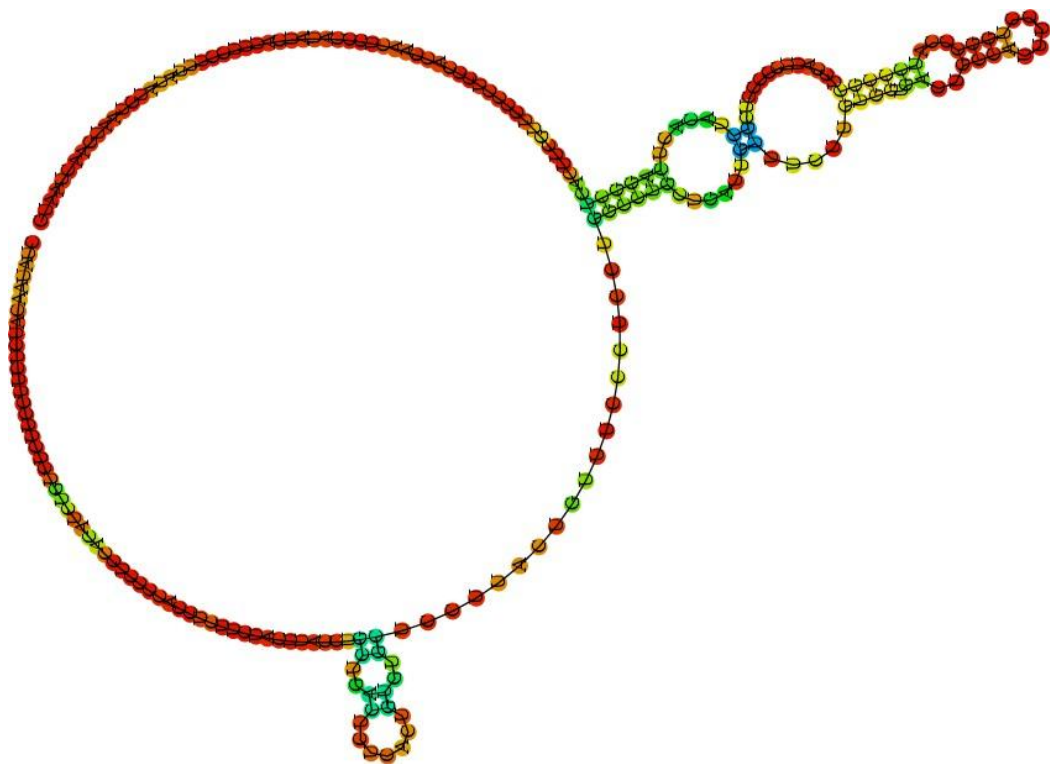


Centroid secondary structure

$\Delta G = -17.96$ kcal/mol

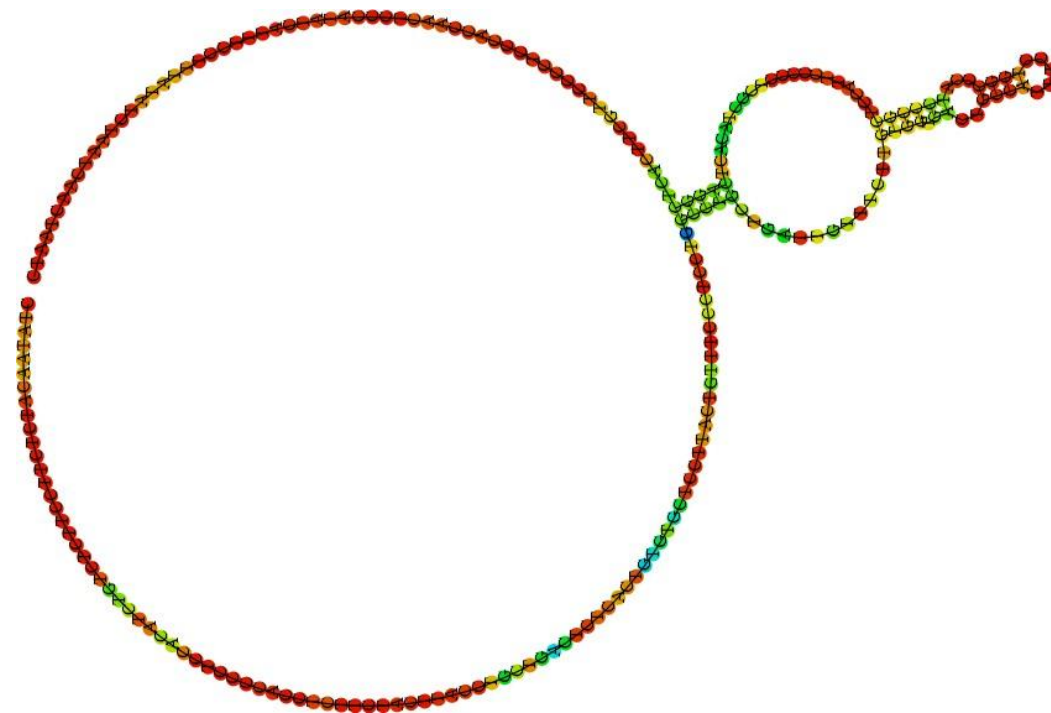


TRINITY_DN118383_c0_g1_i1



MFE secondary structure

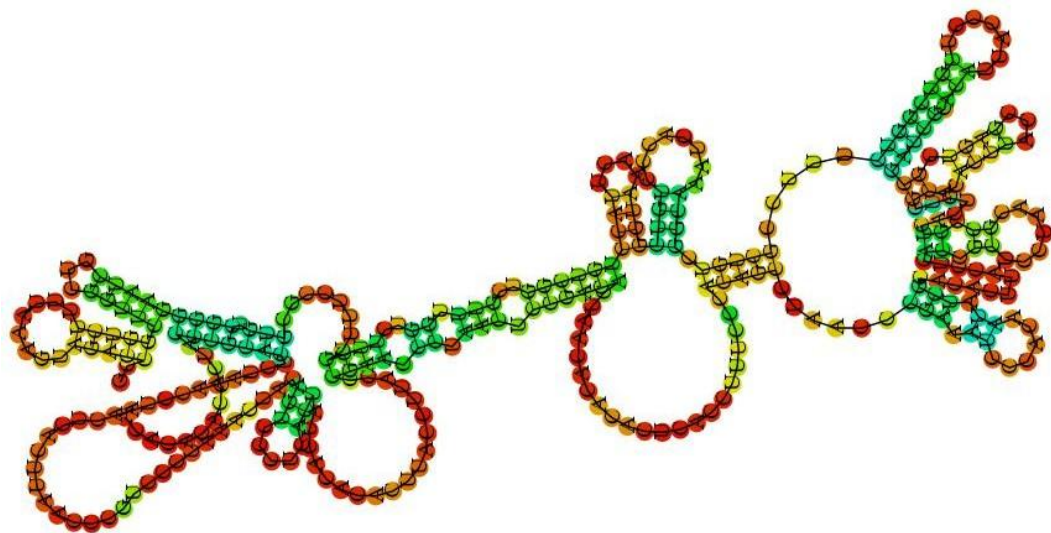
$\Delta G = -18.40$ kcal/mol



Centroid secondary structure

$\Delta G = -14.87$ kcal/mol

TRINITY_DN119279_c0_g5_i1



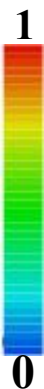
MFE secondary structure

$\Delta G = -52.10$ kcal/mol



Centroid secondary structure

$\Delta G = -26.70$ kcal/mol

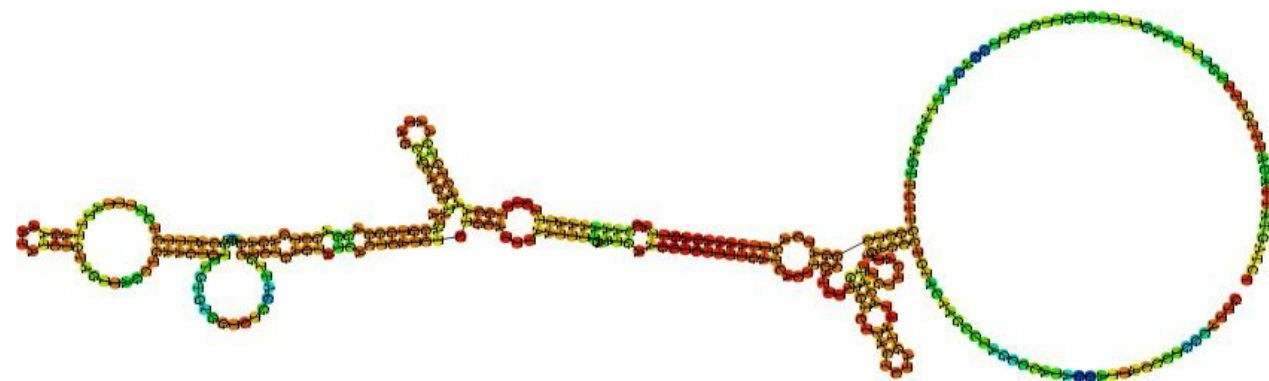


TRINITY_DN119387_c3_g1_i1



MFE secondary structure

$\Delta G = -68.90$ kcal/mol

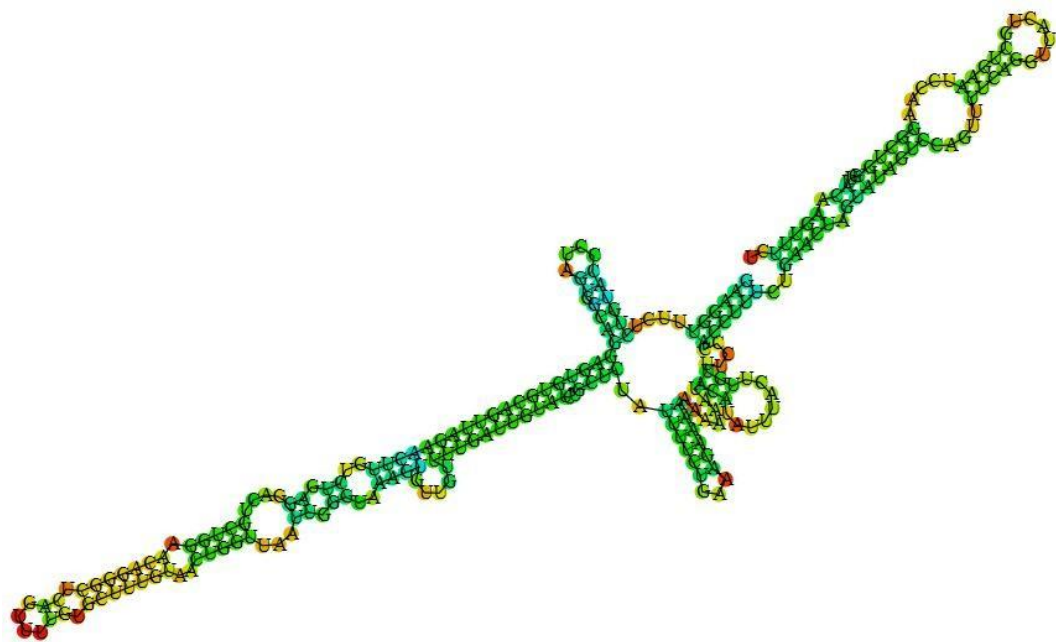


Centroid secondary structure

$\Delta G = -55.90$ kcal/mol

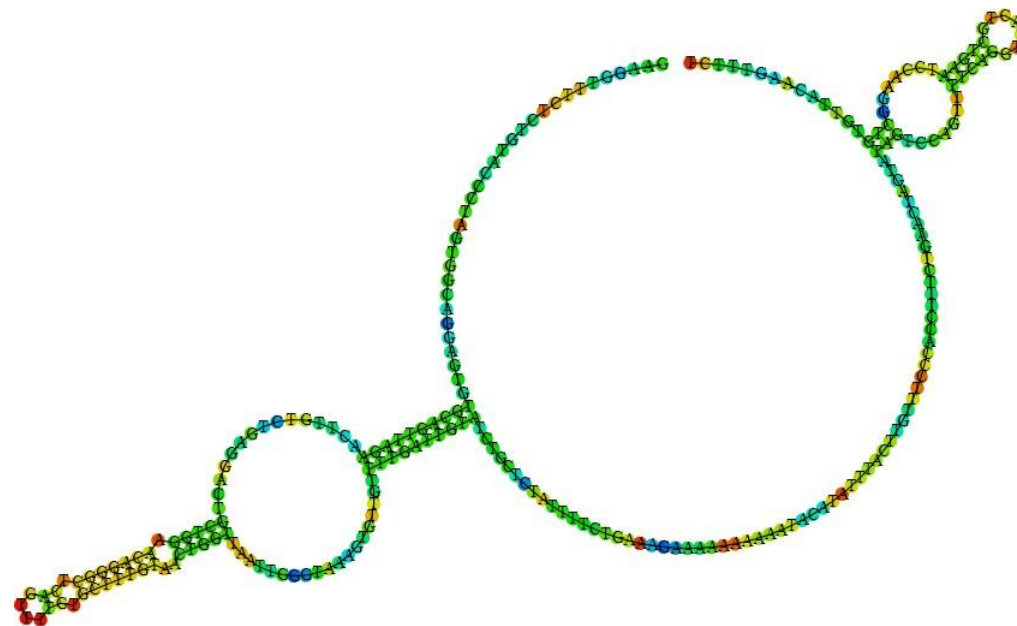


TRINITY_DN119952_c2_g2_i1



MFE secondary structure

$\Delta G = -52.00$ kcal/mol



Centroid secondary structure

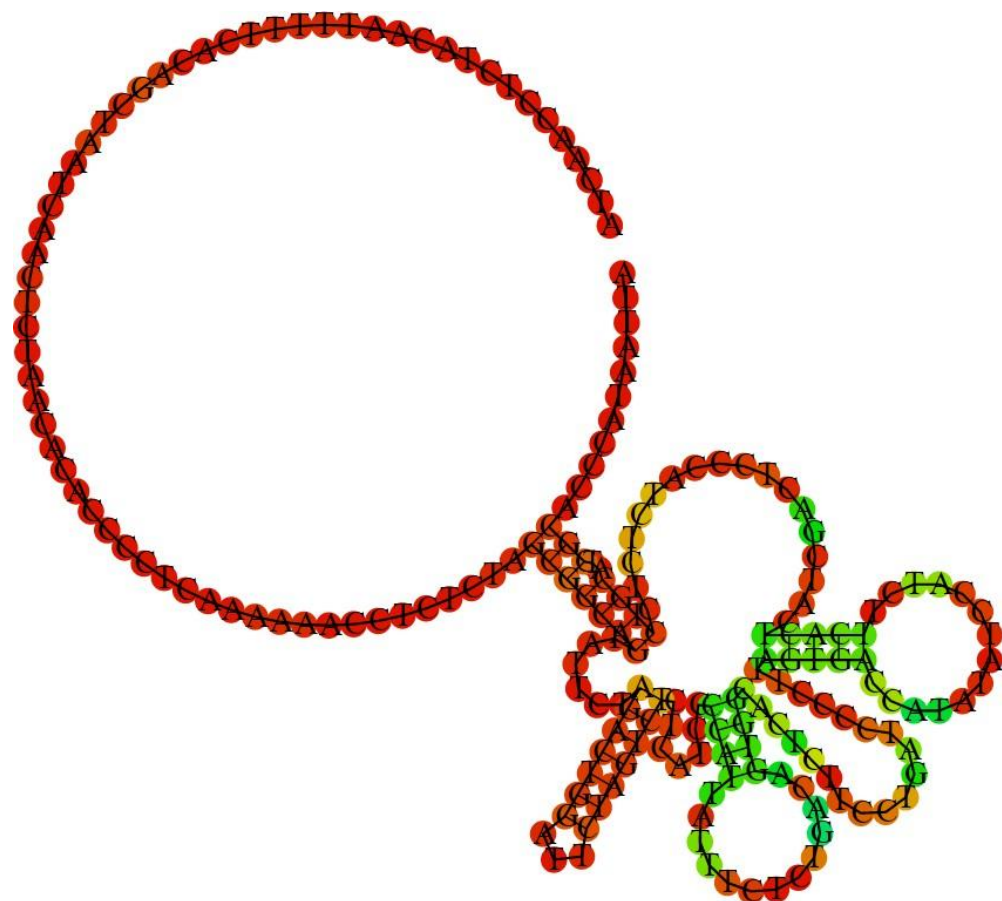
$\Delta G = -20.17$ kcal/mol

TRINITY_DN121120_c4_g5_i2



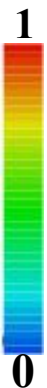
MFE secondary structure

$\Delta G = -19.40$ kcal/mol

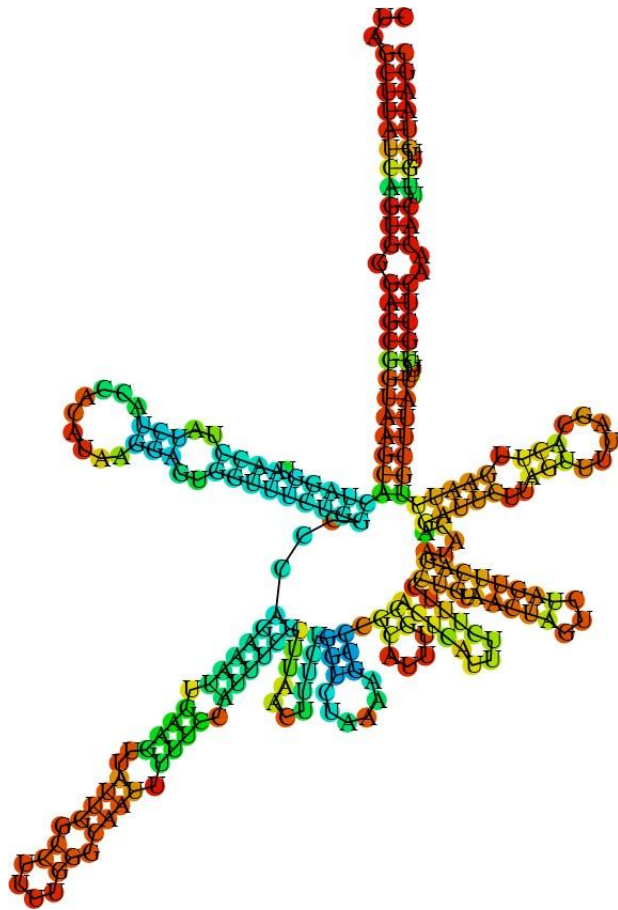


Centroid secondary structure

$\Delta G = -16.90$ kcal/mol

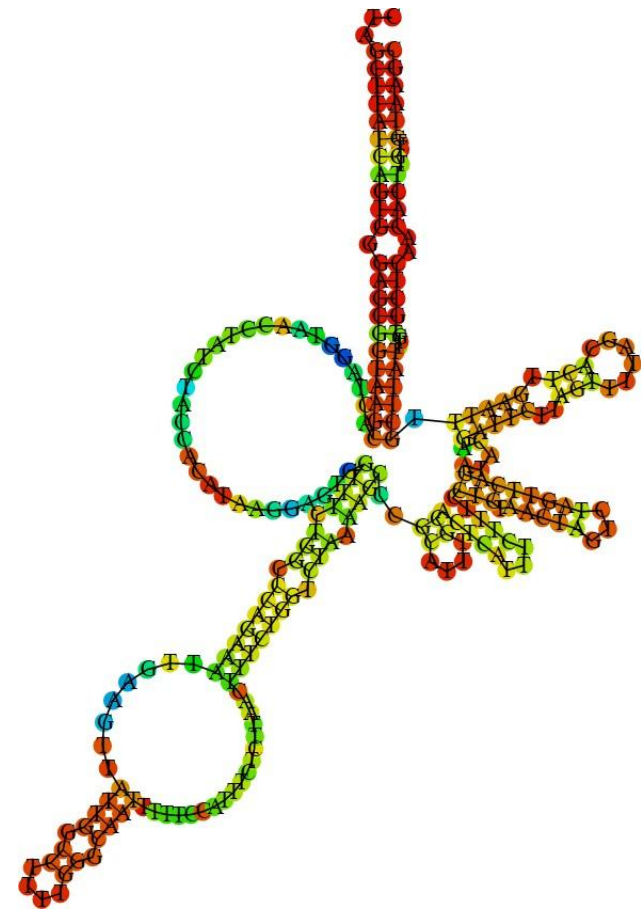


TRINITY_DN121327_c2_g8_i1



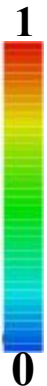
MFE secondary structure

$\Delta G = -56.70$ kcal/mol

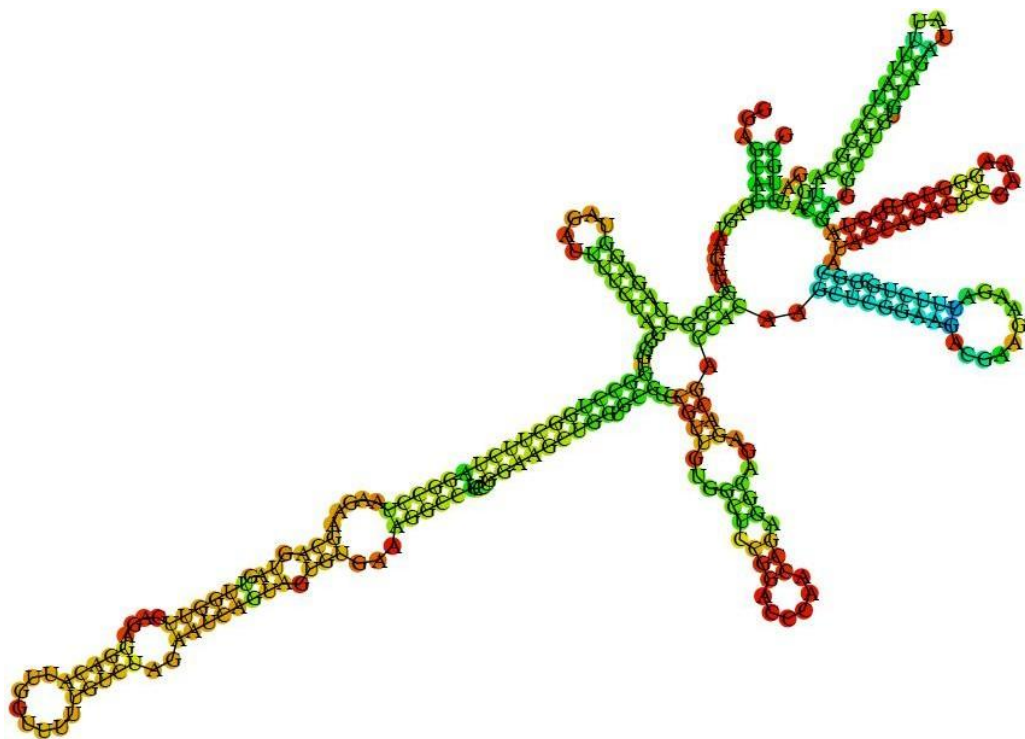


Centroid secondary structure

$\Delta G = -45.20$ kcal/mol

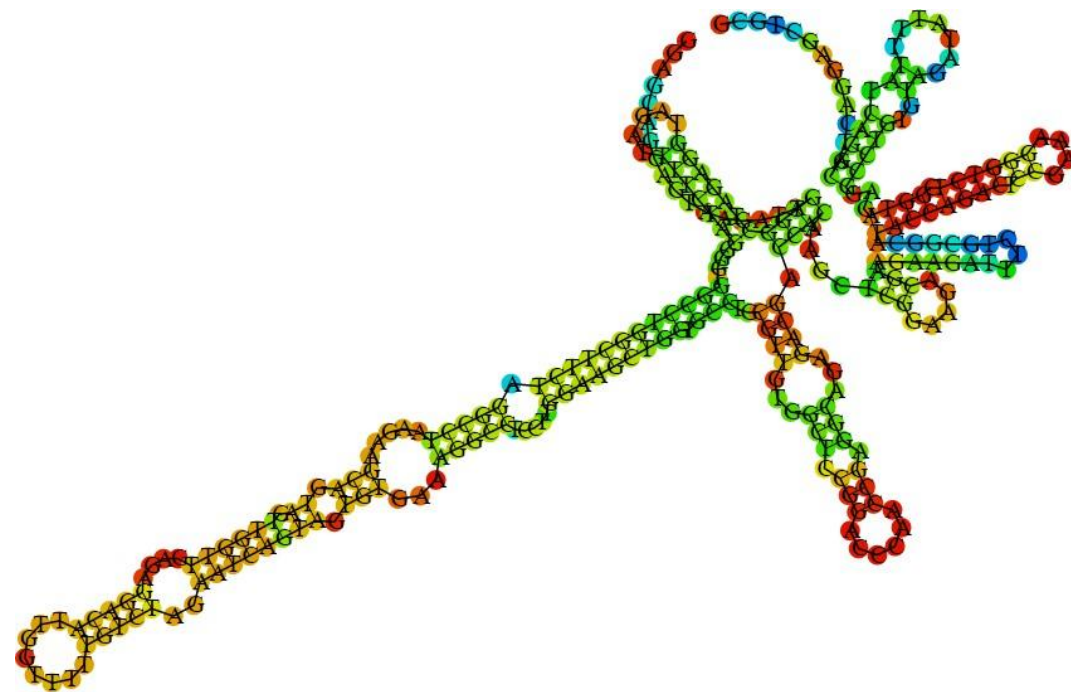


TRINITY_DN121506_c5_g2_i1



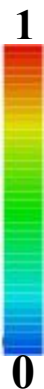
MFE secondary structure

$\Delta G = -94.30$ kcal/mol

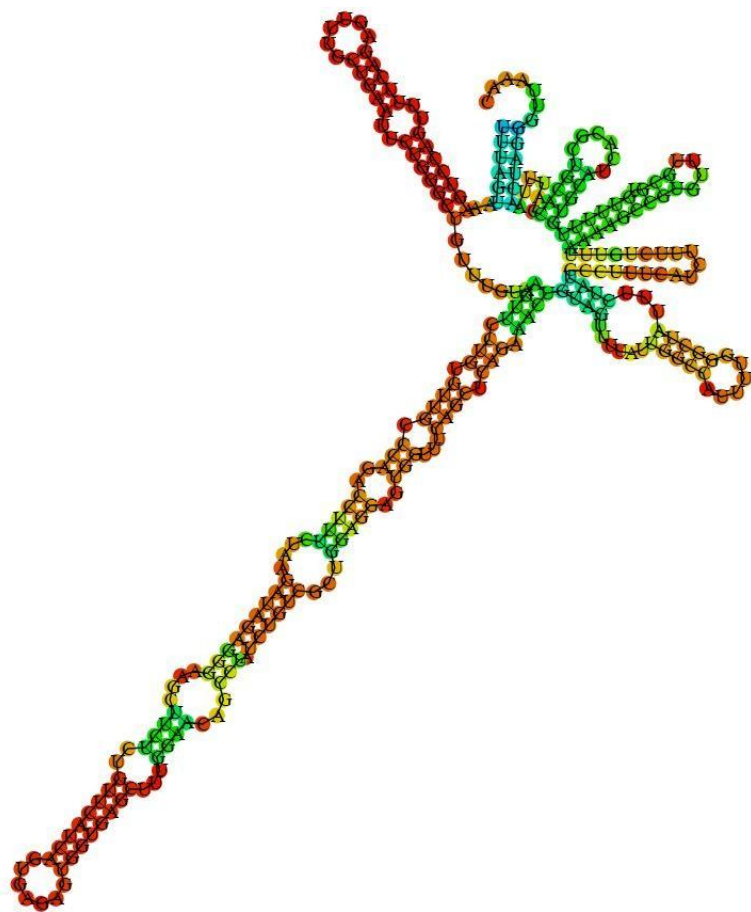


Centroid secondary structure

$\Delta G = -78.70$ kcal/mol

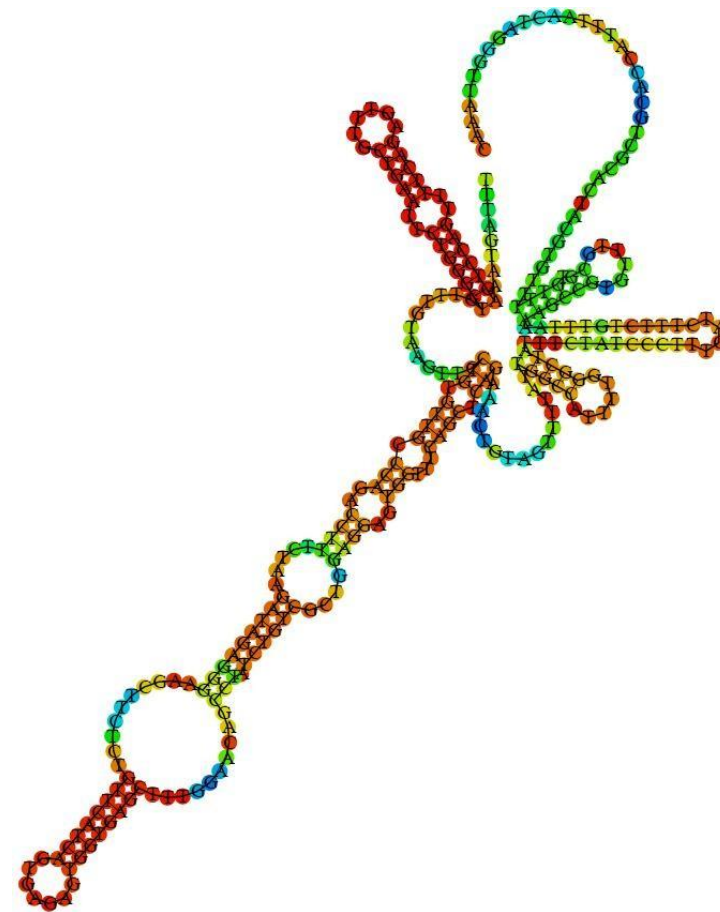


TRINITY_DN121541_c0_g1_i2



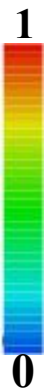
MFE secondary structure

$\Delta G = -79.40$ kcal/mol

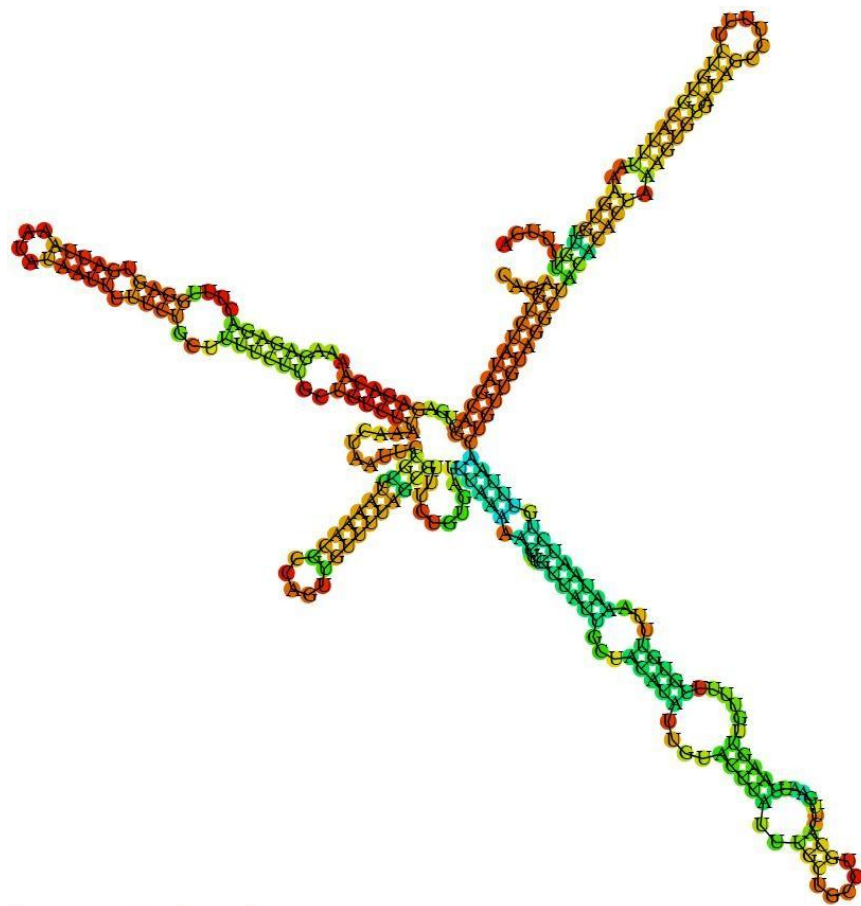


Centroid secondary structure

$\Delta G = -60.50$ kcal/mol



TRINITY_DN121840_c0_g2_i1



MFE secondary structure

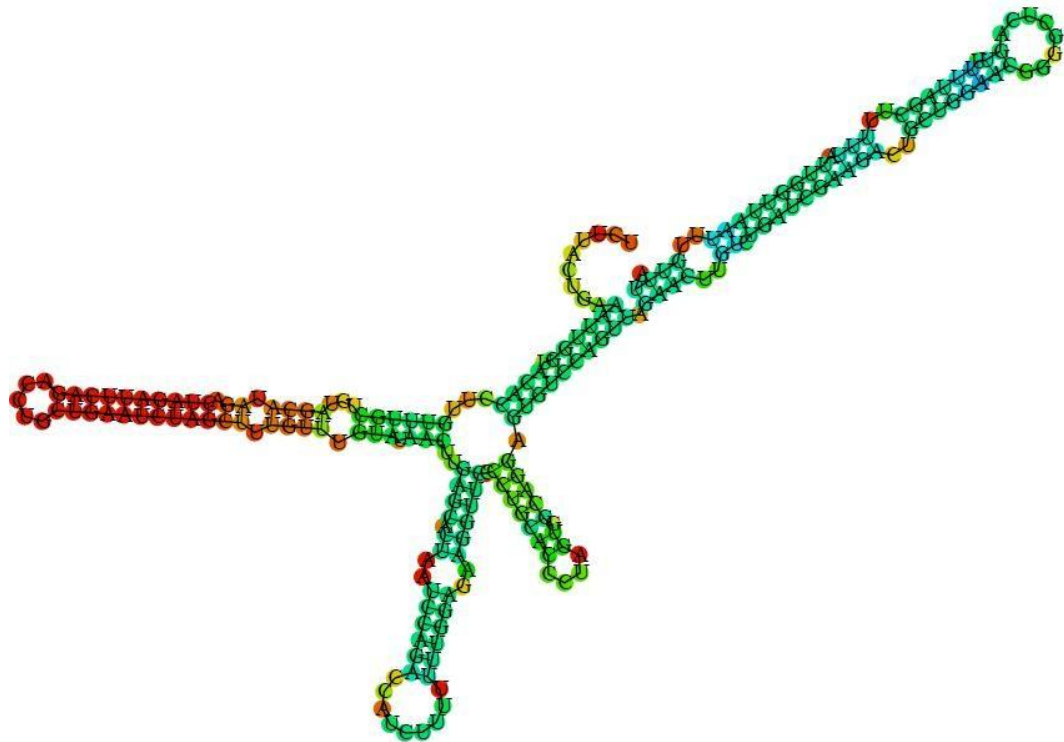
$\Delta G = -55.70$ kcal/mol



Centroid secondary structure

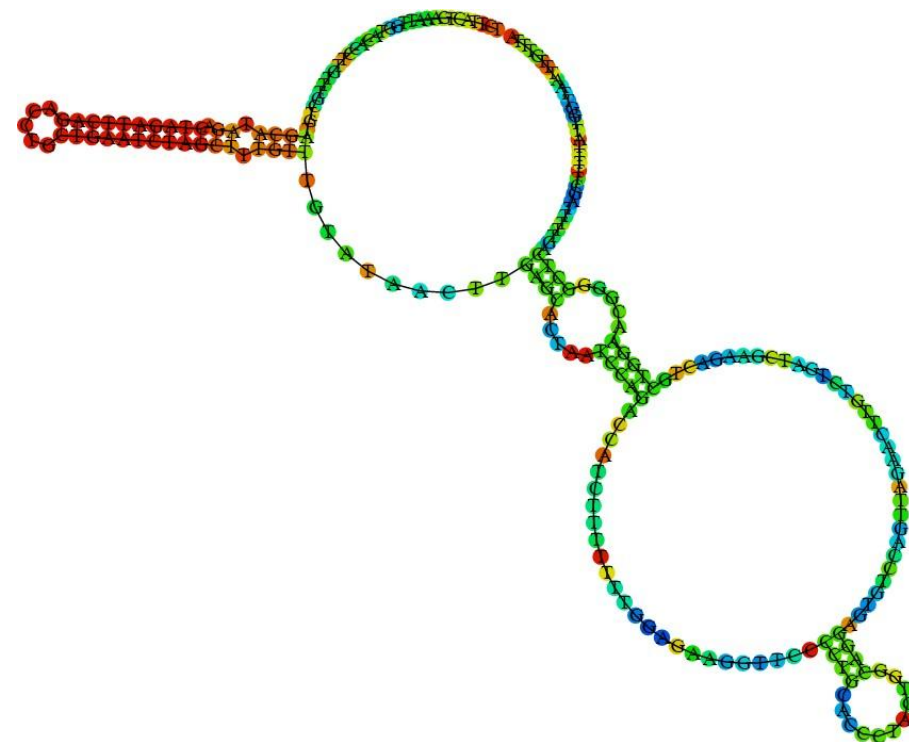
$\Delta G = -38.80$ kcal/mol

TRINITY_DN122360_c4_g2_i2



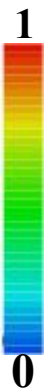
MFE secondary structure

$\Delta G = -59.30$ kcal/mol

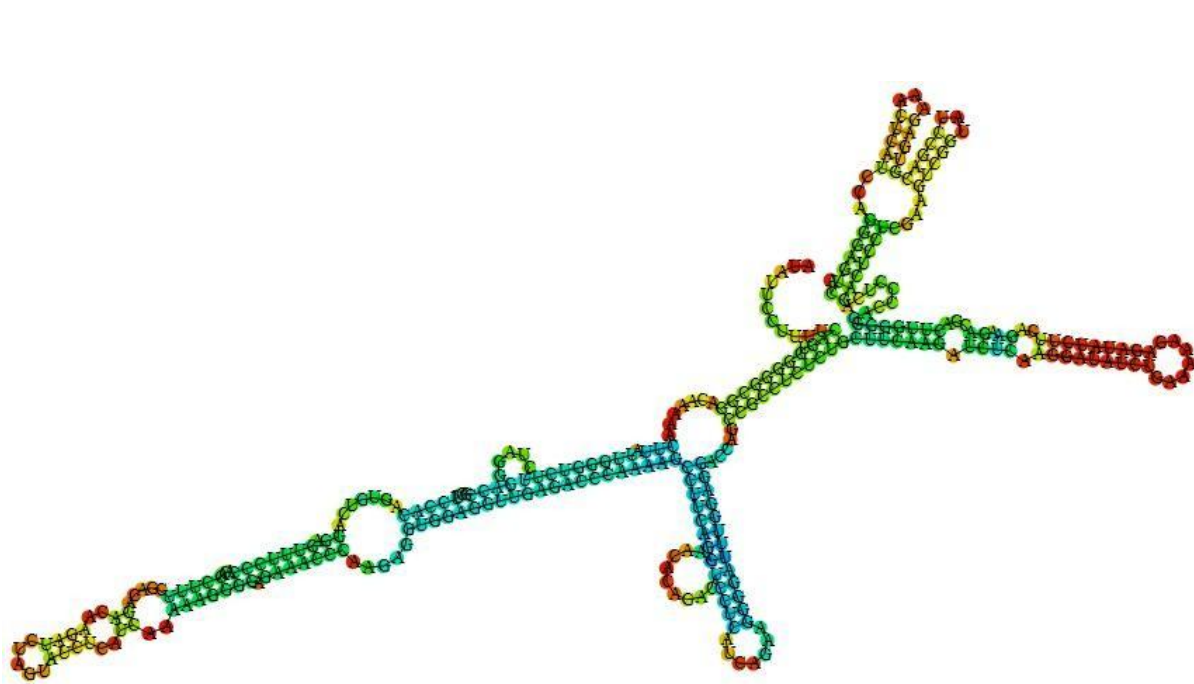


Centroid secondary structure

$\Delta G = -29.78$ kcal/mol

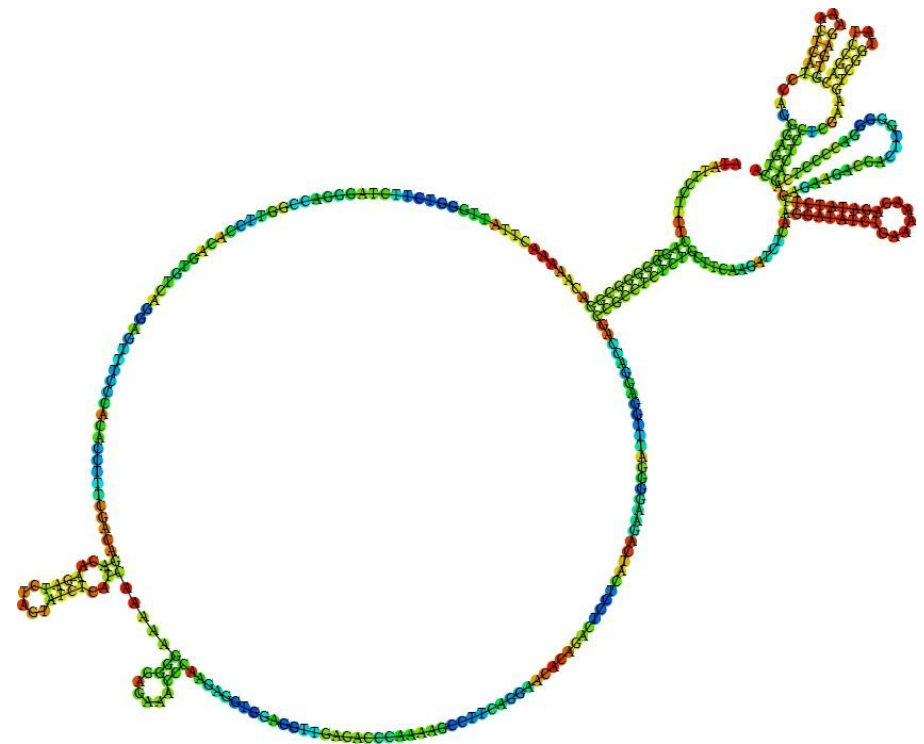


TRINITY_DN122542_c1_g2_i2



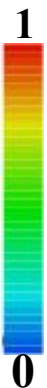
MFE secondary structure

$\Delta G = -97.20$ kcal/mol

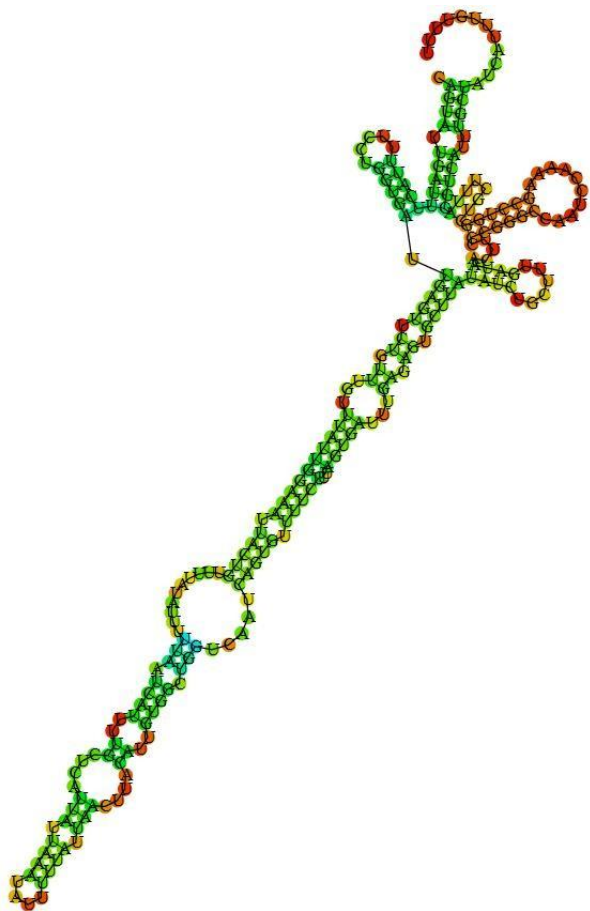


Centroid secondary structure

$\Delta G = -48.20$ kcal/mol

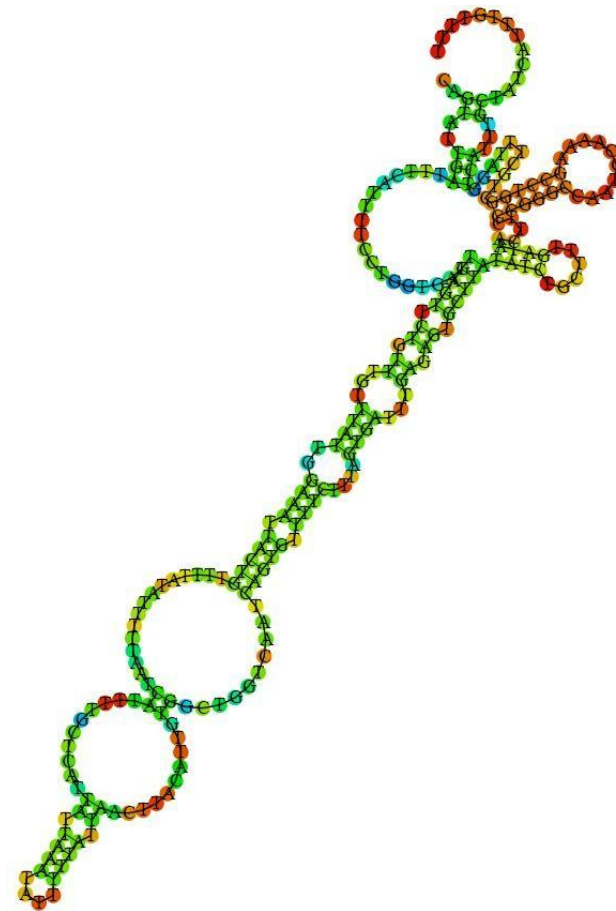


TRINITY_DN122682_c0_g1_i1



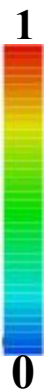
MFE secondary structure

$\Delta G = -34.80$ kcal/mol



Centroid secondary structure

$\Delta G = -18.10$ kcal/mol





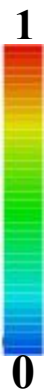
MFE secondary structure

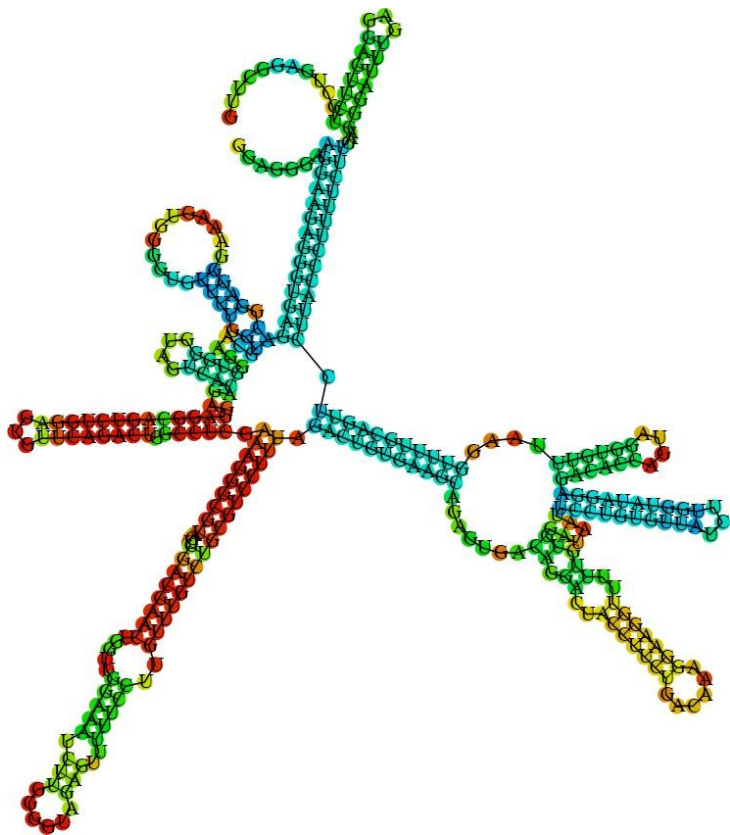
$\Delta G = -75.10$ kcal/mol



Centroid secondary structure

$\Delta G = -68.30$ kcal/mol





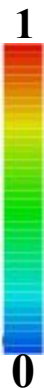
MFE secondary structure

$\Delta G = -109.70$ kcal/mol

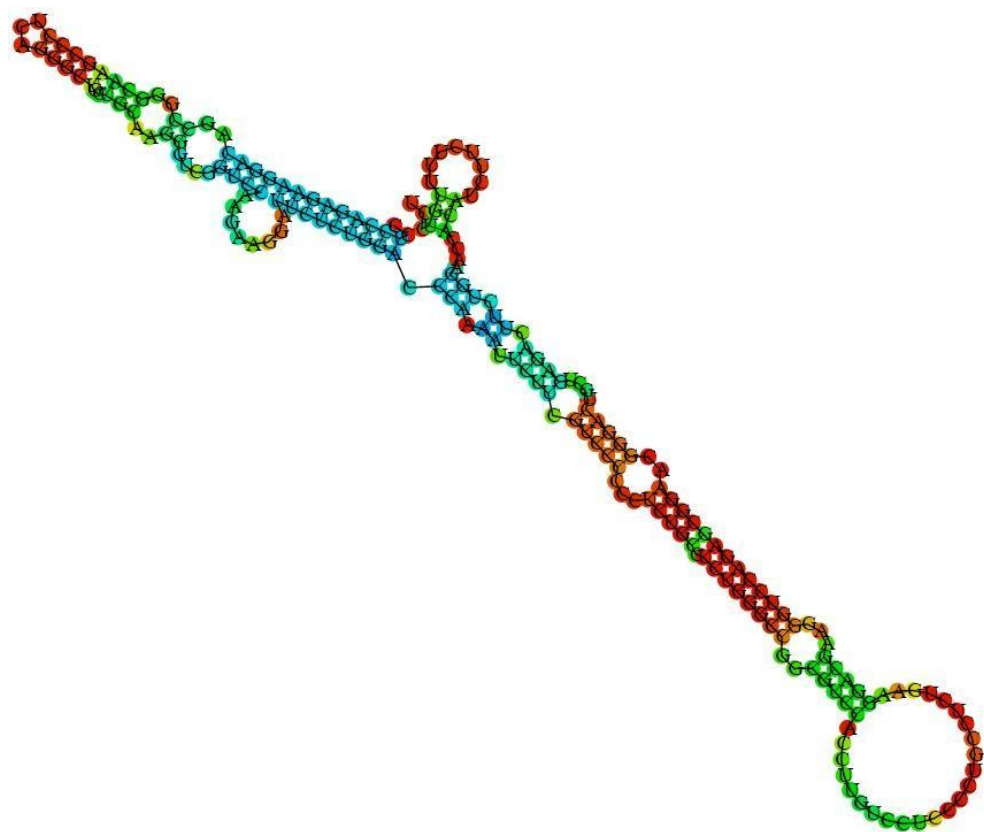


Centroid secondary structure

$\Delta G = -77.80$ kcal/mol

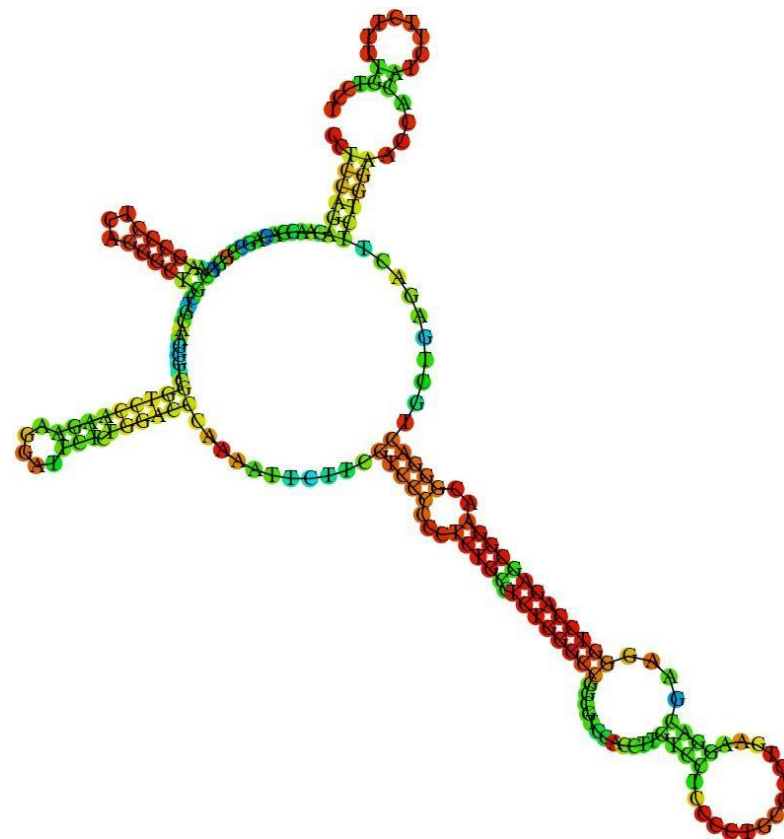


TRINITY_DN125294_c1_g8_i1



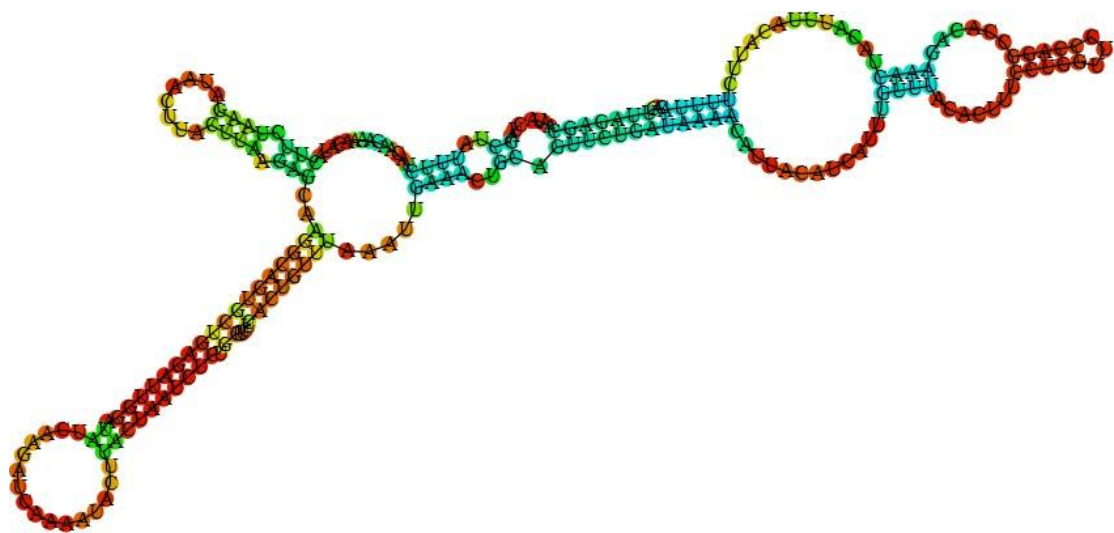
MFE secondary structure

$\Delta G = -72.20$ kcal/mol



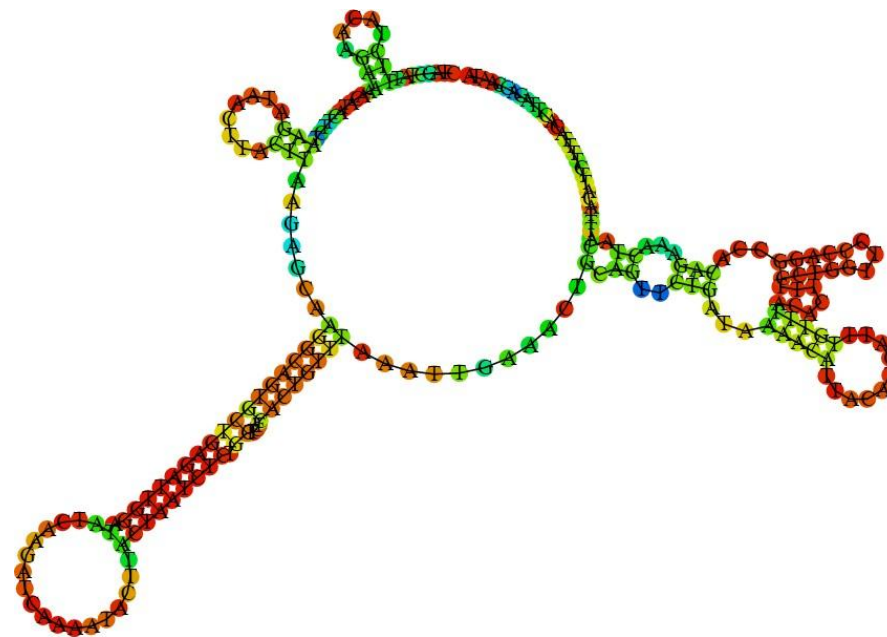
Centroid secondary structure

$\Delta G = -54.50$ kcal/mol



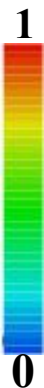
MFE secondary structure

$\Delta G = -33.60$ kcal/mol

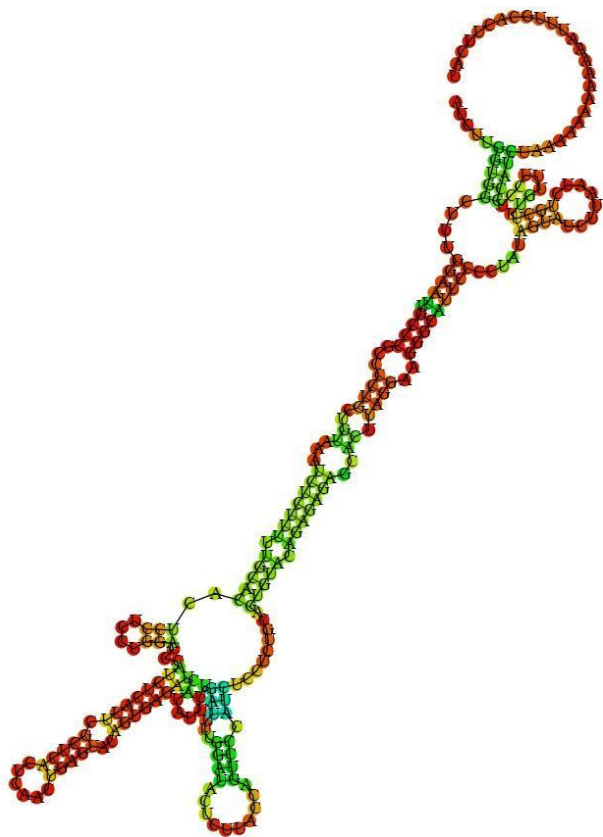


Centroid secondary structure

$\Delta G = -16.00$ kcal/mol

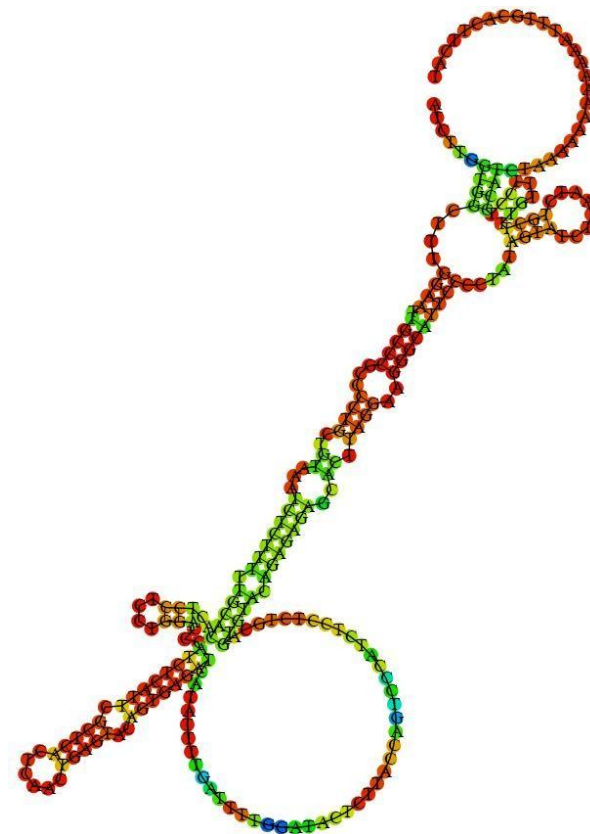


TRINITY_DN126208_c2_g3_i2



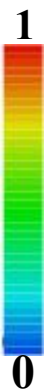
MFE secondary structure

$$\Delta G_{\text{MFE}} = -52.80 \text{ kcal/mol}$$

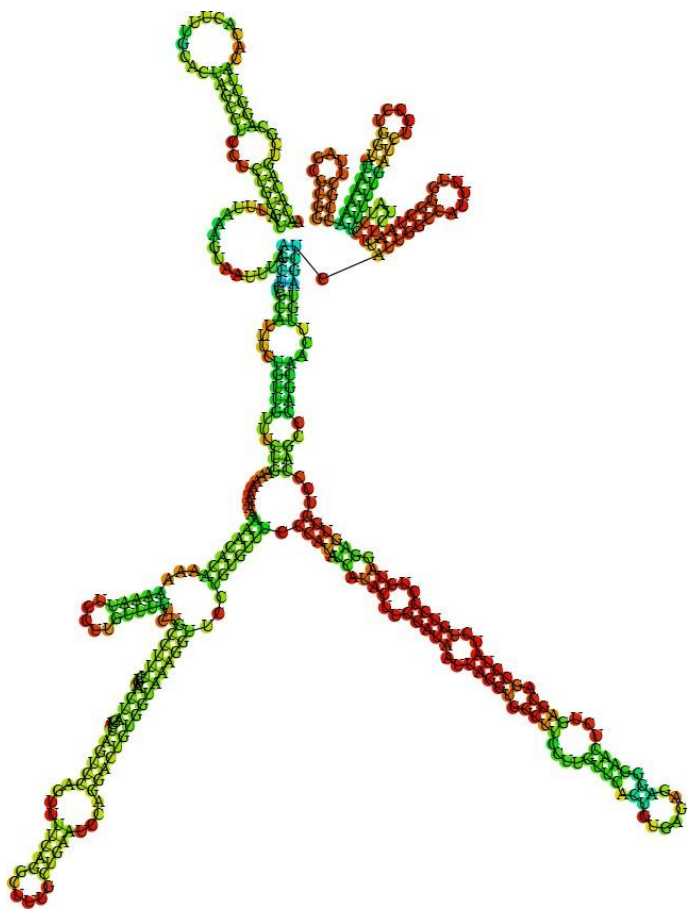


Centroid secondary structure

$$\Delta G_{\text{CENTROID}} = -46.60 \text{ kcal/mol}$$

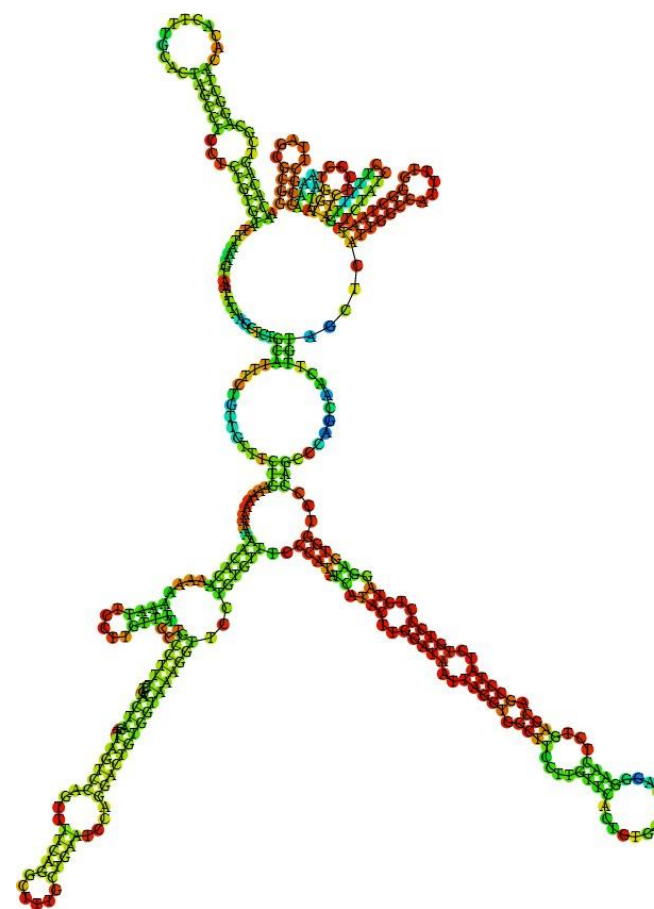


TRINITY_DN126278_c2_g1_i1



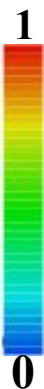
MFE secondary structure

$\Delta G = -90.90$ kcal/mol

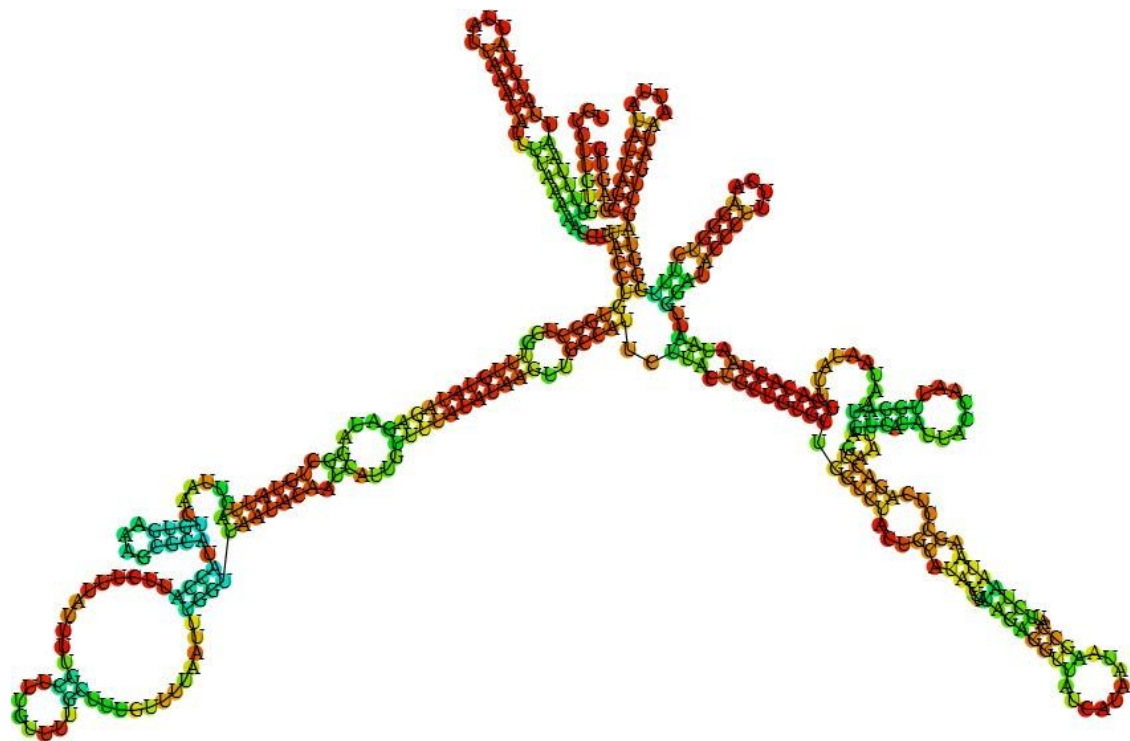


Centroid secondary structure

$\Delta G = -75.30$ kcal/mol

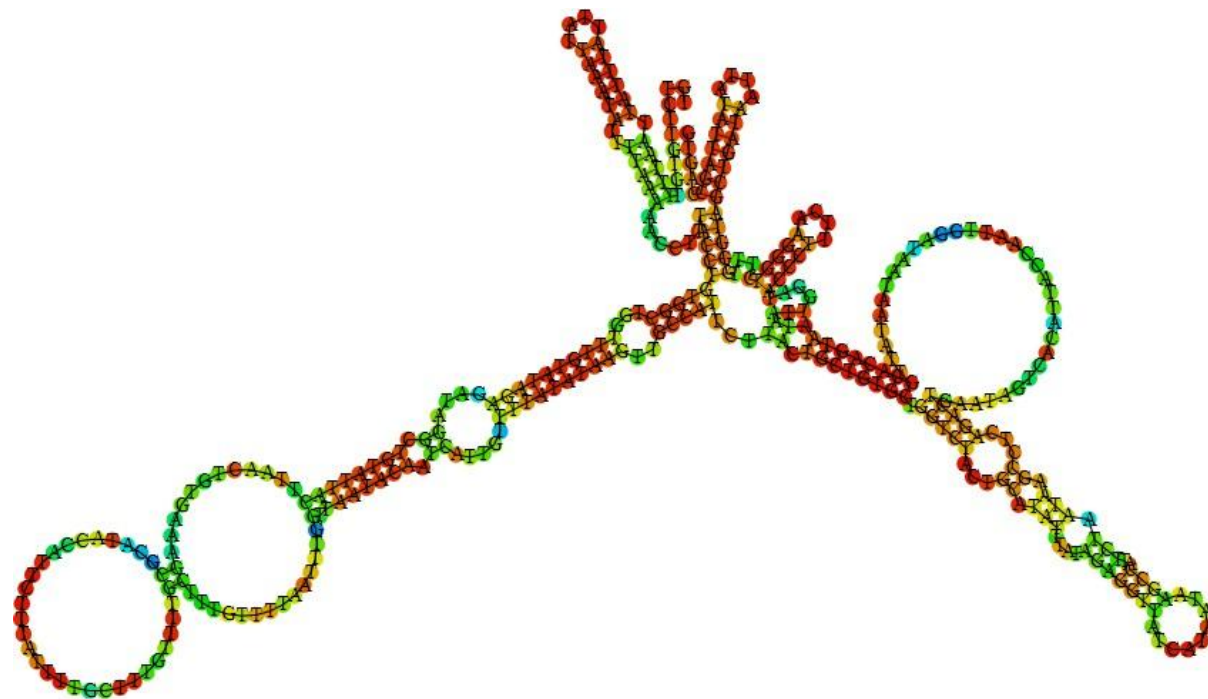


TRINITY_DN126712_c2_g2_i1



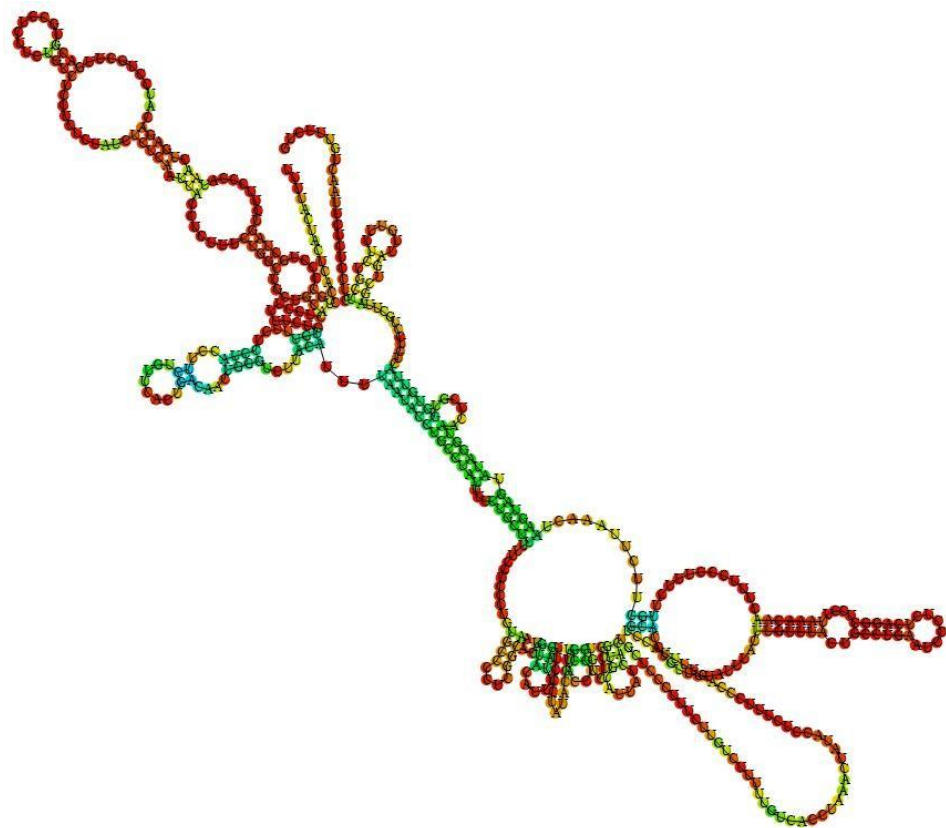
MFE secondary structure

$\Delta G = -63.80$ kcal/mol



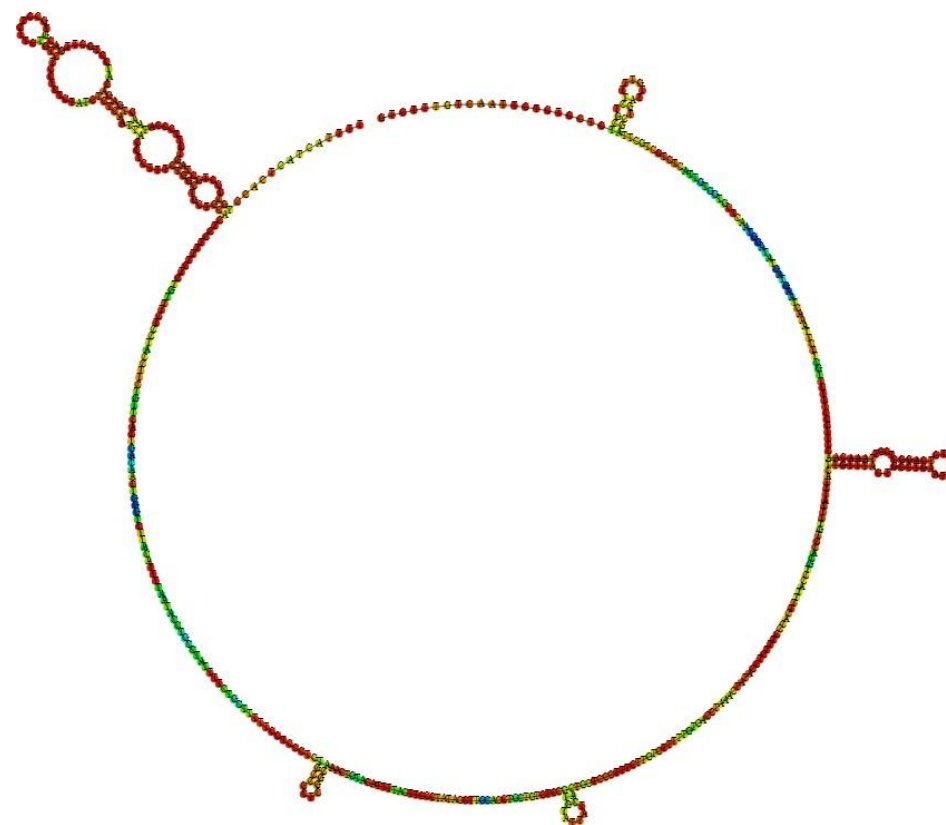
Centroid secondary structure

$\Delta G = -45.47$ kcal/mol



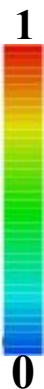
MFE secondary structure

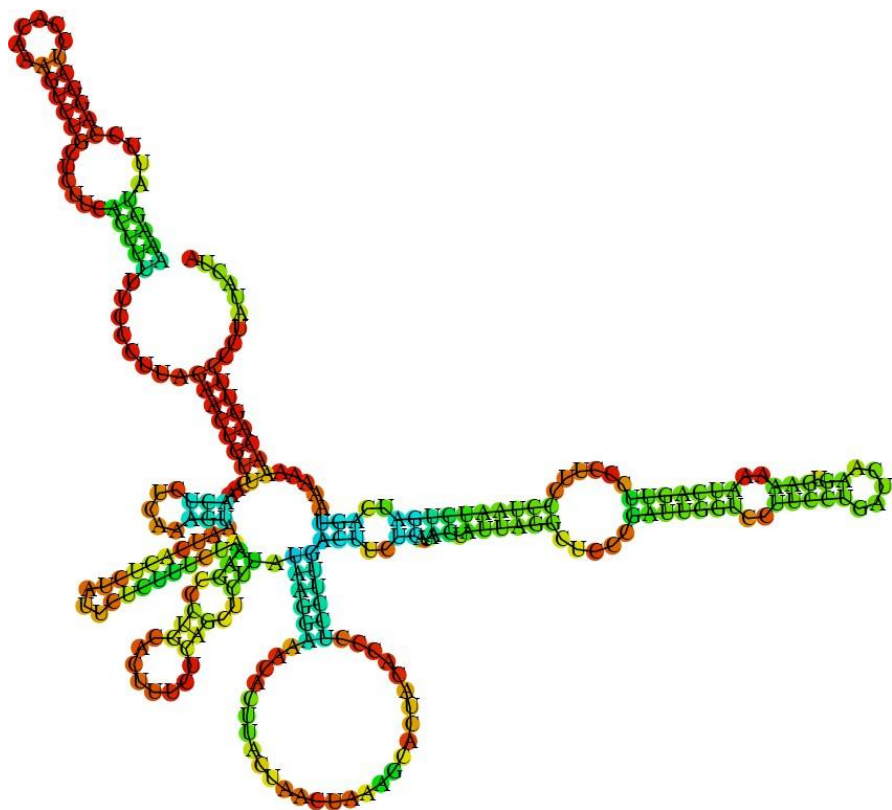
$\Delta G = -59.60$ kcal/mol



Centroid secondary structure

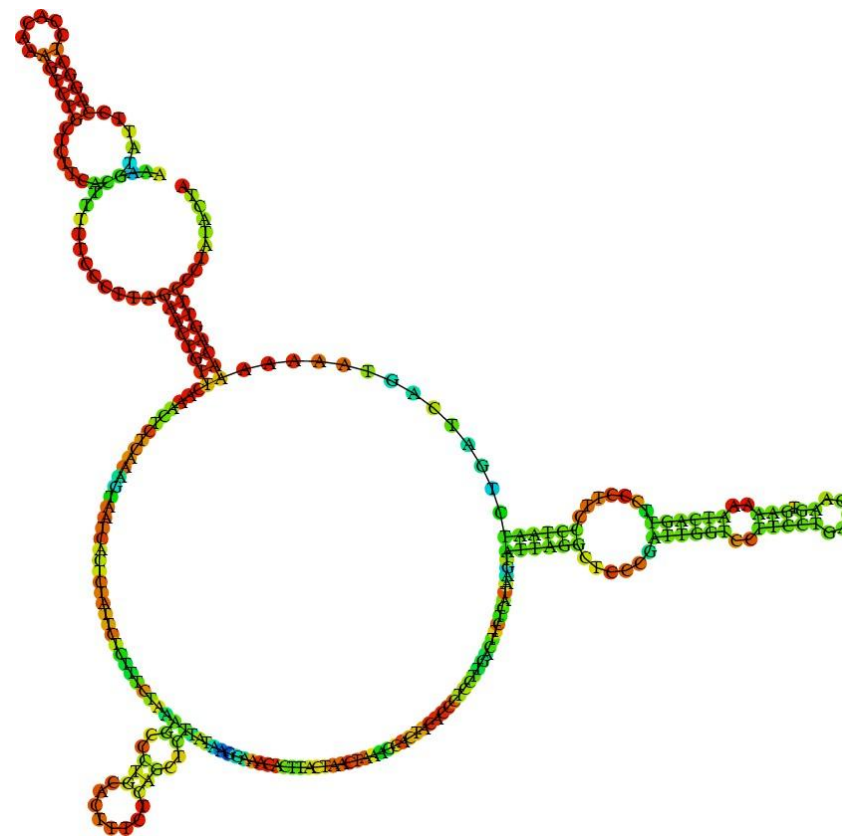
$\Delta G = -37.00$ kcal/mol





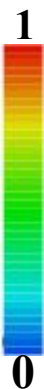
MFE secondary structure

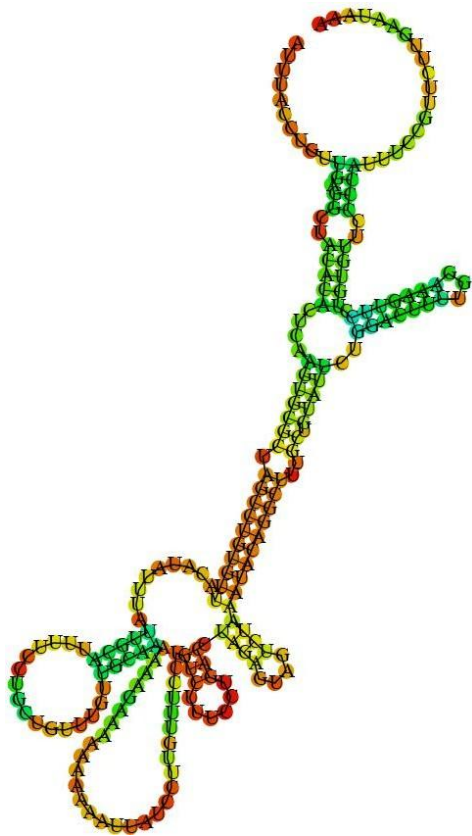
$\Delta G = -32.30$ kcal/mol



Centroid secondary structure

$\Delta G = -20.20$ kcal/mol





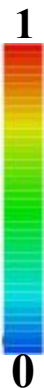
MFE secondary structure

$\Delta G = -35.90$ kcal/mol

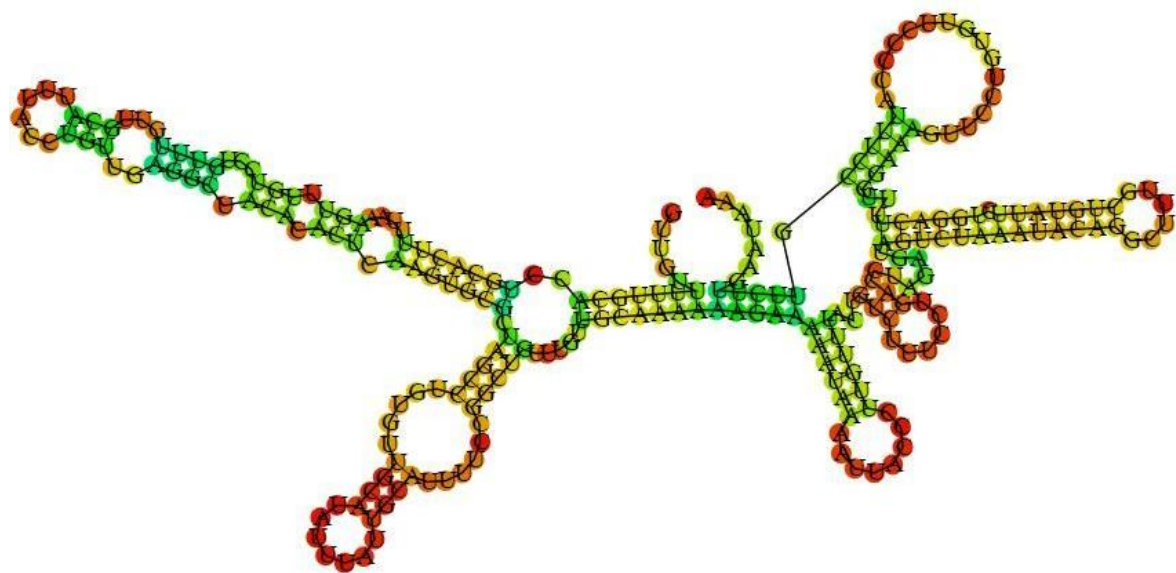


Centroid secondary structure

$\Delta G = -20.20$ kcal/mol

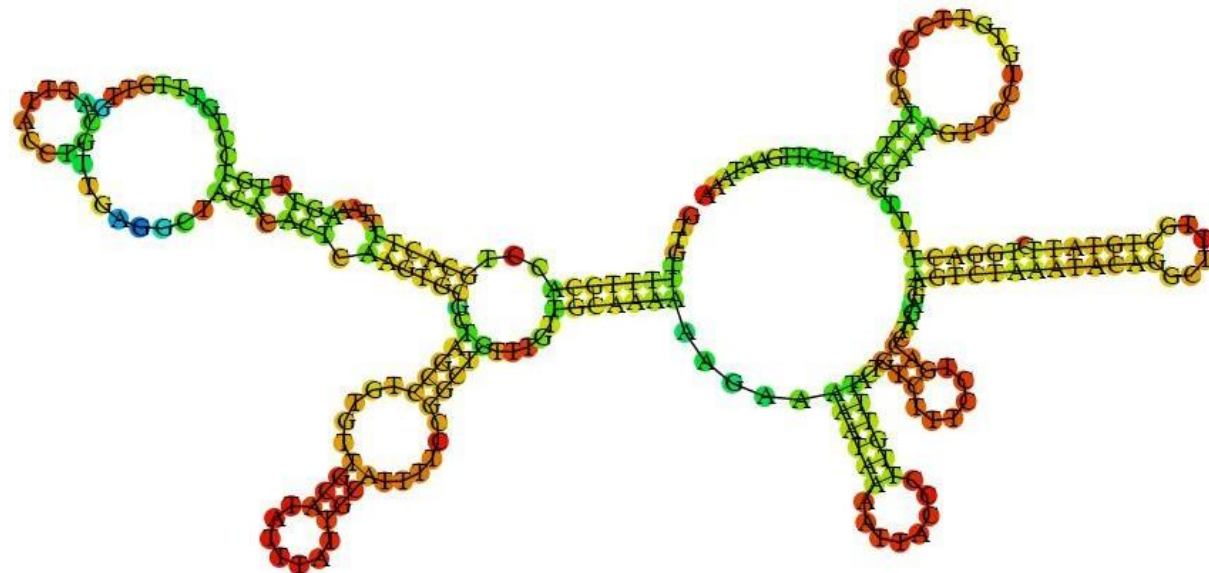


TRINITY_DN127425_c3_g5_i2



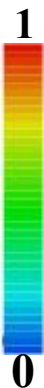
MFE secondary structure

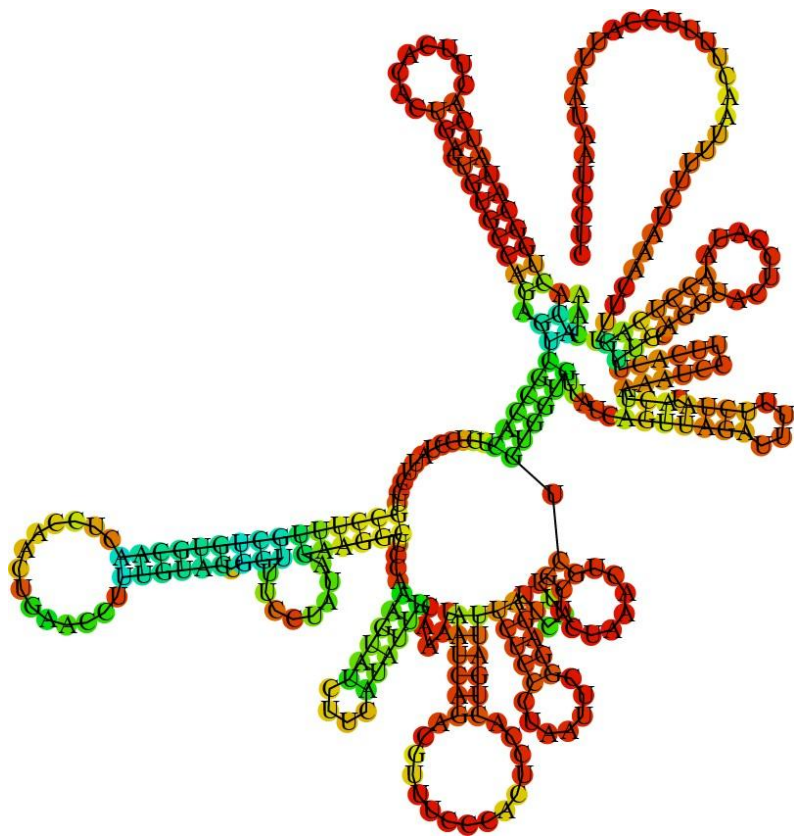
$\Delta G = -46.60$ kcal/mol



Centroid secondary structure

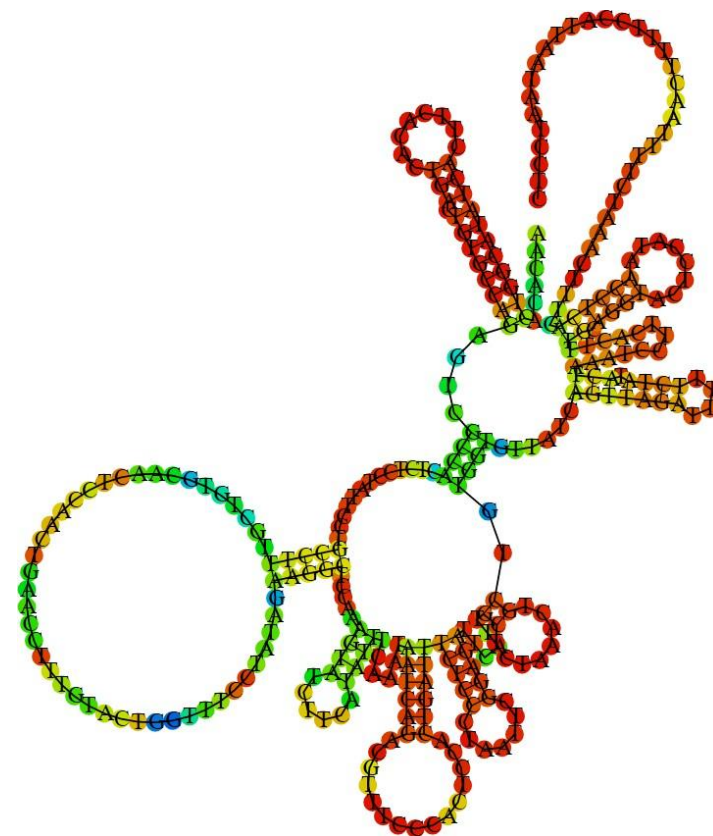
$\Delta G = -37.30$ kcal/mol





MFE secondary structure

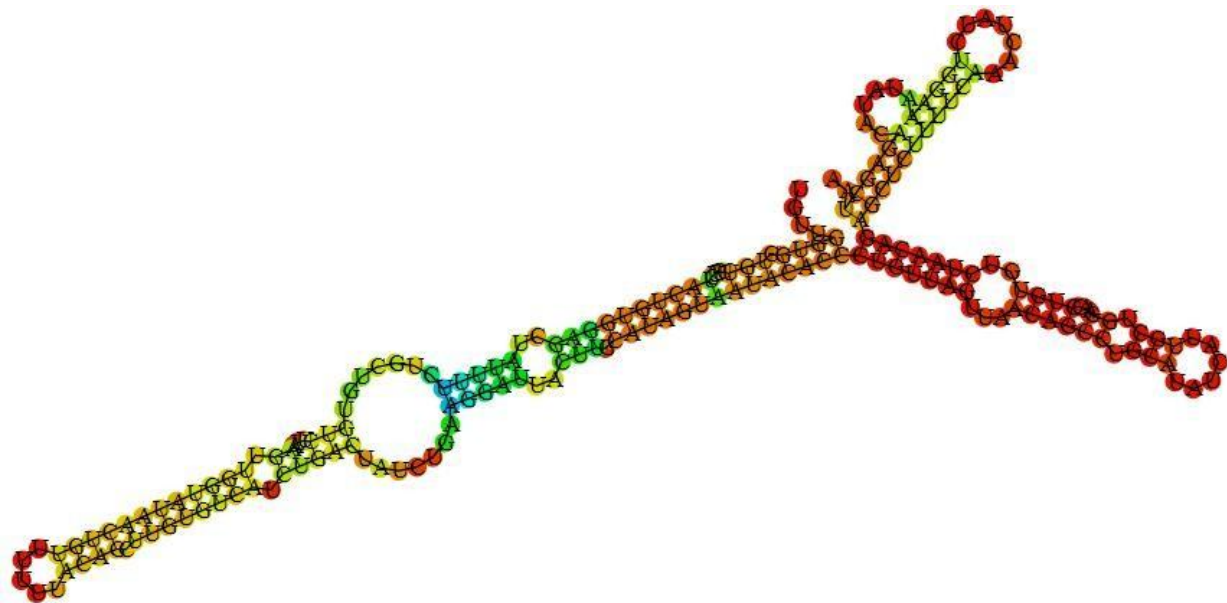
$\Delta G = -50.20$ kcal/mol



Centroid secondary structure

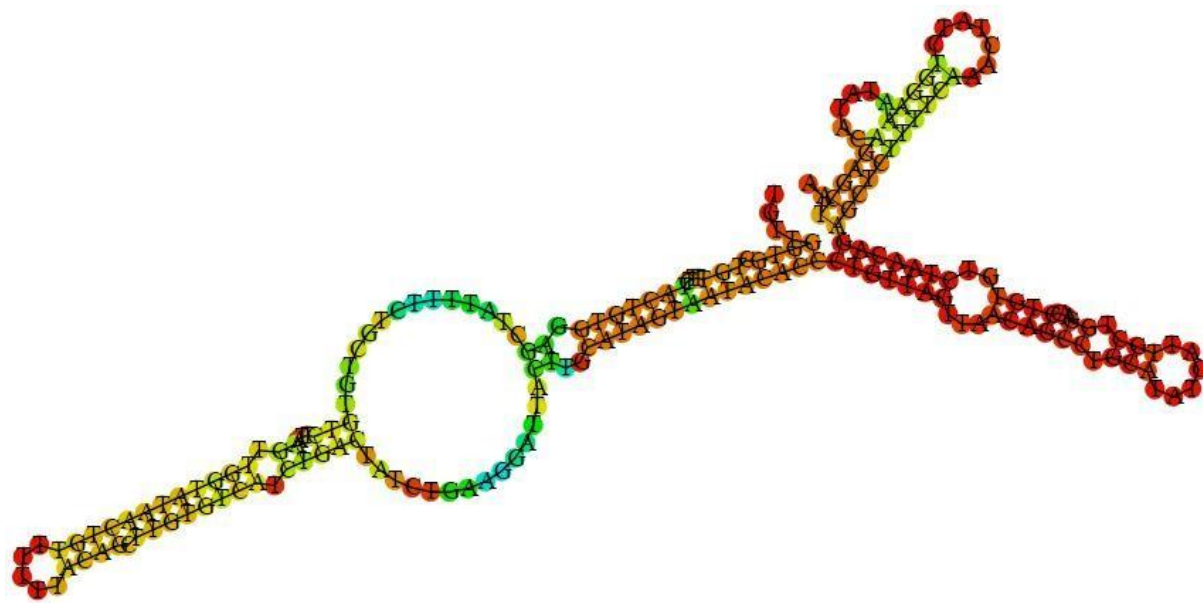
$\Delta G = -35.12$ kcal/mol

TRINITY_DN127747_c5_g1_i1



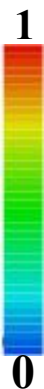
MFE secondary structure

$\Delta G = -49.40$ kcal/mol

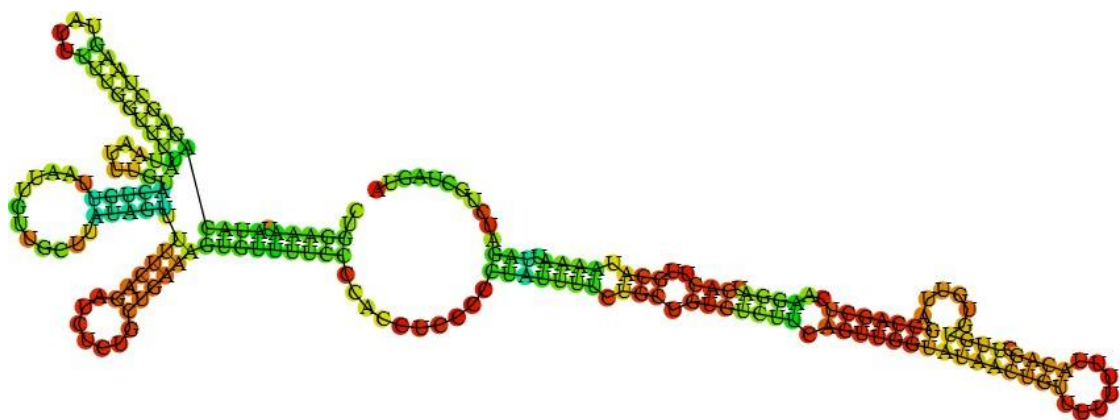


Centroid secondary structure

$\Delta G = -46.00$ kcal/mol

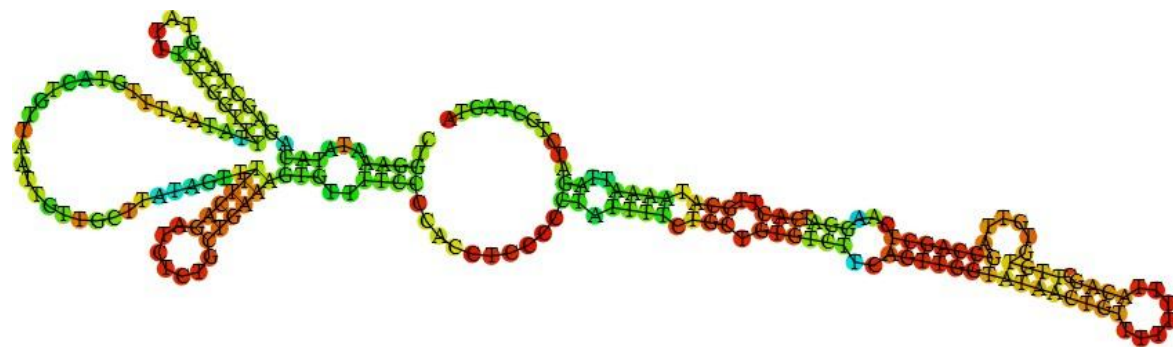


TRINITY_DN127747_c5_g6_i1



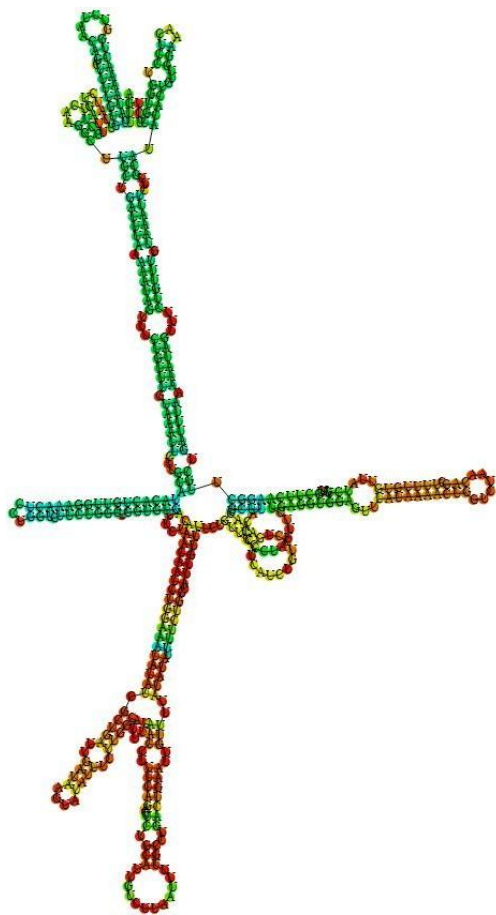
MFE secondary structure

$\Delta G = -38.00$ kcal/mol



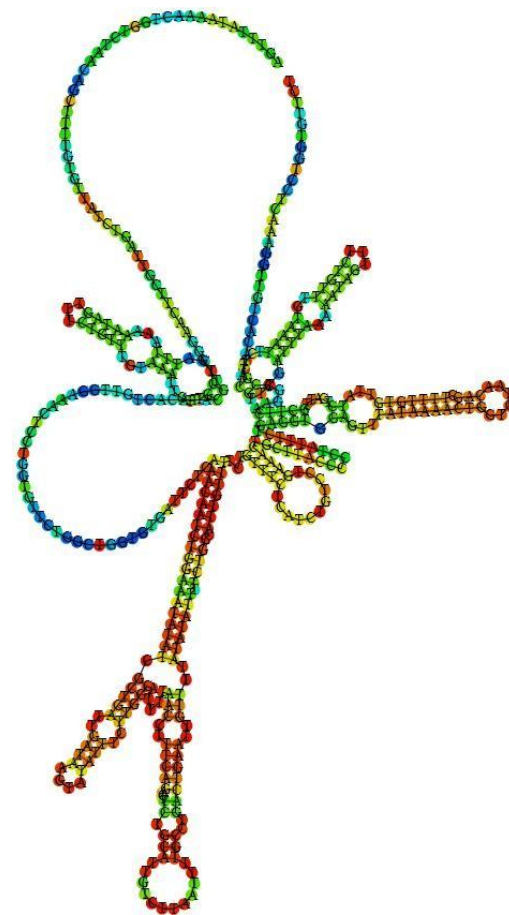
Centroid secondary structure

$\Delta G = -28.90$ kcal/mol



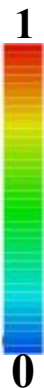
MFE secondary structure

$\Delta G = -81.90$ kcal/mol

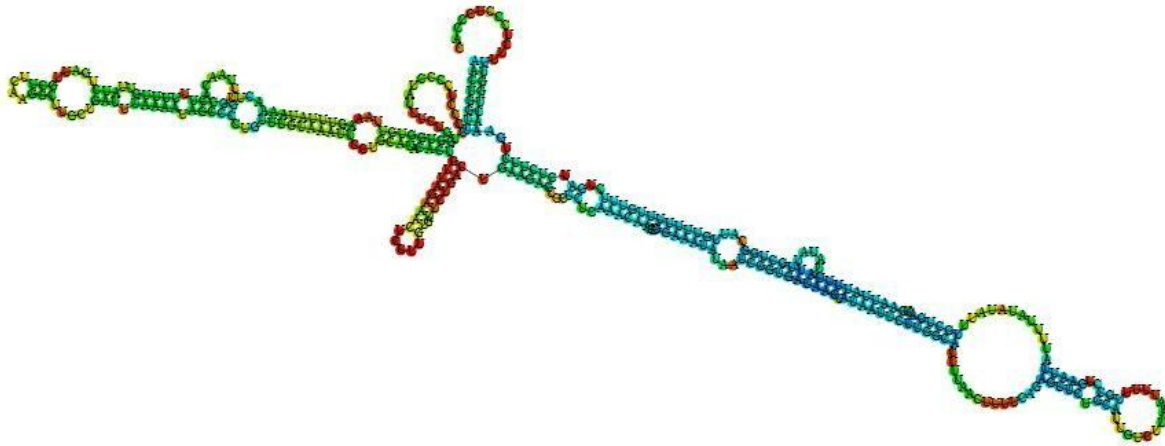


Centroid secondary structure

$\Delta G = -38.20$ kcal/mol

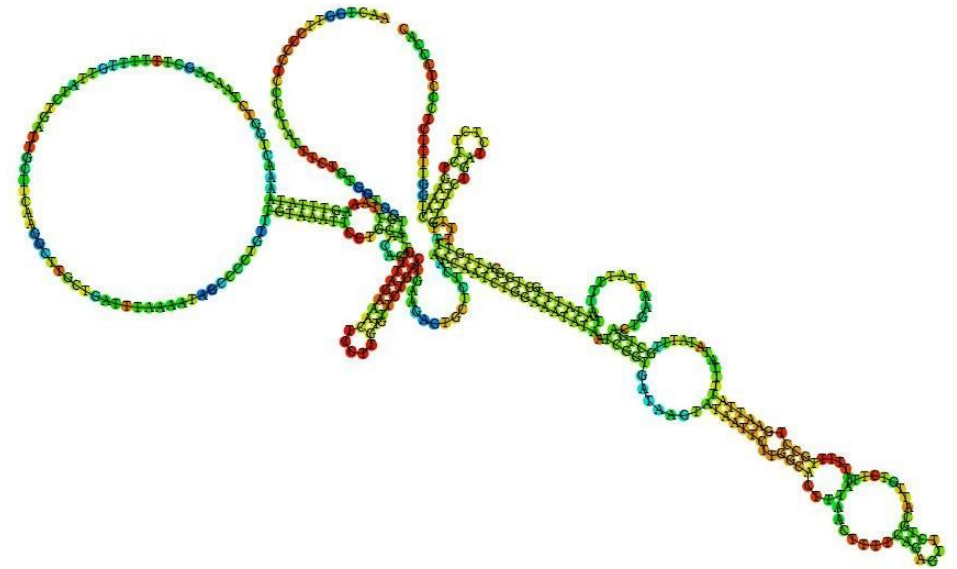


TRINITY_DN127819_c0_g2_i2



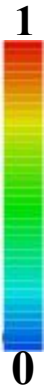
MFE secondary structure

$\Delta G = -62.20$ kcal/mol

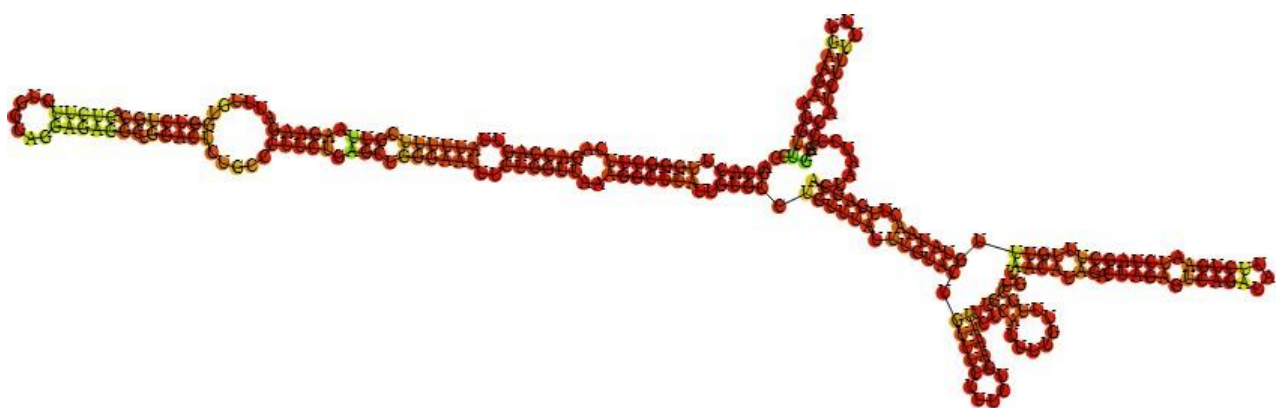


Centroid secondary structure

$\Delta G = -28.42$ kcal/mol

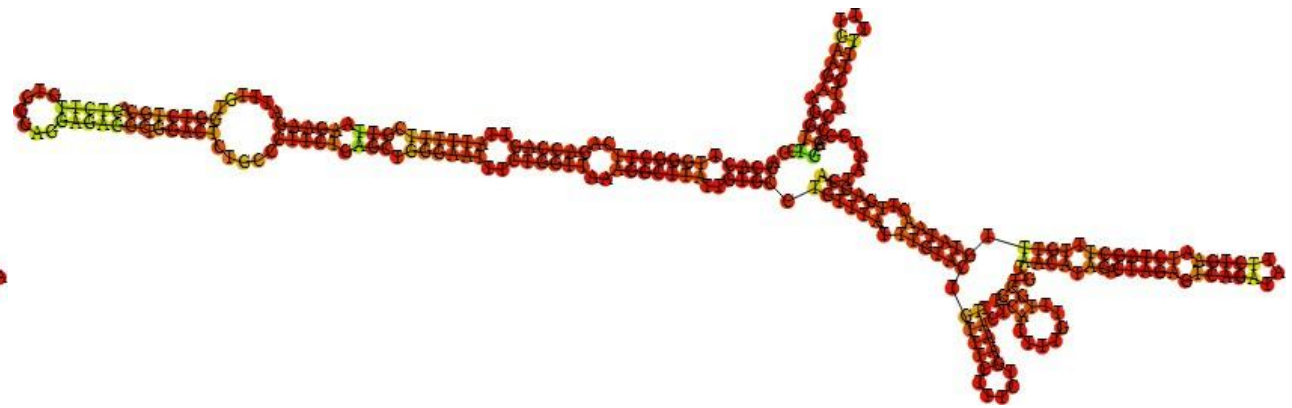


TRINITY_DN128349_c6_g5_i1



MFE secondary structure

$\Delta G = -76.70$ kcal/mol

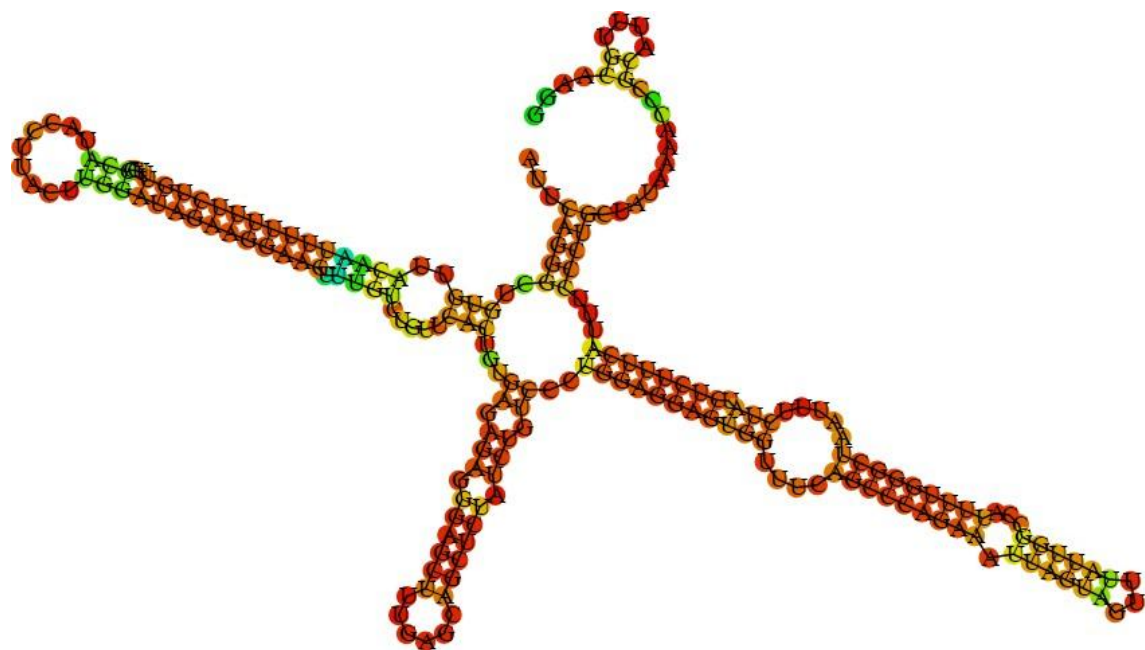


Centroid secondary structure

$\Delta G = -76.70$ kcal/mol

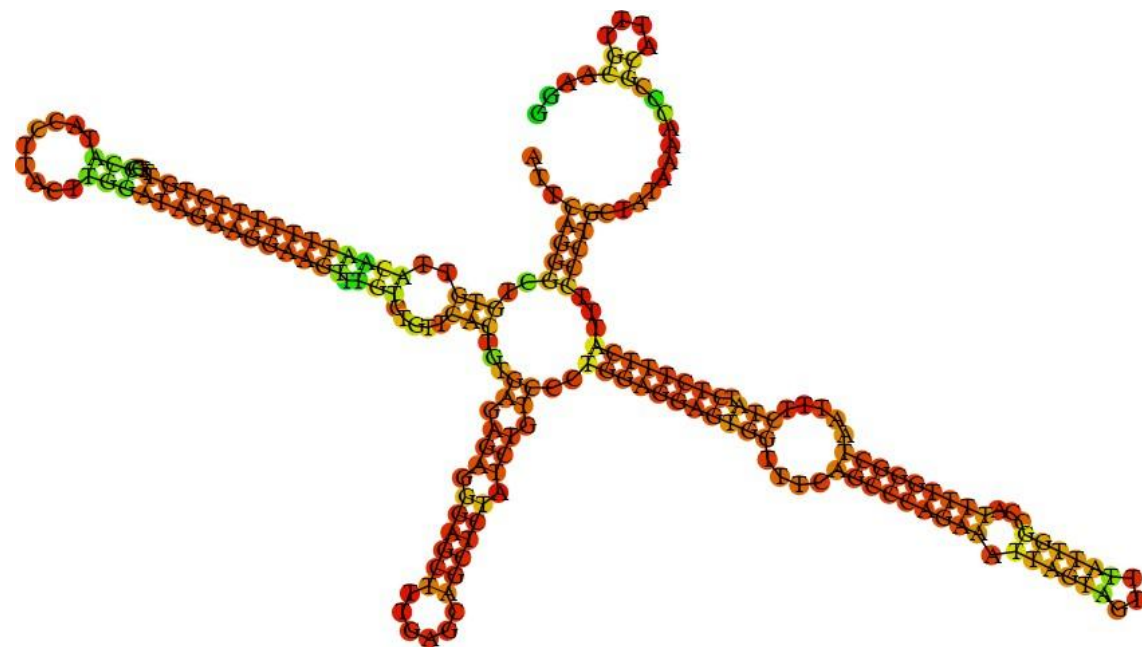


TRINITY_DN128368_c9_g5_i1



MFE secondary structure

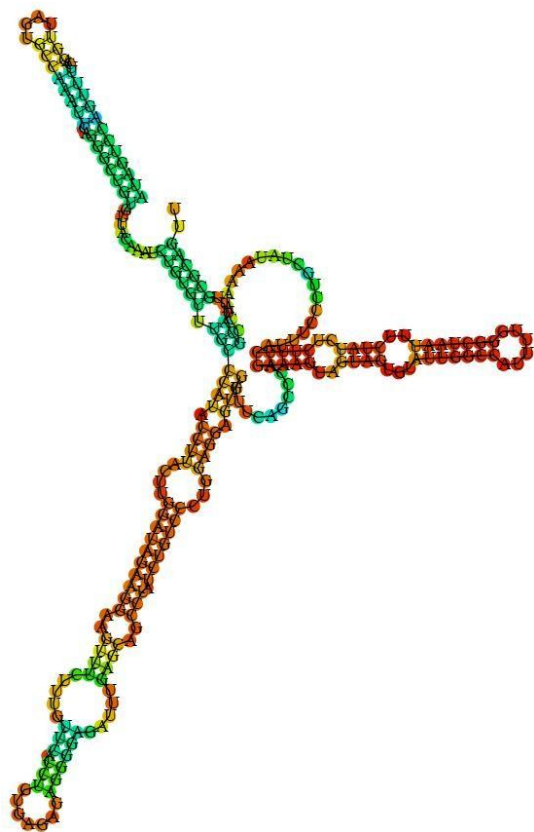
$\Delta G = -55.30$ kcal/mol



Centroid secondary structure

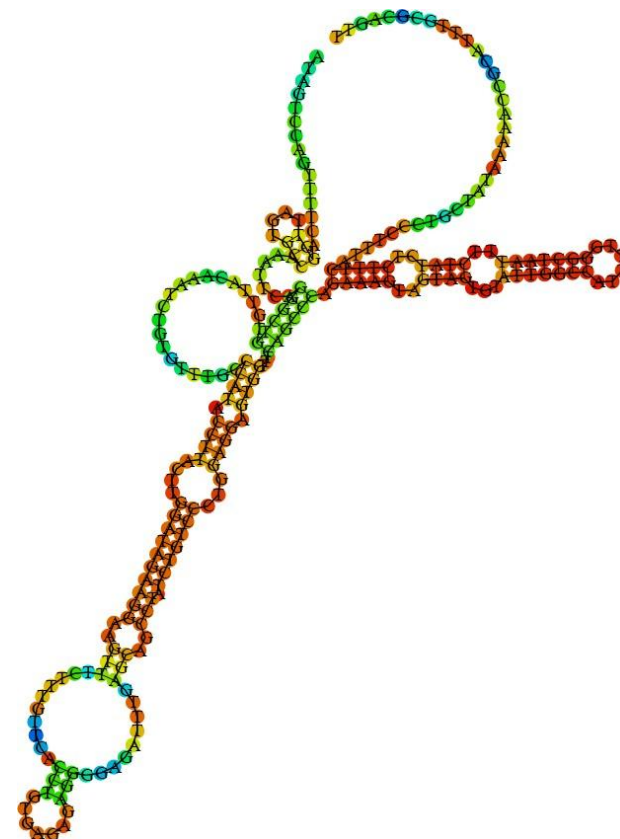
$\Delta G = -55.30$ kcal/mol

TRINITY_DN128368_c9_g5_i3



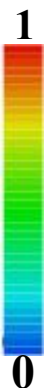
MFE secondary structure

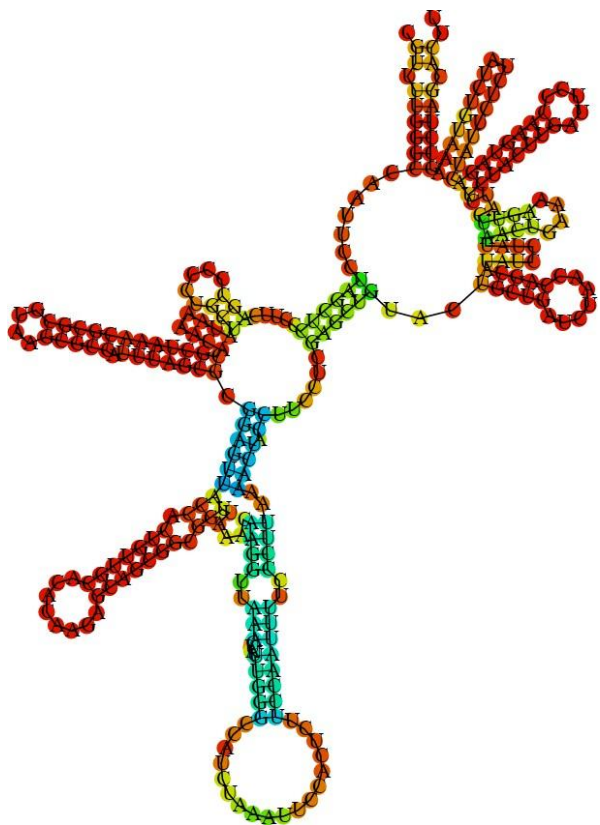
$\Delta G = -63.80$ kcal/mol



Centroid secondary structure

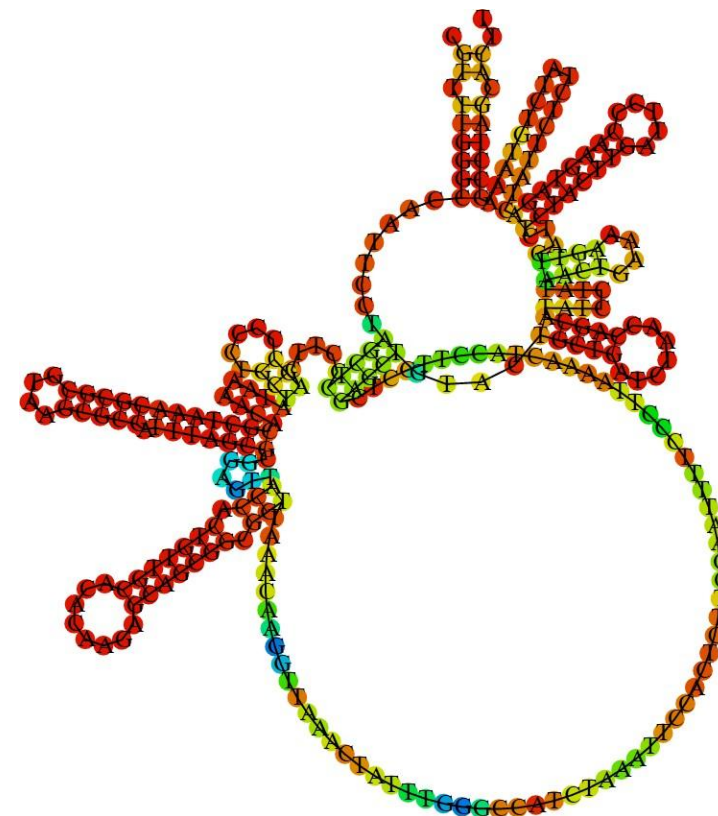
$\Delta G = -53.60$ kcal/mol





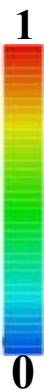
MFE secondary structure

$\Delta G = -70.50$ kcal/mol

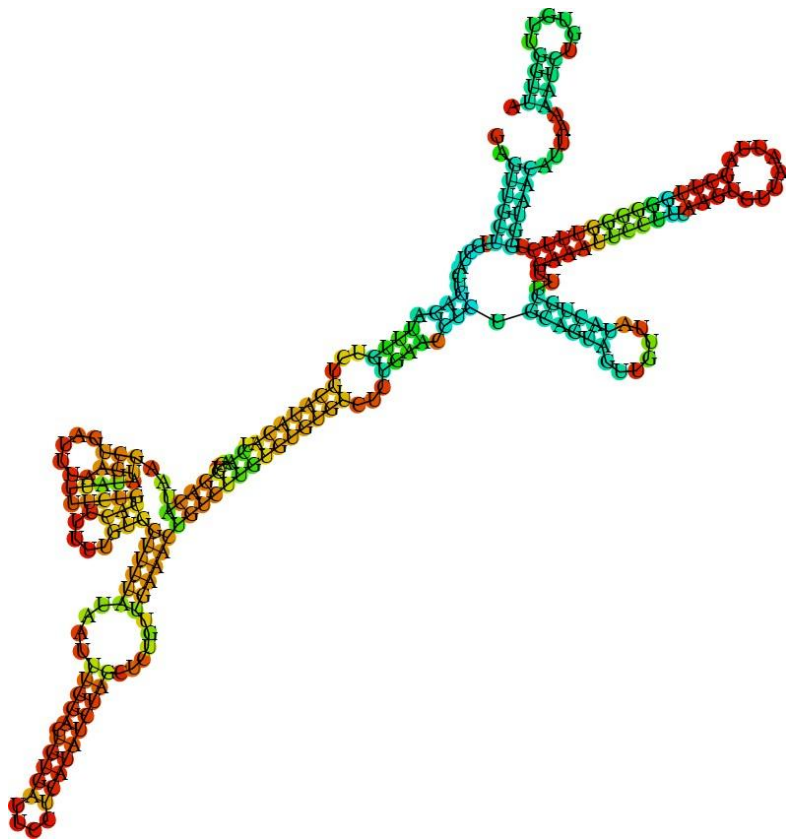


Centroid secondary structure

$\Delta G = -63.40$ kcal/mol

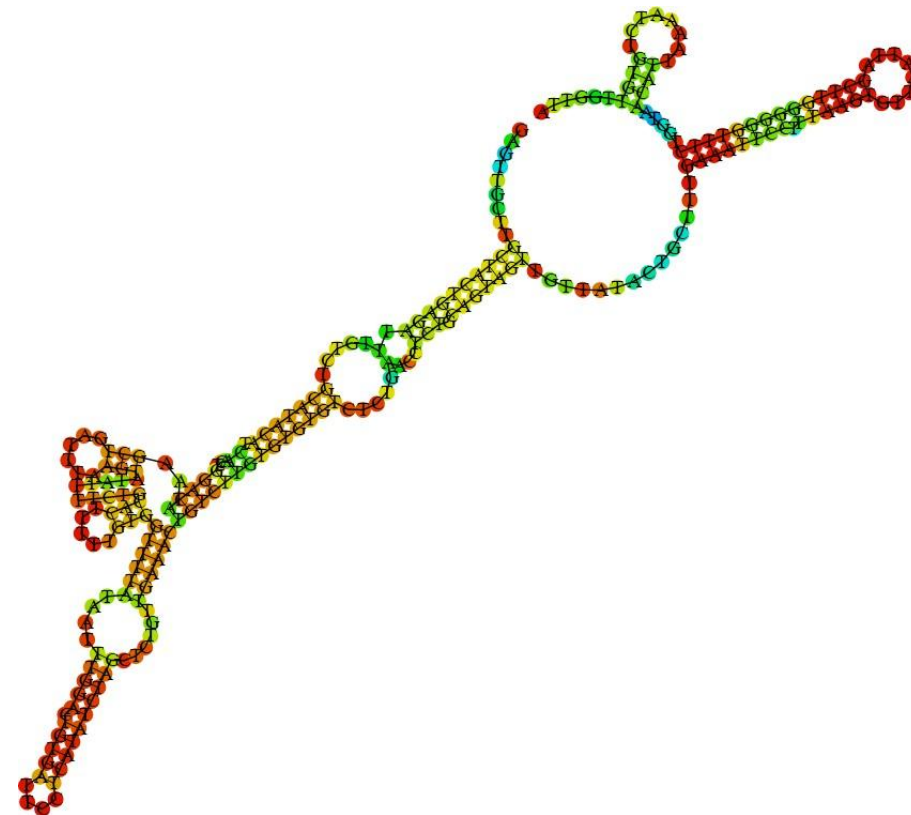


TRINITY_DN128998_c2_g6_i1



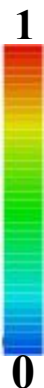
MFE secondary structure

$\Delta G = -50.90$ kcal/mol

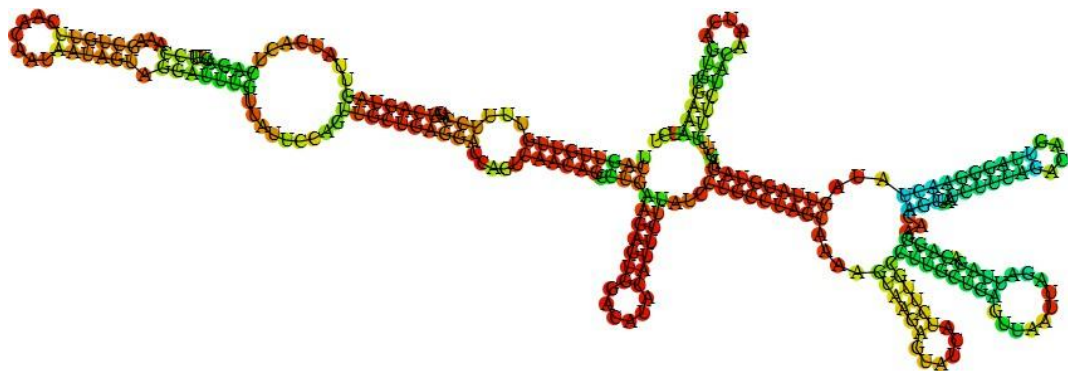


Centroid secondary structure

$\Delta G = -44.00$ kcal/mol

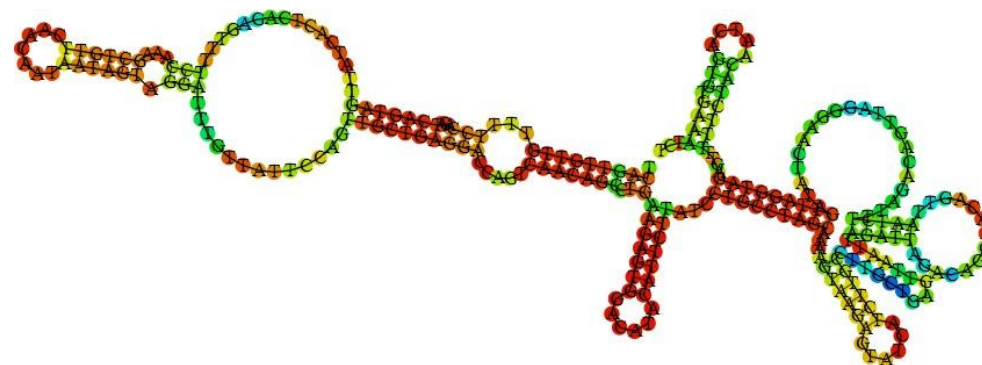


TRINITY_DN21685_c0_g1_i1



MFE secondary structure

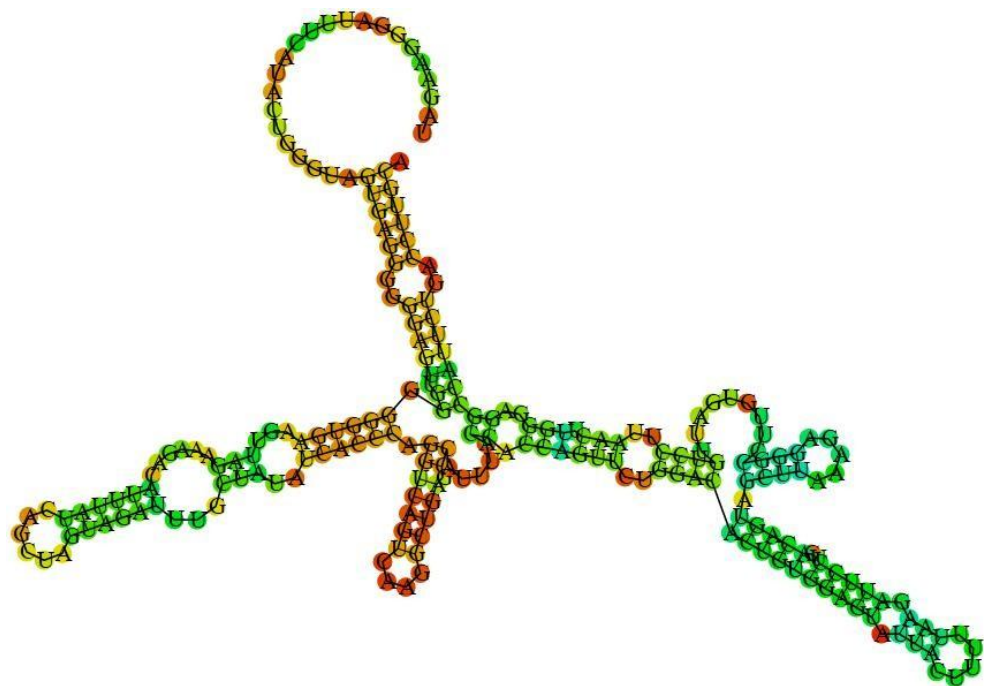
$\Delta G = -56.60$ kcal/mol



Centroid secondary structure

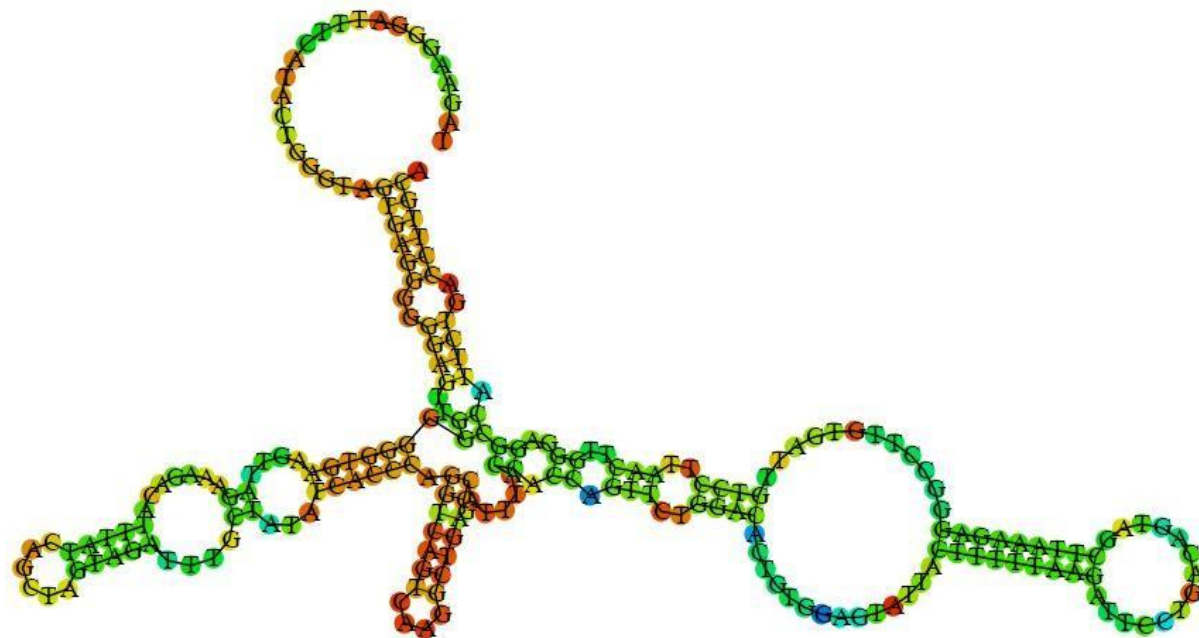
$\Delta G = -46.10$ kcal/mol

TRINITY_DN24628_c0_g1_i1



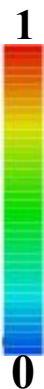
MFE secondary structure

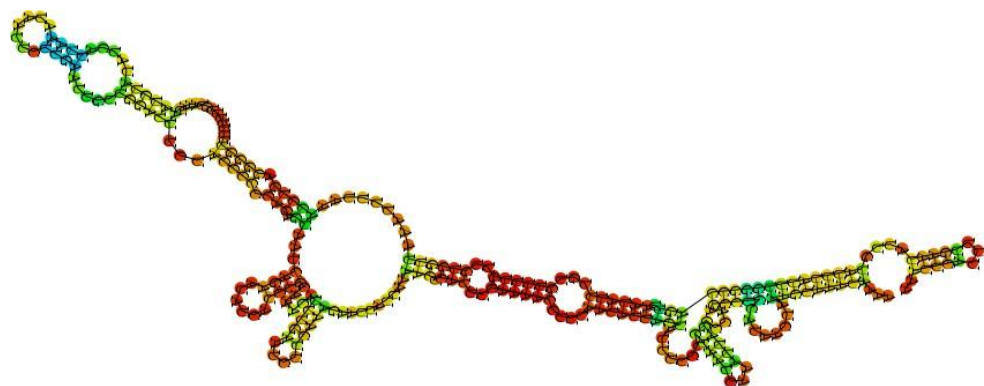
$\Delta G = -54.00$ kcal/mol



Centroid secondary structure

$\Delta G = -41.60$ kcal/mol





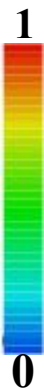
MFE secondary structure

$\Delta G = -38.60$ kcal/mol

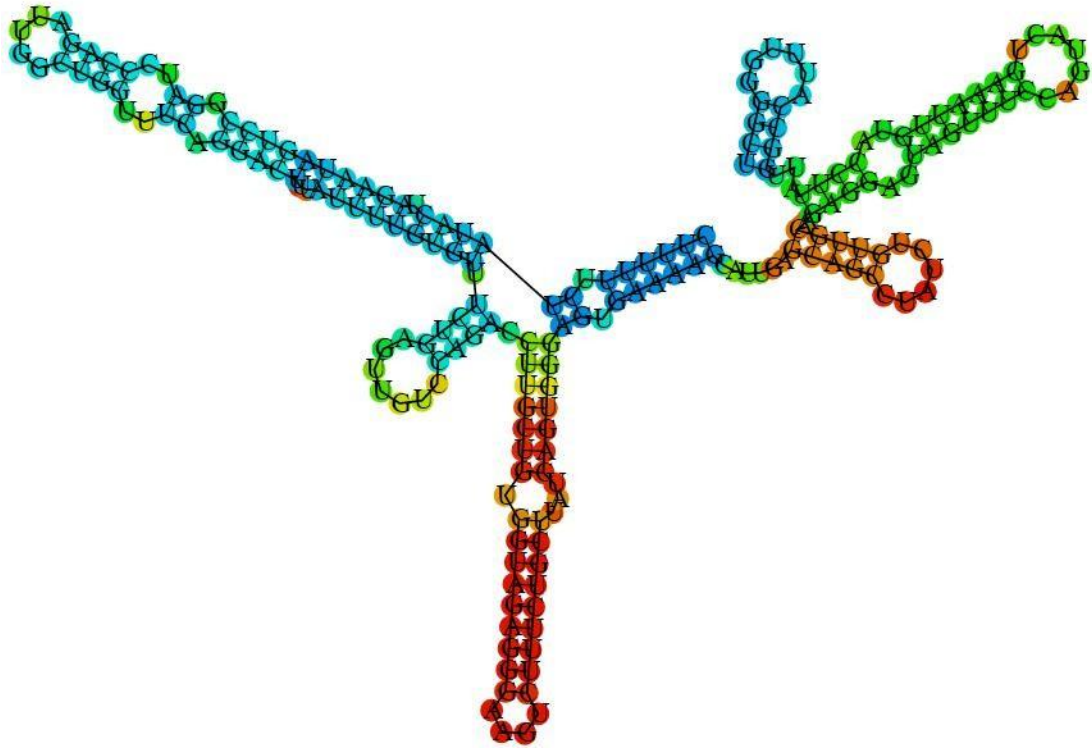


Centroid secondary structure

$\Delta G = -31.30$ kcal/mol



TRINITY_DN59921_c0_g1_i1



MFE secondary structure

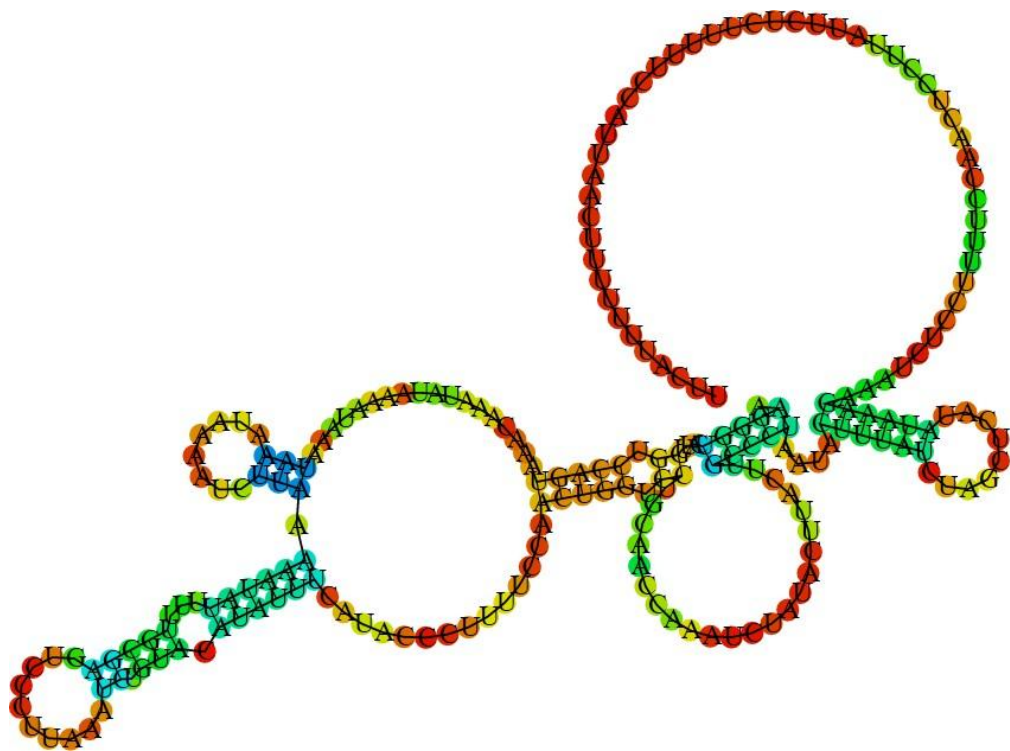
$\Delta G = -55.80$ kcal/mol



Centroid secondary structure

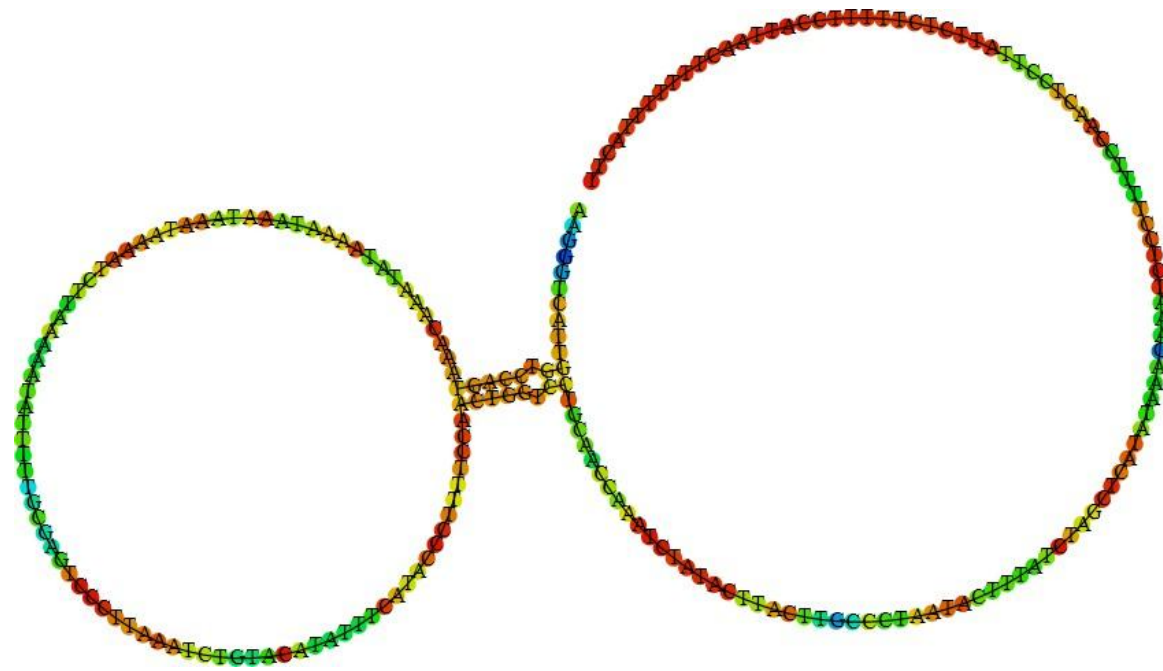
$\Delta G = -37.60$ kcal/mol

TRINITY_DN64687_c0_g1_i1



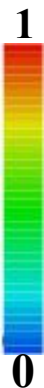
MFE secondary structure

$\Delta G = -12.00$ kcal/mol

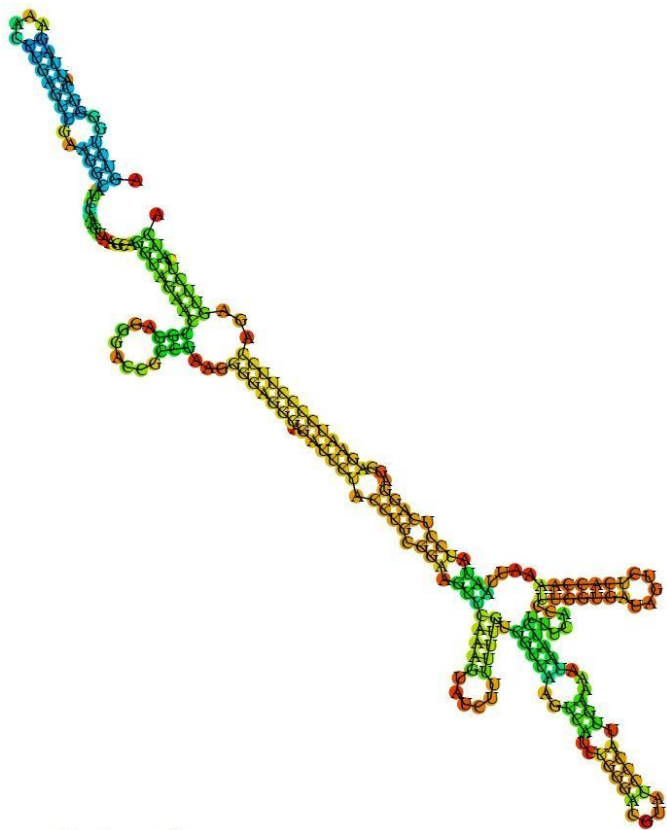


Centroid secondary structure

$\Delta G = -6.49$ kcal/mol



TRINITY_DN68223_c0_g1_i1



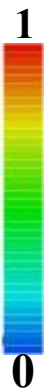
MFE secondary structure

$\Delta G = -59.70$ kcal/mol

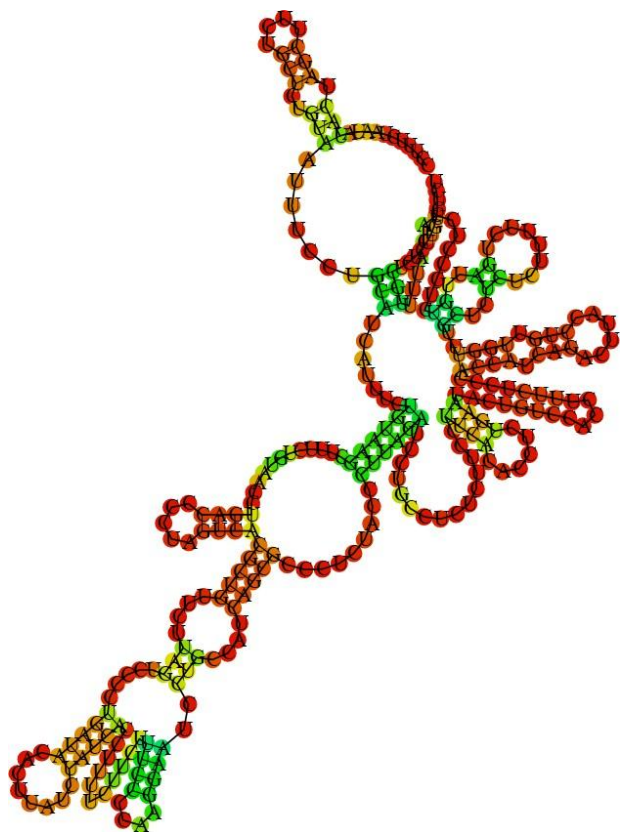


Centroid secondary structure

$\Delta G = -48.30$ kcal/mol

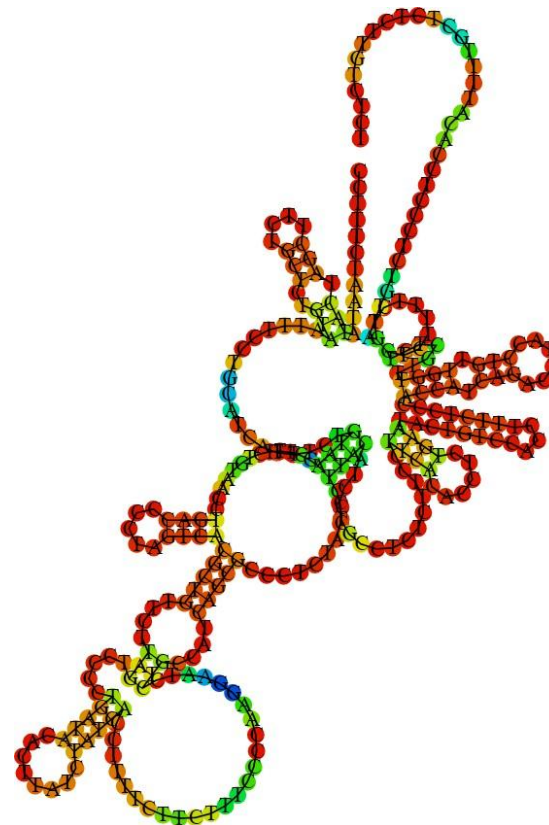


TRINITY_DN76107_c0_g2_i1



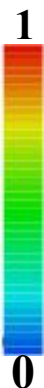
MFE secondary structure

$\Delta G = -29.50$ kcal/mol

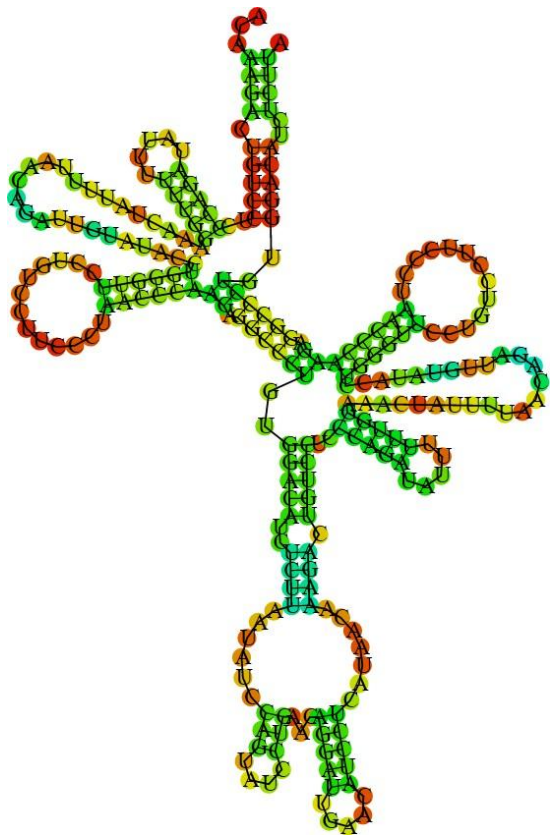


Centroid secondary structure

$\Delta G = -17.40$ kcal/mol

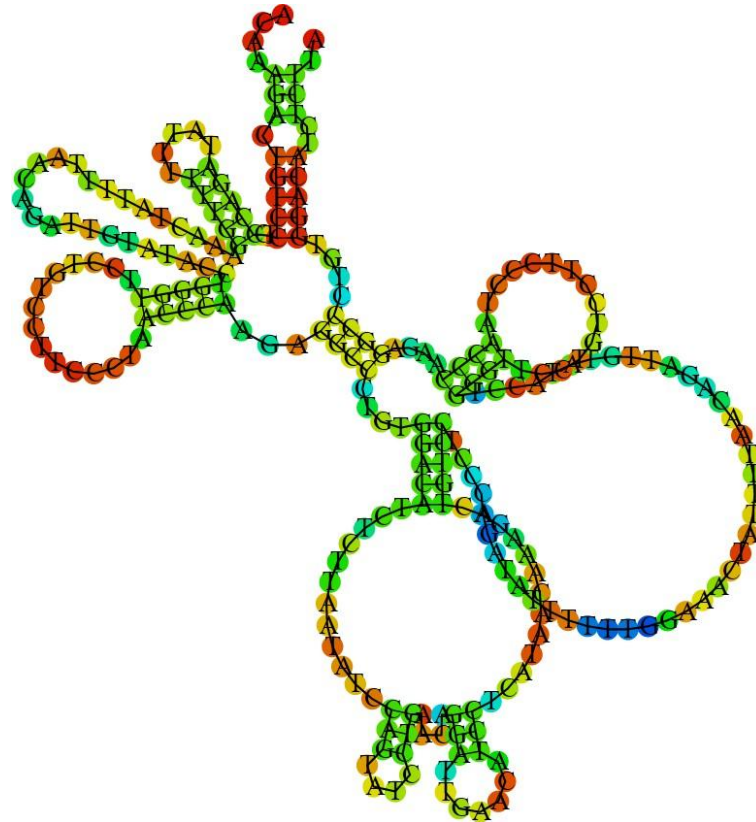


TRINITY_DN90964_c0_g1_i2



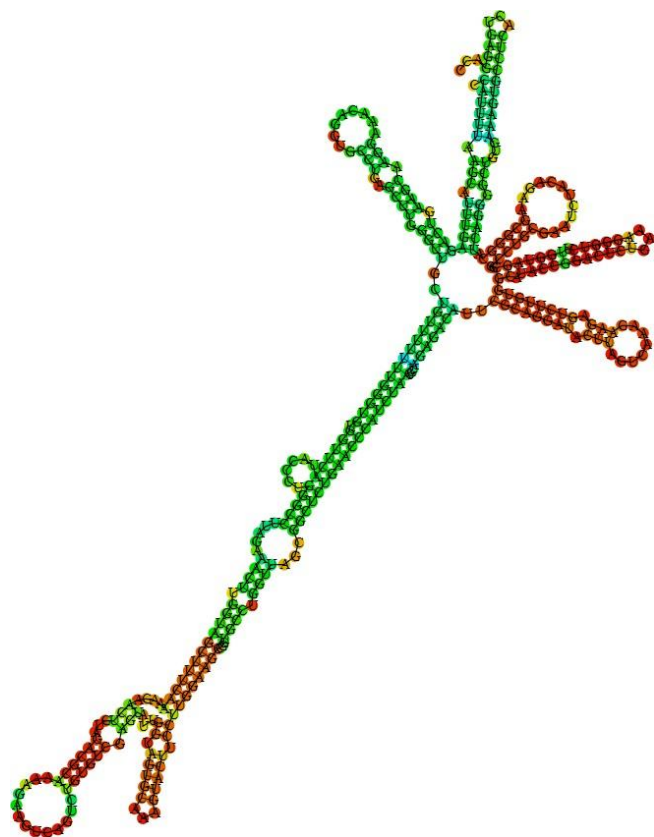
MFE secondary structure

$\Delta G = -47.30$ kcal/mol



Centroid secondary structure

$\Delta G = -31.30$ kcal/mol



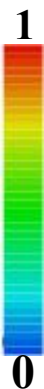
MFE secondary structure

$\Delta G = -109.10$ kcal/mol



Centroid secondary structure

$\Delta G = -64.70$ kcal/mol



Apêndice B - Tabela dos resultados para os 126 genes linhagem-específicos nas análises de: Expressão Diferencial, Pfam e os respectivos E-values associados e o CEMiTool com os respectivos módulos associados a cada transcrito.

Contigs (lineage-specific coding transcript)	1 dpa		14 dpa		21 dpa		CEMiTool Module	Pfam	
	Fold Change	Total Gene Reads	Fold Change	Total Gene Reads	Fold Change	Total Gene Reads		Protein Domain	E-value
TRINITY_DN101228_c0_g1_i1	3,43	60	4,20	68	5,48	135	0	GIY-YIG catalytic domain	8,20E-05
TRINITY_DN101265_c0_g2_i1	29,62	74	40,53	95	27,67	113	0	GIY-YIG catalytic domain	1,10E-05
TRINITY_DN102357_c0_g1_i1	5,30	45	6,01	47	7,28	83	0	Reverse transcriptase (RNA-dependent DNA polymerase)	1,40E-07
TRINITY_DN105290_c0_g6_i1	3,69	169	5,09	211	3,39	215	0	GIY-YIG catalytic domain	8,00E-06
TRINITY_DN105890_c0_g1_i1	16,30	51	15,91	47	11,32	54	0	GIY-YIG catalytic domain	5,30E-05
TRINITY_DN105952_c0_g1_i1	10,17	35	11,71	38	8,17	46	0	GIY-YIG catalytic domain	9,30E-06
TRINITY_DN106345_c0_g1_i1	14,18	88	13,02	78	13,88	112	0	GIY-YIG catalytic domain	3,00E-07
TRINITY_DN108468_c0_g1_i1	6,68	86	7,36	89	4,08	86	0	GIY-YIG catalytic domain	2,10E-05
TRINITY_DN111095_c0_g1_i1	3,35	4429	3,93	4813	3,98	7077	M1	Phage integrase, N-terminal SAM-like domain	5,30E-07
TRINITY_DN112111_c0_g1_i1	3,88	183	4,82	209	3,19	221	0	Endonuclease-reverse transcriptase	3,90E-08
TRINITY_DN112564_c0_g2_i2	9,89	113	6,68	76	7,27	129	0	Reverse transcriptase (RNA-dependent DNA polymerase)	9,80E-10
TRINITY_DN112860_c0_g1_i1	8,59	83	7,12	67	7,54	101	0	zinc-binding in reverse transcriptase	5,20E-08
TRINITY_DN113318_c2_g1_i7	13,69	64	12,80	57	11,01	83	0	Reverse transcriptase (RNA-dependent DNA polymerase)	9,10E-08
TRINITY_DN113521_c0_g1_i1	59,30	176	49,91	141	47,23	215	M1	Endonuclease-reverse transcriptase	2,20E-09
TRINITY_DN113970_c0_g1_i1	11,81	43	10,59	37	11,08	63	0	Endonuclease-reverse transcriptase	5,20E-10
TRINITY_DN115225_c1_g1_i1	10,07	37	6,84	25	4,89	31	0	Endonuclease-reverse transcriptase	1,50E-08
TRINITY_DN115225_c1_g1_i2	16,50	17	35,75	34	19,24	33	0	Endonuclease-reverse transcriptase	3,60E-08
TRINITY_DN115761_c2_g10_i2	9,19	92	10,22	96	9,98	149	0	Endonuclease-reverse transcriptase	2,40E-09
TRINITY_DN115823_c0_g2_i1	22,57	23	32,48	31	26,79	39	0	Retroviral aspartyl protease	4,20E-10
TRINITY_DN115910_c4_g3_i2	14,34	114	9,43	74	11,17	137	0	Endonuclease-reverse transcriptase	8,30E-08
TRINITY_DN115910_c4_g4_i1	11,62	70	14,61	82	13,28	119	0	Endonuclease-reverse transcriptase	1,20E-08
TRINITY_DN115999_c0_g1_i2	11,00	271	10,96	258	10,24	384	M1	Endonuclease-reverse transcriptase	2,00E-07

TRINITY_DN116279_c0_g2_i1	24,12	98	25,56	98	22,07	135	M1	Phage integrase, N-terminal SAM-like domain	5,00E-06
TRINITY_DN116879_c1_g1_i8	9,19	162	11,70	191	8,53	234	M1	Endonuclease-reverse transcriptase	2,30E-09
TRINITY_DN116879_c1_g2_i1	9,62	182	12,67	224	9,17	264	0	Endonuclease-reverse transcriptase	3,40E-08
TRINITY_DN117256_c0_g2_i2	6,07	71	8,48	90	8,08	129	0	Endonuclease-reverse transcriptase	1,20E-08
TRINITY_DN117327_c0_g1_i2	26,95	115	27,71	112	15,25	96	M1	Endonuclease-reverse transcriptase	4,60E-07
TRINITY_DN117839_c0_g2_i2	13,65	95	19,04	124	12,83	138	0	Endonuclease-reverse transcriptase	2,00E-09
TRINITY_DN118390_c1_g4_i2	6,66	355	6,91	349	7,05	519	M1	Endonuclease-reverse transcriptase	1,10E-07
TRINITY_DN118755_c0_g4_i1	10,05	80	9,71	74	11,27	123	0	Endonuclease-reverse transcriptase	3,70E-08
TRINITY_DN119241_c12_g1_i1	6,27	1549	6,47	1519	6,69	2376	M1	Lamina-associated polypeptide 2 alpha	4,00E-07
TRINITY_DN119296_c0_g1_i1	8,71	65	13,37	92	13,24	139	0	Endonuclease-reverse transcriptase	4,10E-09
TRINITY_DN119604_c0_g6_i1	6,07	1003	6,45	1009	6,37	1539	M1	Phage integrase, N-terminal SAM-like domain	7,00E-07
TRINITY_DN120065_c1_g1_i2	8,13	111	6,70	89	7,28	146	0	Endonuclease-reverse transcriptase	2,30E-10
TRINITY_DN120564_c5_g4_i1	3,87	940	3,41	813	4,00	1316	M1	Endonuclease-reverse transcriptase	1,00E-09
TRINITY_DN120790_c0_g3_i2	39,45	79	40,46	77	28,28	93	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	5,80E-08
TRINITY_DN120911_c4_g2_i1	7,20	66	6,03	54	7,15	94	0	Endonuclease-reverse transcriptase	1,80E-08
TRINITY_DN120911_c6_g4_i2	8,51	28	5,43	18	10,12	43	0	Endonuclease-reverse transcriptase	5,20E-09
TRINITY_DN120911_c6_g4_i5	9,87	47	7,48	35	3,22	29	0	Endonuclease-reverse transcriptase	6,50E-09
TRINITY_DN120961_c0_g6_i1	3,51	167	5,65	236	4,45	281	M1	Phage integrase, N-terminal SAM-like domain	5,50E-06
TRINITY_DN121120_c4_g1_i1	6,34	102	6,18	95	5,54	143	0	Endonuclease-reverse transcriptase	5,70E-09
TRINITY_DN121391_c0_g2_i1	14,07	118	9,75	80	9,18	114	0	Endonuclease-reverse transcriptase	2,60E-09
TRINITY_DN121669_c2_g6_i1	1,94	566	2,46	644	1,91	756	0	Reverse transcriptase (RNA-dependent DNA polymerase)	5,10E-08
TRINITY_DN121847_c4_g1_i2	5,06	135	5,19	132	7,29	271	0	Endonuclease-reverse transcriptase	2,20E-09
TRINITY_DN121855_c3_g8_i1	5,34	804	5,43	779	5,12	1093	M1	Endonuclease-reverse transcriptase	6,20E-07

TRINITY_DN121873_c0_g3_i2	10,03	478	8,64	398	9,41	660	M1	Lamina-associated polypeptide 2 alpha	5,40E-06
TRINITY_DN121983_c0_g3_i1	6,81	804	7,78	864	7,55	1263	M1	Lamina-associated polypeptide 2 alpha	4,20E-06
TRINITY_DN122079_c0_g2_i1	5,61	124	1,23	40	-1,65	35	0	GIY-YIG catalytic domain	3,40E-06
TRINITY_DN122239_c3_g3_i1	7,13	237	5,51	180	5,51	268	0	Myb/SANT-like DNA-binding domain	1,10E-07
TRINITY_DN122317_c2_g3_i2	15,07	63	13,29	53	14,94	88	0	Endonuclease-reverse transcriptase	4,50E-08
TRINITY_DN122317_c2_g4_i1	6,03	84	6,60	87	7,71	144	0	Endonuclease-reverse transcriptase	1,60E-07
TRINITY_DN122365_c4_g8_i2	13,09	265	12,20	235	12,53	366	M1	Endonuclease-reverse transcriptase	1,40E-10
TRINITY_DN122412_c1_g3_i1	8,03	201	6,51	160	5,23	199	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	4,20E-07
TRINITY_DN122455_c6_g5_i2	5,95	42	6,45	43	6,08	60	0	Endonuclease-reverse transcriptase	5,20E-11
TRINITY_DN122455_c6_g5_i4	4,22	48	5,63	57	3,45	55	0	Endonuclease-reverse transcriptase	1,40E-10
TRINITY_DN122605_c3_g3_i1	2,55	331	3,01	359	3,18	531	0	Reverse transcriptase (RNA-dependent DNA polymerase)	5,30E-08
TRINITY_DN122959_c0_g5_i2	2,43	388	2,75	408	2,37	514	0	Phage integrase, N-terminal SAM-like domain	7,90E-07
TRINITY_DN123032_c5_g2_i1	7,46	109	6,29	90	5,55	122	0	Reverse transcriptase (RNA-dependent DNA polymerase)	2,50E-08
TRINITY_DN123138_c4_g1_i5	8,95	272	8,48	248	7,47	343	M1	Endonuclease-reverse transcriptase	2,40E-07
TRINITY_DN123203_c5_g1_i1	3,00	966	4,10	1181	3,75	1647	M1	Lamina-associated polypeptide 2 alpha	3,90E-06
TRINITY_DN123421_c0_g1_i1	16,11	263	15,11	234	15,17	362	M1	Endonuclease-reverse transcriptase	2,20E-08
TRINITY_DN123475_c4_g1_i1	3,24	126	2,48	99	2,04	115	0	Reverse transcriptase (RNA-dependent DNA polymerase)	6,60E-08
TRINITY_DN123798_c5_g2_i2	2,36	205	2,60	210	2,46	265	0	Phage integrase, N-terminal SAM-like domain	6,50E-05
TRINITY_DN123892_c1_g1_i2	2,30	31	2,17	29	3,76	66	0	GIY-YIG catalytic domain	7,80E-06
TRINITY_DN124258_c3_g5_i3	5,13	171	4,36	142	5,22	247	0	Endonuclease-reverse transcriptase	1,30E-09
TRINITY_DN124747_c1_g2_i1	3,18	163	3,76	178	3,17	228	0	Endonuclease-reverse transcriptase	5,00E-08
TRINITY_DN124833_c0_g1_i2	15,33	47	25,30	73	23,38	105	0	Endonuclease-reverse transcriptase	5,20E-10
TRINITY_DN124833_c2_g1_i1	3,87	205	3,61	186	3,10	237	0	Endonuclease-reverse transcriptase	1,50E-06

TRINITY_DN124840_c1_g1_i1	5,02	255	6,66	310	7,54	497	M1	Myb/SANT-like DNA-binding domain	1,30E-07
TRINITY_DN124840_c1_g1_i4	13,69	72	18,40	90	22,91	166	M1	Myb/SANT-like DNA-binding domain	3,50E-07
TRINITY_DN124952_c0_g5_i13	4,34	93	5,61	110	3,98	136	0	Endonuclease-reverse transcriptase	2,20E-07
TRINITY_DN124952_c0_g5_i5	6,73	50	9,52	65	8,12	84	0	Endonuclease-reverse transcriptase	1,50E-08
TRINITY_DN124952_c0_g5_i8	16,61	93	14,25	77	17,83	152	0	Endonuclease-reverse transcriptase	1,80E-09
TRINITY_DN125031_c0_g1_i2	7,72	137	7,22	122	5,03	131	0	GIY-YIG catalytic domain	1,90E-05
TRINITY_DN125387_c0_g3_i1	6,47	2199	6,80	2191	6,72	3279	M1	Lamina-associated polypeptide 2 alpha	1,20E-06
TRINITY_DN125387_c0_g4_i1	5,11	466	5,27	457	4,70	625	M1	Lamina-associated polypeptide 2 alpha	5,60E-06
TRINITY_DN125387_c0_g6_i1	2,32	699	2,45	698	2,16	905	0	Lamina-associated polypeptide 2 alpha	6,30E-06
TRINITY_DN125472_c0_g3_i2	6,63	82	8,15	94	8,48	150	0	GIY-YIG catalytic domain	9,20E-05
TRINITY_DN125548_c3_g1_i2	6,99	105	6,43	93	7,30	159	0	Endonuclease-reverse transcriptase	1,80E-08
TRINITY_DN125639_c1_g5_i1	19,08	39	26,16	50	23,12	74	0	Endonuclease-reverse transcriptase	1,20E-07
TRINITY_DN125693_c3_g3_i2	4,81	1498	4,36	1324	5,03	2174	M1	Endonuclease-reverse transcriptase	8,20E-07
TRINITY_DN125727_c0_g2_i1	32,95	33	46,91	44	34,22	52	0	Endonuclease-reverse transcriptase	1,90E-06
TRINITY_DN125735_c4_g1_i1	27,26	83	22,83	66	26,35	136	0	Endonuclease-reverse transcriptase	1,30E-10
TRINITY_DN125883_c2_g20_i5	6,37	3275	7,32	3529	7,02	5124	M1	Phage integrase, N-terminal SAM-like domain	1,00E-07
TRINITY_DN125994_c1_g3_i1	14,96	334	10,40	227	12,52	421	M1	Endonuclease-reverse transcriptase	2,50E-11
TRINITY_DN126164_c5_g6_i1	69,03	153	48,31	102	44,78	159	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	5,70E-09
TRINITY_DN126236_c2_g2_i1	17,25	221	18,98	230	18,85	338	M1	Myb/SANT-like DNA-binding domain	1,30E-06
TRINITY_DN126384_c1_g2_i1	13,56	100	10,85	77	11,10	115	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	3,40E-12
TRINITY_DN126393_c0_g4_i1	16,34	138	14,20	116	14,24	183	M1	Endonuclease-reverse transcriptase	2,10E-09
TRINITY_DN126549_c2_g12_i2	13,66	281	11,94	236	10,40	340	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	3,40E-07
TRINITY_DN126787_c3_g3_i6	11,56	377	12,80	392	11,11	547	M1	Endonuclease-reverse transcriptase	5,80E-07
TRINITY_DN126806_c3_g1_i2	7,37	44	15,82	85	16,86	143	0	Endonuclease-reverse transcriptase	1,10E-08

TRINITY_DN126806_c3_g3_i3	7,63	170	9,43	196	7,32	246	M1	Endonuclease-reverse transcriptase	4,10E-07
TRINITY_DN127354_c2_g1_i2	15,72	65	17,40	68	16,73	102	0	Endonuclease-reverse transcriptase	9,40E-11
TRINITY_DN127354_c2_g10_i1	19,67	60	23,44	68	17,09	85	0	Endonuclease-reverse transcriptase	2,00E-07
TRINITY_DN127977_c3_g5_i1	10,02	1267	10,81	1294	9,83	1813	M1	Myb/SANT-like DNA-binding domain	7,20E-09
TRINITY_DN128058_c4_g1_i1	20,82	179	24,44	199	24,08	318	M1	Endonuclease-reverse transcriptase	1,30E-06
TRINITY_DN128224_c2_g7_i1	3,93	81	3,68	74	4,38	125	0	Reverse transcriptase (RNA-dependent DNA polymerase)	1,40E-07
TRINITY_DN128256_c2_g1_i1	7,08	74	7,55	75	6,46	106	0	Reverse transcriptase (RNA-dependent DNA polymerase)	1,00E-09
TRINITY_DN128338_c2_g7_i1	6,54	202	6,08	180	6,78	276	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	1,10E-09
TRINITY_DN128703_c0_g2_i1	6,07	181	6,51	184	4,02	179	0	Retroviral aspartyl protease	0,00024
TRINITY_DN128930_c0_g4_i2	9,07	30	8,34	26	8,87	40	0	Endonuclease-reverse transcriptase	2,40E-09
TRINITY_DN128939_c0_g2_i3	6,68	288	6,29	263	7,25	421	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	3,30E-08
TRINITY_DN128979_c3_g2_i1	4,34	197	3,00	141	3,90	255	0	Reverse transcriptase (RNA-dependent DNA polymerase)	4,20E-10
TRINITY_DN129087_c4_g3_i1	2,82	237	2,70	221	2,73	305	0	Myb/SANT-like DNA-binding domain	2,30E-05
TRINITY_DN129346_c3_g5_i1	11,43	125	8,01	86	8,83	161	0	Reverse transcriptase (RNA-dependent DNA polymerase)	6,50E-09
TRINITY_DN129369_c0_g1_i1	7,01	35	10,39	47	6,47	46	0	Reverse transcriptase (RNA-dependent DNA polymerase)	1,20E-09
TRINITY_DN129445_c0_g4_i2	18,38	194	17,27	173	19,20	294	0	Endonuclease-reverse transcriptase	1,20E-09
TRINITY_DN129445_c0_g4_i5	25,21	88	23,45	78	19,80	107	0	Endonuclease-reverse transcriptase	6,20E-09
TRINITY_DN129448_c4_g2_i3	30,43	503	27,45	433	25,31	621	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	2,90E-09
TRINITY_DN129521_c7_g1_i1	19,45	80	26,74	103	21,82	137	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	3,50E-09
TRINITY_DN129843_c1_g2_i2	3,81	445	3,84	429	3,99	657	M1	Endonuclease-reverse transcriptase	5,70E-07
TRINITY_DN130095_c0_g8_i2	9,54	280	6,37	187	8,22	340	M1	Endonuclease-reverse transcriptase	5,80E-11

TRINITY_DN130179_c4_g1_i3	21,68	148	18,86	123	14,69	170	M1	Endonuclease-reverse transcriptase	4,70E-09
TRINITY_DN142749_c0_g1_i1	24,61	25	24,09	23	11,41	17	0	GIY-YIG catalytic domain	4,60E-06
TRINITY_DN158655_c0_g1_i1	13,14	41	10,81	33	8,76	42	0	GIY-YIG catalytic domain	1,10E-05
TRINITY_DN16787_c0_g1_i1	22,74	46	50,36	95	24,82	73	0	GIY-YIG catalytic domain	3,60E-06
TRINITY_DN36440_c0_g1_i1	4,06	147	2,37	93	1,61	97	0	GIY-YIG catalytic domain	4,50E-05
TRINITY_DN45963_c0_g1_i1	27,01	27	45,86	43	21,73	31	0	GIY-YIG catalytic domain	1,10E-05
TRINITY_DN62813_c0_g1_i1	23,79	24	16,31	16	19,48	29	0	GIY-YIG catalytic domain	1,10E-05
TRINITY_DN79505_c1_g2_i1	3,88	67	4,88	77	5,31	127	0	GIY-YIG catalytic domain	8,70E-06
TRINITY_DN81696_c0_g2_i1	12,21	26	13,07	26	10,54	31	0	GIY-YIG catalytic domain	1,40E-05
TRINITY_DN85912_c0_g3_i1	44,01	87	41,70	79	23,16	71	M1	Reverse transcriptase (RNA-dependent DNA polymerase)	1,10E-07
TRINITY_DN86140_c0_g1_i1	5,77	920	6,01	911	6,03	1378	M1	Endonuclease-reverse transcriptase	8,50E-08
TRINITY_DN96530_c0_g1_i1	3,96	95	2,88	71	2,76	94	0	GIY-YIG catalytic domain	3,40E-05
TRINITY_DN99783_c0_g1_i1	10,73	23	20,12	39	19,38	55	0	GIY-YIG catalytic domain	5,40E-07

Apêndice C - Tabela dos resultados para os 46 candidatos a lncRNAs linhagem-específicos nas análises de: Expressão Diferencial, o CEMiTool com os respectivos módulos associados a cada transcrito e o Annocript com as probabilidades de codificação. Os candidatos marcados na coloração verde são os que, após análise no RNAfold web suit apresentaram estrutura secundária termodinamicamente viável.

Contigs (putative lncRNA)	1 dpa		14 dpa		21 dpa		CEMiTool Module	Annocript Coding Probability (%)
	Fold Change	Total Gene Reads	Fold Change	Total Gene Reads	Fold Change	Total Gene Reads		
TRINITY_DN104611_c0_g2_i1	22,43	23	27,05	26	13,24	22	0	8,49E-01
TRINITY_DN108590_c1_g2_i1	9,06	51	6,83	38	9,79	94	0	4,68E-05
TRINITY_DN116435_c0_g1_i2	10,54	550	8,85	447	8,19	657	M1	5,69E-04
TRINITY_DN116435_c0_g3_i1	5,87	1769	5,24	1536	4,84	2285	M1	7,95E-04
TRINITY_DN118383_c0_g1_i1	9,59	75	4,34	36	6,62	69	M1	4,67E-05
TRINITY_DN119279_c0_g5_i1	10,95	248	6,77	154	11,17	352	M1	4,73E-05
TRINITY_DN119387_c3_g1_i1	8,29	115	6,91	93	5,19	113	0	2,55E-01
TRINITY_DN119952_c2_g2_i1	2,87	271	2,29	221	2,21	306	0	1,57E+00
TRINITY_DN121120_c4_g5_i2	9,55	87	7,15	64	6,38	91	0	1,49E+00
TRINITY_DN121327_c2_g8_i1	4,24	551	3,99	502	4,16	769	M1	5,31E-01
TRINITY_DN121506_c5_g2_i1	3,97	398	5,33	484	4,76	672	M1	4,89E-01
TRINITY_DN121541_c0_g1_i2	5,40	527	5,55	517	4,94	684	M1	6,38E-05
TRINITY_DN121840_c0_g2_i1	5,26	604	5,04	556	4,64	773	M1	3,43E-04
TRINITY_DN122360_c4_g2_i2	5,40	1426	5,15	1307	5,47	2042	M1	2,91E-04
TRINITY_DN122542_c1_g2_i2	3,24	289	3,98	327	4,22	515	M1	3,24E-01
TRINITY_DN122682_c0_g1_i1	4,81	166	4,08	139	3,29	172	0	7,47E-05
TRINITY_DN123033_c16_g1_i1	3,86	2993	4,57	3283	4,32	4593	M1	3,59E-01
TRINITY_DN125013_c0_g2_i1	6,44	124	5,74	107	4,87	134	0	3,12E-01
TRINITY_DN125294_c1_g8_i1	4,03	197	5,91	260	5,46	356	M1	1,86E-04
TRINITY_DN125898_c0_g2_i1	4,35	305	4,35	291	4,28	422	M1	3,30E+00
TRINITY_DN126208_c2_g3_i2	9,14	291	7,42	230	8,90	411	M1	6,38E-01
TRINITY_DN126278_c2_g1_i1	7,26	289	8,51	316	7,17	412	M1	5,08E-01
TRINITY_DN126712_c2_g2_i1	4,94	331	4,85	312	4,71	454	M1	5,43E-04
TRINITY_DN127136_c1_g3_i4	9,53	73	12,74	91	10,80	117	0	8,44E-04
TRINITY_DN127306_c2_g7_i1	3,44	1098	3,46	1059	2,77	1270	M1	4,81E+00
TRINITY_DN127425_c3_g3_i1	15,39	792	15,43	757	12,91	1047	M1	4,77E-05
TRINITY_DN127425_c3_g5_i2	10,15	351	8,11	272	8,30	429	M1	1,19E-03
TRINITY_DN127545_c6_g1_i1	2,13	142	1,79	122	2,32	203	0	5,75E-01
TRINITY_DN127747_c5_g1_i1	10,04	171	9,75	160	7,92	202	M1	4,82E-04
TRINITY_DN127747_c5_g6_i1	10,49	233	9,68	205	7,73	265	M1	2,34E+00
TRINITY_DN127819_c0_g2_i1	21,67	265	17,92	210	16,13	297	M1	1,04E-04
TRINITY_DN127819_c0_g2_i2	27,91	266	14,51	136	15,83	229	M1	2,86E-01
TRINITY_DN128349_c6_g5_i1	7,63	3025	7,92	2982	7,31	4165	M1	1,43E-04
TRINITY_DN128368_c9_g5_i1	4,25	1013	4,48	1011	3,99	1375	M1	8,18E-04
TRINITY_DN128368_c9_g5_i3	5,65	1341	5,38	1229	5,00	1733	M1	4,09E-01
TRINITY_DN128556_c1_g7_i1	16,91	805	15,65	711	16,32	1128	M1	6,44E-05
TRINITY_DN128998_c2_g6_i1	19,97	1713	24,00	1930	20,19	2580	M1	1,05E-04
TRINITY_DN21685_c0_g1_i1	6,87	51	6,91	49	9,91	106	0	2,83E-01
TRINITY_DN24628_c0_g1_i1	17,55	36	12,47	25	15,49	49	0	2,21E-04
TRINITY_DN24674_c0_g1_i1	22,62	23	4,34	5	1,39	4	0	8,09E-04
TRINITY_DN59921_c0_g1_i1	15,03	66	14,57	61	14,10	93	M1	2,45E+00
TRINITY_DN64687_c0_g1_i1	63,64	63	45,65	43	40,58	66	M1	1,01E+00
TRINITY_DN68223_c0_g1_i1	13,12	31	14,02	31	19,37	70	0	3,96E+00
TRINITY_DN76107_c0_g2_i1	2,88	24	5,04	36	2,92	34	0	5,05E-05
TRINITY_DN90964_c0_g1_i2	8,63	28	6,51	21	10,45	48	0	4,81E-05
TRINITY_DN96821_c0_g2_i1	3,81	1200	4,34	1278	4,42	1898	M1	6,73E-01