UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METRÓPOLE DIGITAL
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA
MESTRADO ACADÊMICO EM BIOINFORMÁTICA

RNA-Gatherer: uma ferramenta computacional para anotação de RNAs não-codificantes

Pitágoras de Azevedo Alves Sobrinho

Natal-RN, Brasil
2021

Pitágoras de Azevedo Alves Sobrinho

RNA-Gatherer: uma ferramenta computacional para anotação de RNAs não-codificantes

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do título de Mestre em Bioinformática.

Linha de pesquisa: Genômica

Orientador: Wilfredo Blanco Figuerola
Coorientador: Jorge Estefano S. de Souza

Natal-RN, Brasil
2021

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Leopoldo Nelson - -Centro de Biociências - CB

Alves Sobrinho, Pitágoras de Azevedo.
RNA-Gatherer: uma ferramenta computacional para anotação de RNAs não-codificantes / Pitágoras de Azevedo Alves Sobrinho. - Natal, 2021.
75 f.: il.

Dissertação (Mestrado) - Universidade Federal do Rio Grande do Norte. Instituto Metrópole Digital. Programa de Pós-graduação em Bioinformática.
Orientador: Prof. Dr. Wilfredo Blanco Figuerola.
Coorientador: Prof. Dr. Jorge Estefano S. de Souza.

1. Anotação de genes - Dissertação. 2. ncRNA - Dissertação. 3. Predição de funções - Dissertação. I. Figuerola, Wilfredo Blanco. II. Souza, Jorge Estefano S. de. III. Universidade Federal do Rio Grande do Norte. IV. Título.

RN/UF/BSCB CDU 575.113

Elaborado por KATIA REJANE DA SILVA - CRB-15/351

PITÁGORAS DE AZEVEDO ALVES SOBRINHO

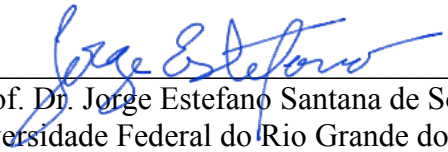
"RNA-GATHERER: UMA FERRAMENTA COMPUTACIONAL PARA ANOTAÇÃO DE RNAS NÃO-CODIFICANTES EM ORGANISMOS POUCO CONHECIDOS"
Defesa de Mestrado apresentanda ao Programa de Pós-Graduação em Bioinformática da Universidade Federal do Rio Grande do Norte.

Área de concentração: Bioinformática
Linha de Pesquisa: Genômica
Orientador: Prof. Dr. Wilfredo Blanco Figuerola

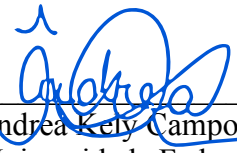
Natal, 29 de janeiro de 2021.

BANCA EXAMINADORA

Prof. Dr. Wilfredo Blanco Figuerola
Universidade Federal do Rio Grande do Norte
(Presidente)



Prof. Dr. Jorge Estefano Santana de Souza
Universidade Federal do Rio Grande do Norte
(Avaliador Interno)



Prof. Dra. Ândrea Kely Campos Ribeiro dos Santos
Universidade Federal do Pará
(Examinador Externo à Instituição)

Esta dissertação é dedicada a todos os amigos, familiares e colegas que acreditaram em mim e me apoiaram como puderam. Eu não teria chegado onde cheguei sem eles.

Agradecimentos

Primeiramente, agradeço à instituição que me permitiu crescer academicamente e que por muito tempo foi a minha casa, a **Universidade Federal do Rio Grande do Norte** - UFRN. Também agradeço à **Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior** - CAPES, pelo fomento durante o desenvolvimento deste projeto. Agradeço ao **Bioinformatics Multidisciplinary Environment** - BioME, do **Instituto Metrópole Digital**, e toda sua equipe, que me deram suporte, estrutura física e os recursos disponíveis para realizar minha pesquisa.

Este trabalho nunca teria sido possível sem as contribuições, ensinamentos, as discussões, orientações e a paciência dos professores **Jorge Estefano de Souza**, meu co-orientador, e **Wilfredo Blanco**, meu orientador acadêmico no mestrado. Também agradeço aos alunos **Inácio Gomes**, do Doutorado em Bioinformática, e a **Diego Marques**, da Universidade Federal do Pará, pelas contribuições no desenvolvimento do artigo que também resultou desta pesquisa. Também agradeço aos meus colegas de pós-graduação **Danilo Lopes**, **Tayná Fiúza**, **Renata Cavalcante**, **Marília Dantas** e **Ricardo Almeida** por compartilharem suas expertises comigo quando precisei, mas também pela companhia nesta difícil jornada.

Não posso deixar de agradecer aos meus grandes amigos e amigas **Priscila Bessa**, **Heloysa Galvão**, **Ronaldo Silveira**, **Ana Caroline**, **Morgana Dias**, **Isaac Oliveira** e **Eros Gibson** que sempre estiveram ao meu lado nos melhores e em muitos dos piores momentos, sem nunca deixar de me apoiar. Aos meus pais, **Ana Cristina da Silva** e **Abraão Alves**, que sempre batalharam e fizeram tudo o que podiam por mim. E por último, mas não menos importante, quero agradecer à **Carlos Barata**, por sempre ter sido um exemplo e inspiração na minha vida.

Somos uma forma do cosmos conhecer a si mesmo.
Carl Sagan.

Resumo

RNAs não-codificantes são moléculas que desempenham papéis decisivos na regulação de genes, então identificá-los é essencial para entender a genética de uma espécie. Diversos fatores, como: baixo nível de expressão, amplo espectro de subtipos, atributos diversos, funções heterogêneas e ausência de homologia entre espécies; fazem a detecção de ncRNAs um desafio. Recentemente, novas estratégias de bioinformática vem tentando vencer esses desafios usando modelos de covariância e inteligência artificial. A co-expressão desses genes também vem sendo analisada computacionalmente para revelar quais são suas funções. No entanto, não há consenso sobre quais métricas e parâmetros usar no processo de prever funções. Em organismos pouco conhecidos, como *Arapaima gigas*, a falta de informações de referência aumenta essa dificuldade. Além disso, principalmente para RNAs longos não-codificantes, há poucas funções conhecidas, o que torna difícil explicar os papéis desses genes e avaliar a qualidade das predições. Neste trabalho, é descrito um *software* para descobrir os genes não-codificantes, de diversos tipos, e suas funções em espécies de eucariotos. Este foi validado com uma espécie modelo, o camundongo, e utilizado para explorar o panorama de ncRNAs numa espécie pouco estudada, o *Arapaima gigas*. A comparação da semelhança entre funções de genes co-expressos nos permitiu definir níveis de confiança para as métricas de calcular co-expressão, e assim, desenvolver uma *pipeline* de predição funções para lncRNA, a qual inclui métricas para calcular correlações não-lineares. O pacote de *software* descrito aqui fez 63307 anotações não-codificantes em *A. gigas*, incluindo 11 tipos de ncRNA e 4 de regiões cis-regulatórias. Dessas anotações, apenas 706 eram similares a ncRNAs já conhecidos em outras espécies e os restantes não haviam sido descritos anteriormente. A análise exploratória dos lncRNAs também revelou 19854 lncRNAs de tecido específico e 256 lncRNAs expressos de forma onipresente. Prever as funções dessas moléculas também revelou que elas estão envolvidas na pigmentação da pele, diferenciação sexual, crescimento e defesa contra tumores.

Palavras-chave: ncRNA. *Arapaima Gigas*. Anotação de Genes. Predição de Funções.

Abstract

Non-coding RNAs are molecules that play decisive roles in several types of gene regulation. Identifying them is necessary for understanding the genetics of a species. Several factors, such as: low level of expression, the broad spectrum of subtypes, diverse attributes, heterogeneous functions and absence of homology between species; make the detection of ncRNAs genes a challenge. The latest bioinformatics strategies for detecting ncRNA genes have tried to identify their locations in the genomes and their secondary structures, using covariance models and artificial intelligence. The co-expression of these genes has been computationally analyzed in order to reveal their functional annotations. However, there is no consensus on which metrics and parameters to use in the process of predicting the functions of these molecules. In organisms little known, such as *Arapaima gigas*, the lack of reference information increases the difficulty. Additionally, even for known long non-coding RNAs, there is little functional information, which makes it difficult to explain the roles of these genes. In this work, we describe a software for discovering the non-coding genes, including their diverse types, and their functions in eukaryotic genomes. It was validated by annotating a model species (*Mus musculus*) and then used to explore the landscape of ncRNA in *Arapaima gigas*. Comparing the similarity between the functions of co-expressed genes allowed us to define confidence levels for the metrics that measure co-expression, and thus, develop a pipeline for predicting lncRNA functions, which includes metrics for non-linear correlations. The described software suite made 63307 non-coding annotations in *A. gigas*, including 11 types of ncRNA and 4 types of cis-regulatory regions. Of these annotations, only 706 are similar to ncRNAs already known in other species and the remaining were never described before. The exploratory analysis of lncRNA also revealed 19854 tissue specific lncRNAs and 256 lncRNAs ubiquitously expressed. Predicting the functions of these molecules revealed RNAs involved in skin pigmentation, sex differentiation, growth and defense against tumors.

Keywords: ncRNA. Arapaima Gigas. Gene Annotation. Function Prediction.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

DNA	*Deoxyribonucleic Acid* (ácido desoxirribonucleico)
cDNA	DNA Complementar
RNA	*Ribonucleic Acid* (ácido ribonucleico)
ncRNA	RNA Não-Codificante
nt	Nucleotídeos
aa	Aminoácidos
BLAST	*Basic Local Alignment Search Tool*
MC	Modelo de Covariância
HMM	*Hidden Markov Model* (modelo oculto de Markov)
GO	*Gene Ontology* (Ontologia de Genes)
tRNA	RNA de Transferência
mRNA	RNA Mensageiro
lncRNA	RNA longo Não-Codificante
ORF	*Open Reading Frame* (fase de leitura aberta)
3‘ UTR	*3’ Untranslated Region* (região não-traduzida 3’)
MIC	*Maximal Information Coefficient* (coeficiente de informação máxima)
DC	*Distance Correlation* (correlação de distância)
UML	*Unified Modeling Language* (linguagem de modelagem unificada)
API REST	*Application Programming Interface for Representational State Transfer* (interface de programação de aplicações para transferência representacional de estado)
NC	Nível de Confiança
FDR	*False Discovery Rate* (taxa de falsas descobertas)
RNA-Seq	*RNA Sequencing* (sequenciamento de RNA)

TPM	Transcritos Por Milhão
NR	*Non-Redundant Sequence Database* (banco de dados de sequências não-redundantes)
DE	Diferencialmente Expresso
Log2	Logaritmo de Base 2
Cis-reg	Elemento Cis-Regulador
MF	*Molecular Function* (função molecular)
BP	*Biological Process* (processo biológico)
CC	*Cellular Component* (componente celular)

Lista de símbolos

$\mathbb{R}$	Conjunto dos Números Reais
$\emptyset$	Conjunto Vazio

Sumário

1 Introdução

Desde o primeiro sequenciamento completo de um genoma, feito da bactéria *Haemophilus influenzae* em 1995 (FLEISCHMANN et al., 1995), e passando pelo Projeto do Genoma Humano, concluído em 2003 (NHGRI, 2020), o campo da genômica tem avançado muito. Hoje, no ano de 2021, 15333 genomas eucarióticos e 298676 genomas de procariotos diferentes já foram sequenciados (COORDINATORS, 2018). No entanto, para compreender a genética de um organismo, não basta ter uma montagem do genoma dele. Pois, dentro desse genoma, há segmentos gênicos diversos, que precisam ser preditos e anotados.

Predizer um gene é o ato de detectar uma parte do genoma que pode estar sendo expressa para criar uma sequência de RNA e, talvez, também codificar uma proteína. Uma *pipeline* para anotação de genes é algo mais complexo que uma predição, pois ela precisa incluir (além das predições) informações de homologia e de expressão gênica (YANDELL; ENCE, 2012).

Essas homologias podem ser estabelecidas com bancos de dados de sequência como RefSeq (PRUITT; TATUSOVA; MAGLOTT, 2007) e RNACentral (CONSORTIUM, 2019). Além disso, há diversas outras informações que podem acrescentar detalhes na anotação. Bancos de dados como PFAM (FINN et al., 2016) e RFAM (KALVARI et al., 2018) descrevem famílias de genes codificantes e não-codificantes. Também há ontologias, que podem ser utilizadas para descrever as funções dos genes anotados. Um exemplo é o KEGG *Pathway Database*, que busca representar o conhecimento atual sobre vias moleculares, com suas interações, relações e reações (KANEHISA; GOTO, 2000). Outra iniciativa para representar funções é o *Gene Ontology*, que compreende um vocabulário universal de termos para descrever processos biológicos, funções moleculares e componentes celulares, através de um grande grafo que relaciona esses termos (ASHBURNER et al., 2000). Quando faltam funções conhecidas, esses termos e vias também podem ser preditos, utilizando metodologias diversas, realizando uma predição funcional (YANG et al., 2018; BAEK et al., 2018; EHSANI; DRABLøS, 2018; JIANG et al., 2015). Uma boa anotação deve integrar todos esses dados, provenientes de muitos esforços diferentes para compreender a genética dos seres vivos.

1.1 O *Arapaima Gigas*

A espécie *A. gigas* (Figura 1), conhecida popularmente como Pirarucu, possui poucos parentes próximos. Ela é a única remanescente do gênero *Arapaima* e compartilha a família *Arapaimidae* apenas com *Heterotis niloticus*, uma espécie presente em diversas partes da Africa (BETANCUR-R et al., 2017; NELSON; GRANDE; WILSON, 2016).

Estimativas indicam que essas duas espécies, localizadas em continentes diferentes, tenham se tornado geograficamente isoladas entre 50 e 85 milhões de anos atrás (HAO et al., 2020; LAVOUÉ, 2016). A estimativa mais recente, feita por Hao et al. (2020), aponta que essa divergência teria 59 milhões de anos. Essa família faz parte da sub-ordem *Osteoglossoidei*, que contém também os *Osteoglossidae* (aruanãs) (BETANCUR-R et al., 2017; NELSON; GRANDE; WILSON, 2016; LI, 1996). Estudos recentes estimmaram a idade da divergência entre o Aruanã Dourado (*Scleropages formosus*) e o Pirarucu em 106 a 91 milhões de anos (HAO et al., 2020; VIALLE et al., 2018).

Figura 1 – Peixe da espécie *Arapaima gigas*, ou Pirarucu, com escamas apresentando coloração avermelhada

Fonte: foto obtida de Bjoertvedt (2009)

Diversas características tornam esta uma espécie singular. Primeiramente, os *A. gigas* são peixes capazes de respiração aérea (VIALLE et al., 2018; ALMEIDA, 2006), assim como os *H. niloticus* (FAGBENRO et al., 2000; D'AUBENTON, 1955). Com um comprimento entre 2 e 4.5 metros (NELSON; GRANDE; WILSON, 2016; HRBEK; CROSSA; FARIAS, 2007; DE-GROOT, 1991), e um peso superior a 200kg (HRBEK; CROSSA; FARIAS, 2007), é um dos maiores peixes na América do Sul. O crescimento também é muito acelerado, chegando a 10kg em um ano, o que torna essa espécie a mais promissora para o desenvolvimento de piscicultura intensiva na região amazônica (ALMEIDA, 2006; IMBIRIBA, 2001; CARVALHO; NASCIMENTO, 1992).

Também há pouco dimorfismo sexual aparente, o que dificulta diferenciar machos e fêmeas. O único dimorfismo sexual apresentado é uma maior intensidade de coloração avermelhada (Figura 1) desenvolvida pelos machos na época de reprodução (ALMEIDA, 2006). O comportamento reprodutivo, de acordo com Imbiriba (2001), consiste de uma etapa de cortejamento, seguida da formação de um casal, a construção de um ninho, o acasalamento e a defesa da prole. Como indivíduos precisam formar casais para reprodução, a aparente falta de diferenças entre machos e fêmeas antes da temporada de reprodução pode dificultar a tarefa de colocar um número igual de indivíduos de ambos os sexos em tanques, o que pode ser um problema para a piscicultura.

Estudos genéticos sobre *A. gigas* são essenciais, dado sua importância ecológica e comercial para a região da Amazônia. Dois estudos publicaram montagens de genoma (VIALLE et al., 2018; DU et al., 2019) e um estudo investigou o transcriptoma da espécie (MARTINS et al., 2020). Tais pesquisas analisaram a filogenia e gigantismo da espécie, assim como possíveis diferenças entre os genomas de macho e fêmea. No entanto, mesmo com diversos dados de genoma e transcriptoma publicados (DU et al., 2019; VIALLE et al., 2018; WATANABE et al., 2018; RAMÍREZ et al., 2018; PEREIRA et al., 2017), até o momento os esforços para entender a genética do Pirarucu foram focados apenas nos genes codificantes, de tal forma que o panorama dos RNAs não-codificantes nessa espécie permanece um campo pouco explorado. Como as informações são escassas, para estudar esses RNAs se faz necessário desenvolver um *pipeline* suficientemente genérico, que possa encontrar informações novas mesmo num peixe como o *A. gigas*, que é pouco coberto pelos bancos de dados.

1.2 RNAs Não-Codificantes

Conforme o nome sugere, RNAs não-codificantes (ncRNAs) consistem em RNAs que não codificam proteínas. Estes RNAs tem bastante diversidade, totalizando 3941 famílias distintas, sendo produzidos por genes não-codificantes e introns autocatáliticos (KALVARI et al., 2018). Eles agem de muitas formas na manutenção das células, como por exemplo regulação epigenética, regulação de transcrição, regulação pós-transcricional, defesa do genoma, entre outras funções (SIGNAL; GLOSS; DINGER, 2016; HUNG et al., 2011). Já foi observada uma correlação entre a complexidade dos organismos e a quantidade de genes não-codificantes presentes no genoma (MATTICK; TAFT; FAULKNER, 2010; TAFT; MATTICK, 2003). Também já foi sugerido que os RNAs não-codificantes poderiam estar por trás da complexidade dos eucariotos (MATTICK, 2004). Por isso, hoje os genes que codificam ncRNAs são entendidos como importantes reguladores dos mais diversos processos biológicos. Isso não seria possível senão pela capacidade das moléculas de ncRNA para adquirir conformações chamadas de estruturas secundárias, formadas por pareamento Watson-Crick de nucleotídeos (BRION; WESTHOF, 1997).

1.2.1 Anotação de RNAs Não-Codificantes

A quantidade dos genes de ncRNA pode também superar a de genes codificantes. No genoma humano (montagem GRCh38.p13), por exemplo, além dos 20,438 genes codificantes, há 24,000 genes não-codificantes identificados (YATES et al., 2020). Apesar de corresponderem a uma porção expressiva do genoma em eucariotos (AL-TOBASEI; PANERU; SALEM, 2016; QUINN; CHANG, 2016; HUNG et al., 2011), seu processo de identificação ainda enfrenta diversas dificuldades, dado seus baixos níveis de expressão e conservação(BAEK et al., 2018; QUINN; CHANG, 2016). Diversos trabalhos têm buscado

nos últimos anos identificar tais RNAs em organismos diversos (MA et al., 2018; KERN et al., 2018; HARRIS; KOVACS; LONDO, 2017; SCOTT et al., 2017; AL-TOBASEI; PANERU; SALEM, 2016), sendo esses esforços de anotação importantes para a expansão dos nossos conhecimentos sobre os ncRNAs.

1.2.1.1 Busca por Homologia

Uma forma comum de anotar genes codificantes é usar algoritmos de alinhamento como BLAST para buscar homologias com sequencias de genes já conhecidos e caracterizados, os quais podem ser utilizados para inferir funções (CONESA et al., 2005). Essa forma de anotação é ideal para encontrar homologias altamente conservadas, mas as sequências de genes não-codificantes não apresentam o mesmo nível de conservação que genes codificantes (FREYHULT; BOLLBACK; GARDNER, 2007). Assim, se apenas esse método for utilizado, a grande diversidade dos ncRNAs acaba sendo subestimada (FREYHULT; BOLLBACK; GARDNER, 2007).

No entanto, há uma estratégia de anotação que pode identificar uma diversidade maior de ncRNAs: os Modelos de Covariância (MC). Eles são uma generalização (EDDY; DURBIN, 1994) dos *Hidden Markov Models* (HMMs), rigorosos modelos probabilísticos que já foram aplicados no reconhecimento de fala (RABINER, 1989), reconhecimento de padrões de comportamento animal (TENNESSEN et al., 2019) e na análise de sequências de proteínas (BALDI et al., 1994). Um MC é construído a partir do alinhamento múltiplo de sequências de ncRNA semelhantes entre si, usando uma ferramenta como o *cmbuild* do pacote *Infernal* (NAWROCKI; EDDY, 2013). O resultado é um modelo probabilístico que descreve a estrutura secundária e o consenso entre as sequências de RNA analisadas (EDDY; DURBIN, 1994).

O projeto RFAM (rfam.xfam.org) é um banco de dados que, no momento da escrita deste texto, contém MCs para 3941 famílias não-codificantes distintas (KALVARI et al., 2018). Além de famílias para RNAs não-codificantes, o RFAM também inclui famílias de elementos cis-regulatórios, que são regiões de DNA não-codificante importantes na regulação de genes próximos. Esses modelos podem ser buscados em genomas através da ferramenta *cmscan*, também do pacote *Infernal*. Dessa forma, a combinação desse pacote com o banco de dados RFAM permite, além de prever novos ncRNAs, identificar a qual família tais RNAs pertencem (NAWROCKI, 2014). Essas famílias também podem ter funções já conhecidas, descritas por termos de *Gene Ontology* (GO) (ASHBURNER et al., 2000), o que permite associar funções ao novo ncRNA (KALVARI et al., 2018). Outra ferramenta que faz uso de modelos de covariância é o *tRNAscan-SE*. Este software é especializado em RNAs de transferência (tRNAs) e busca no genoma modelos criados para os diversos isótipos desses RNA. Além disso, a ferramenta também informa os anticódons dessas moléculas, necessários para determinar quais RNAs mensageiros (mRNAs) podem ser transportados por eles (CHAN; LOWE, 2019).

1.2.1.2 Predição de RNAs longos Não-Codificantes

Os RNAs longos não-codificantes (lncRNA) são um tipo de ncRNA particularmente numeroso nos genomas, mas que apresenta dificuldades particulares de anotação (JOHNSSON et al., 2014). Esses RNAs possuem baixíssimo nível de conservação, o que dificulta a anotação por homologia de sequência (CAMARGO et al., 2020; BAEK et al., 2018; QUINN; CHANG, 2016; JOHNSSON et al., 2014). Além disso, a estrutura desse tipo de ncRNA ainda é pouco compreendida, o que também dificulta a anotação (JOHNSSON et al., 2014). Em organismos pouco estudados (como o *A. gigas*), os quais tem baixa representação nos bancos de dados, a busca por homologia se torna ainda mais limitada (CAMARGO et al., 2020; WANG et al., 2019). Para complicar ainda mais a tarefa de identifica-los, os lncRNA contêm *Open Reading Frames* (ORFs) fracos e características semelhantes aos 3‘ UTRs de RNAs mensageiros, o que os torna muito semelhantes a genes codificantes no nível de sequência primária (NIAZI; VALADKHAN, 2012). Consequentemente, métodos para identificação precisam ser capazes de detectar genes pouco conservados e também de diferenciar essas moléculas dos mRNAs.

Já foram propostas diversas estratégias de predição de lncRNA livres de alinhamento (CAMARGO et al., 2020; YANG et al., 2018; BAEK et al., 2018; WANG et al., 2013; SUN et al., 2013), as quais não precisam realizar extensas buscas por homologia em bancos de dados. No entanto, esses algoritmos precisam ser treinados com dados de alta qualidade sobre lncRNAs e RNAs codificantes, disponíveis para apenas algumas espécies, o que dificulta o treinamento de um algoritmo que seja capaz de prever novos lncRNA em espécies diversas (WANG et al., 2019). Um preditor de lncRNAs recente chamada *RNA Samba*, baseado em aprendizado profundo, foi criado de forma a ser genérico o suficiente para realizar predições em diversas espécies (CAMARGO et al., 2020). Diferentemente de preditores baseados em aprendizado profundo como *lncRNAnet* (BAEK et al., 2018) e *mRNN* (hendrixlab.cgrb.oregonstate.edu/mRNN.html), os quais usam redes neurais recorrentes, o *RNA Samba* usa a arquitetura IGLOO, que olha para as sequências como um todo (SOURKOV, 2018). Ele foi comparado, em múltiplas espécies, com outras ferramentas que tem o mesmo objetivo: *CPAT* (WANG et al., 2013), *CPC2* (KANG et al., 2017), *FEELnc* (WUCHER et al., 2017), *lncRNAnet* e *mRNN*. *RNA Samba* obteve melhor performance que as demais ferramentas na predição de lncRNAs em *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* e *Saccharomyces cerevisiae*; a única exceção foi a espécie *Danio rerio*, para a qual a performance foi muito próxima do *FEELnc* (CAMARGO et al., 2020).

Além de usar apenas uma ferramenta de predição para identificar transcritos de lncRNA, pesquisadores também aplicam outros filtros nos resultados delas para obter resultados mais precisos. É o caso da anotação de lncRNAs em Truta-arco-íris (*Oncorhynchus mykiss*) feita por Al-Tobasei, Paneru e Salem (2016), que usaram a ferramenta CPC, mas também removeram potenciais lncRNAs que se encaixassem nos seguintes critérios:

menos de 200nt (nucleotídeos) de comprimento, ORF com comprimento superior a 100aa (aminoácidos) e alinhamento bem sucedido a uma sequência de proteína conhecida ou um ncRNA de outro tipo.

1.2.1.3 Integração de Estratégias de Anotação

As diferentes estratégias de anotação introduzidas até aqui possuem vantagens e limitações próprias. Considerando isso, alguns estudos tem tentado realizar uma integração das diferentes estratégias, de forma a anotar tipos diversos de ncRNA em genomas (MOSTAJO et al., 2020; ANTHON et al., 2014). Para anotar os ncRNAs no genoma de porco (*Sus scrofa*), Anthon et al. (2014) utilizou busca por homologia de sequência, busca por homologia de estrutura (através dos modelos de covariância) e métodos para tipos específicos de RNA, como *tRNAscan-SE*. Infelizmente, esse estudo não disponibilizou o *pipeline* utilizado na forma de um *software*, o que permitiria anotar outros genomas da mesma forma. No entanto, foi criado um *pipeline* chamado GORAP (github.com/koriege/gorap, Riege & Marz, não publicado) para a anotação de ncRNAs utilizando ferramentas diversas, o qual está disponível na forma de um *software* livre. Essa ferramenta foi utilizada no trabalho de Mostajo et al. (2020) para anotar os ncRNAs de 16 espécies de morcego. Adicionalmente, os autores combinaram as novas anotações das espécies de morcego com anotações de referência, de forma a torná-las mais completas e confiáveis (MOSTAJO et al., 2020).

Um desafio enfrentado tanto por Anthon et al. (2014) quanto por Mostajo et al. (2020) é a resolução de redundâncias. No caso da busca por homologia de sequência, múltiplos genes dos bancos de dados podem ser indicados como semelhantes a um único novo gene (ANTHON et al., 2014). Ferramentas diferentes também podem acabar por anotar o mesmo gene, com coordenadas genômicas ligeiramente diferentes (MOSTAJO et al., 2020). Nesses casos, é necessário escolher uma anotação para representar cada grupo de anotações redundantes. No trabalho de Anthon et al. (2014), métricas de alinhamento como porcentagem de identidade, cobertura e e-value foram usadas para classificar as anotações em três grupos: confiança alta, média ou baixa. Assim, entre as anotações conflitantes, a de mais alta confiança era escolhida. Já nas anotações de Mostajo et al. (2020), foi estabelecida uma lista de prioridades para as diferentes ferramentas usadas na anotação, a qual é usada para escolher as melhores anotações e resolver os conflitos.

1.2.1.4 Contaminação de Anotações

As ferramentas apresentadas anteriormente analisam sequências de DNA ou cDNA passadas como dados de entrada para identificar RNAs não-codificantes, sem considerar a origem dessas sequências. Num cenário ideal, onde as sequências incluídas na montagem de genoma utilizada pertencem todas à espécie original, isso não é um problema. No entanto, análises recentes de genomas disponíveis em bancos de dados públicos revelam

que grande parte deles possuem contaminações (FRANCOIS et al., 2020; STEINEGGER; SALZBERG, 2020). Cerca de 95% das sequências contaminadas ocorrem em genomas eucarióticos, pois eles são mais longos e fragmentados (STEINEGGER; SALZBERG, 2020). Numa busca por contaminação em 43 montagens de genomas de artrópodes, Francois et al. (2020) identificou que 28 delas estavam completamente livres de contaminações, enquanto 4 montagens continham mais de 150 regiões de codificação vindas de contaminantes. Apesar da frequência preocupante de genomas contaminados, os estudos de anotação de ncRNAs previamente citados (MOSTAJO et al., 2020; AL-TOBASEI; PANERU; SALEM, 2016; ANTHON et al., 2014) não buscaram possíveis contaminantes dentre os RNAs identificados.

1.2.2 Prevendo Funções de lncRNAs

Segundo o portal RNACentral (CONSORTIUM, 2019), que indexa dezenas de bancos de dados para ncRNAs, há mais de 1.299 milhões de lncRNA identificados nas mais variadas espécies. Existem bancos de dados que disponibilizam associações entre lncRNAs e doenças, como NONCODE (FANG et al., 2018) e lncRNAdb (QUEK et al., 2015). Porém, quando se trata de associações à termos de *Gene Ontology* (ASHBURNER et al., 2000), apenas 3,253 destes (0.25%) lncRNA estão anotados. Isso significa que há um vasto número desses genes anotados, porém tais anotações não descrevem quais são as funções moleculares, processos biológicos ou componentes celulares aos quais eles estão associados. Essa falta de associações funcionais dificulta a compreensão de como essas sequências de RNA interagem com o resto do organismo. Infelizmente, a determinação das funções dos lncRNAs através de métodos experimentais ainda é limitada por produzir poucos resultados e demandar conhecimentos a priori de possíveis mecanismos candidatos (SIGNAL; GLOSS; DINGER, 2016). Alguns estudos já trabalharam em metodologias para tentar identificar computacionalmente as funções desses lncRNAs. Porém há alguns problemas com as ferramentas já desenvolvidas: problemas de distribuição, falta de interfaces e especificidade para um único organismo.

Atualmente, enquanto este documento é escrito, *KATZLGO*, *LncRNA2Function* e outras ferramentas de identificar funções de lncRNAs (JIANG et al., 2015; ZHANG; ZOU; DENG, 2018; ZHANG et al., 2019; GUO et al., 2013) não estão mais disponíveis de forma pública para os usuários, o que é de fato mais uma dificuldade na caracterização de RNAs não-codificantes em organismos não-modelo. Uma opção disponível seria o *LNCRNA2GOA*, porém este *software* não possui uma interface gráfica ou de linha de comando para o usuário. Consequentemente, essa aplicação requer conhecimentos técnicos de programação para ser utilizada, o que é uma barreira relevante para aqueles que não sabem programar.

Outra opção disponível é o *software LncADeep*, uma ferramenta baseada em aprendizado profundo que foi treinada apenas com sequências de *H. sapiens*. As sequências de nucleotídeos dos lncRNAs tem baixa conservação entre espécies diferentes, o que limita

a aplicabilidade dessa ferramenta em organismos evolutivamente distantes da espécie humana. Já a ferramenta *KATZLGO* (atualmente indisponível) utiliza redes de interação de proteínas e lncRNAs para prever funções, mas os autores só recomendam que a ferramenta seja usada para prever funções em *H. sapiens* (ZHANG et al., 2019). Essa especificidade impede que descobertas sejam feitas em organismos muito distantes do *H. sapiens*. Para isso, é necessário uma estratégia de predição mais genérica.

Estudos recentes sobre as funções de ncRNAs se basearam na ideia de que genes com padrões de expressão similares tem maior probabilidade de estarem cooperando em vias biológicas relacionadas (EHSANI; DRABLøS, 2018; JIANG et al., 2015; ZHAO et al., 2015), um conceito chamado culpa-por-associação. Por isso esses trabalhos usam co-expressão (EISEN et al., 1998; LEE et al., 2004) para determinar suas funções.

1.2.2.1 Determinando Co-Expressões entre Genes Codificantes e lncRNAs

Uma vantagem da culpa-por-associação é que as funções são preditas baseando-se apenas nos níveis de expressão gênica e na anotação de genes codificantes do organismo sendo estudado, sem que seja necessário treinar os algoritmos (JIANG et al., 2015; EHSANI; DRABLøS, 2018; ZHAO et al., 2015). Isso se baseia no conceito de correlação, uma relação estatística de dependência ou associação entre duas variáveis. Neste caso, as duas variáveis são a lista de níveis de expressão de um gene codificante e a de um lncRNA. Comparando essas duas listas com uma métrica de correlação, é produzido um coeficiente de correlação: um número real ($\mathbb{R}$) que indica o quanto a expressão de um lncRNA está associada com a expressão de um gene codificante.

Já foram utilizadas diversas métricas para calcular esses coeficientes. Zhao et al. (2015) utilizou a métrica Spearman para calcular correlação entre genes e Jiang et al. (2015) fez uso da métrica Pearson com o mesmo objetivo. A ferramenta *LNCRNA2GOA* aplicou essas duas últimas, mas também adicionou as métricas Fisher e Sobolev (EHSANI; DRABLøS, 2018). Nesta ferramenta, a correlação entre dois genes só precisa ser alta em uma única métrica para ser considerada uma co-expressão. No entanto, essas 4 métricas não são formuladas pensando em dados não-lineares em geral.

Dependências não lineares existem na biologia e métricas para correlação não-linear podem ser mais adequadas (BRUNEL et al., 2010). Técnicas que aplicam métricas para calcular correlações em dados não-lineares, como *Maximal Information Coefficient* (MIC) e *Distance Correlation* (DC), também existem e já foram empregadas na criação de redes de regulação para genes codificantes (RESHEF et al., 2011; GUO et al., 2014), porém até o momento faltam trabalhos na literatura que apliquem essas métricas na predição de funções de lncRNA.

Uma dificuldade inerente ao uso de qualquer métrica de correlação é a escolha *thresholds* (critérios de corte) para decidir quais correlações caracterizam co-expressões. A escolha de *threshold* geralmente é um coeficiente mínimo, como em *LncRNA2Function*

(JIANG et al., 2015), que adota um valor mínimo de 0.9, ou escolhendo uma quantidade fixa de correlações com os melhores coeficientes, como em *LNCRNA2GOA* (EHSANI; DRABLøS, 2018), que escolhe as 250 melhores. Em ambos os casos, a escolha desses critérios foi arbitrária: nenhum desses trabalhos faz uma análise sistemática de como obter correlações melhores, comparando critérios mais relaxados e mais restritivos. Selecionar poucas correlações pode excluir resultados importantes, enquanto incluir correlações demais pode prejudicar a predição ao incluir muitos falsos positivos.

Ao avaliar a performance de suas metodologias, estudos anteriores se limitaram a uma avaliação qualitativa da predição funcional de 5 lncRNAs humanos bem estudados, dada a baixa quantidade de lncRNAs com caracterização funcional disponível (EHSANI; DRABLøS, 2018; JIANG et al., 2015). Além disso, Ehsani e Drabløs (2018) fez a predição de funções para 352 genes codificantes humanos funcionalmente anotados, de forma a comparar a performance das diferentes métricas. A performance foi calculada determinando, para as co-expressões selecionadas por cada métrica, suas respectivas similaridades semânticas. As métricas de similaridade semântica exercem a função de comparar a anotação funcional de dois genes, resultando num número que indica o quão parecidas são as funções dos dois (MINA, 2019). Ao calcular, para todas as anotações funcionais preditas, suas similaridades funcionais com as anotações reais, Ehsani e Drabløs (2018) puderam mensurar a qualidade das anotações criadas pela ferramenta. Nos resultados desse *benchmark*, a combinação de 4 métricas (Spearman, Pearson, Fisher e Sobolev) resultou em predições mais similares às funções reais do que usar apenas uma métrica.

2 Objetivos

O objetivo central deste trabalho é a criação de um *software* com rotinas para a anotação de ncRNAs no genoma de espécies pouco estudadas, que deve ser validado numa espécie modelo. Este pode então ser utilizado para encontrar ncRNAs de *A. gigas* e estudar os papeis deles no organismo do peixe. Os objetivos específicos são listados abaixo:

a) Criar um *pipeline* de anotação de ncRNAs em genomas, que deve incluir a detecção de ncRNAs usando homologia de sequência e de estrutura, como já foi implementado em ferramentas disponíveis atualmente;
   - Incluir certas funcionalidades novas: uma estratégia de anotação de lncRNAs inspirada em Al-Tobasei, Paneru e Salem (2016), combinação das novas predições com dados de referência, resolução automática de anotações redundantes e remoção de contaminantes;
   - Anotar os ncRNAs de uma espécie modelo, de forma a validar quantas das anotações criadas por essa *pipeline* correspondem às informações de referência atuais;

b) Comparar o quão confiáveis são os conjuntos de genes co-expressos encontrados pelas diferentes métricas de correlação;

c) Criar um *pipeline*, baseado na estratégia de culpa-por-associação, para prever funções de lncRNAs utilizando qualquer combinação entre as métricas Spearman, Pearson, Fisher, Sobolev, MIC e DC;
   - Avaliando a qualidade de predições funcionais usando *thresholds* mais restritivos e mais relaxados, determinar quais devem ser os critérios padrão dessas métricas;

d) Ambos os *pipelines* devem incluir interfaces de linha de comando e serem facilmente configuráveis;

e) Disponibilizar a ferramenta como um software livre no repositório público *GitHub* (github.com), para garantir o acesso de qualquer usuário e possibilitar futuras contribuições;

f) Estudo dos ncRNAs de *A. gigas*:
   - Utilizar dados disponíveis de genoma e transcriptoma para identificar ncRNAs nesta espécie e remover possíveis contaminantes;
   - Prever funções para os lncRNAs encontrados;
   - Buscar homologias entre os ncRNAs encontrados e os ncRNAs de outras espécies;

- Classificar os padrões de expressão dos lncRNAs nos diferentes tecidos de *A. gigas*;
- Buscar lncRNAs relacionados ao crescimento e desenvolvimento deste espécie;
- Buscar lncRNAs que possam servir de marcadores genéticos para determinação do sexo do Pirarucu;

3 Metodologia

3.1 Funcionamento do *RNA-Gatherer*

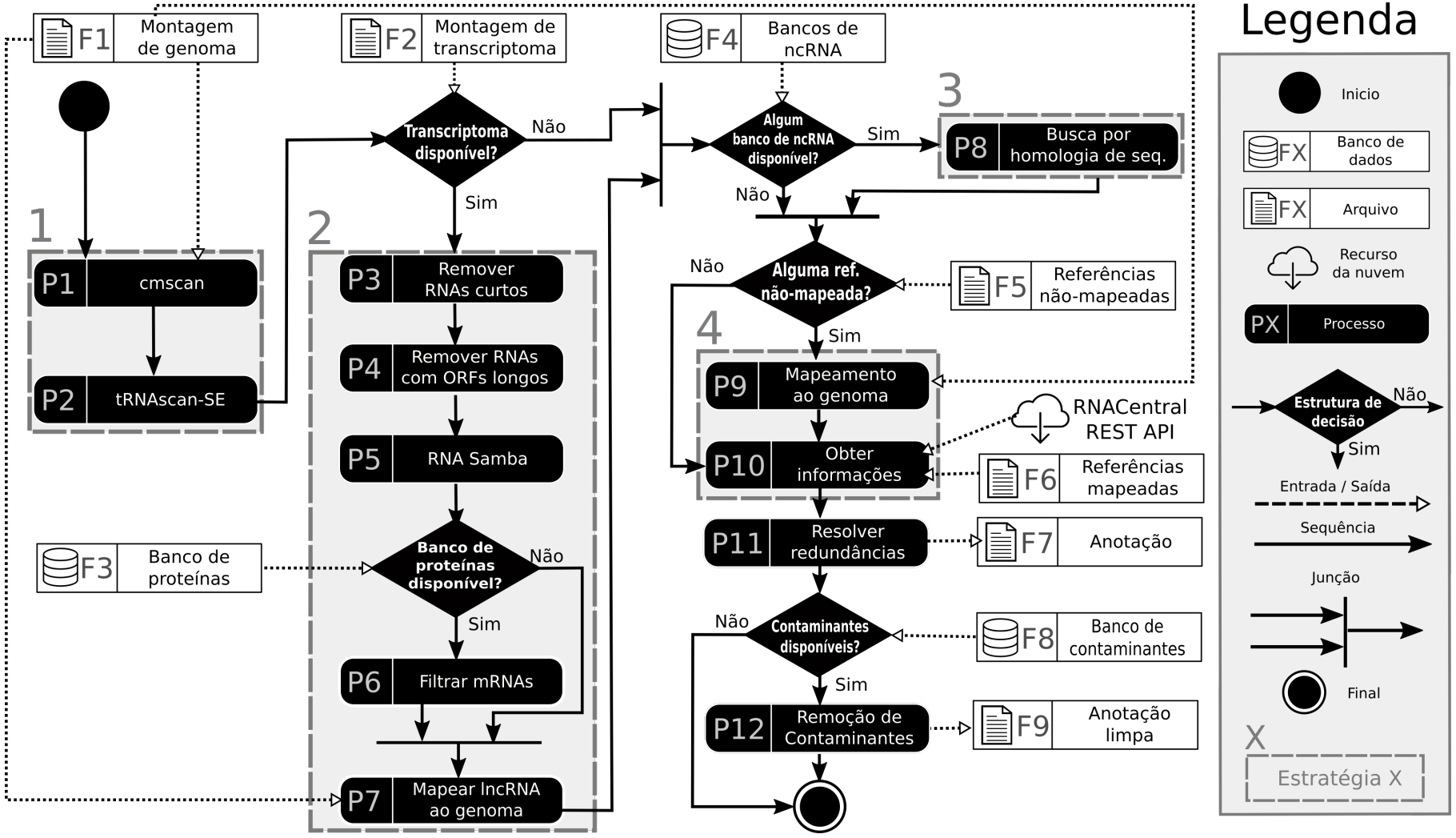
RNA-Gatherer é um pacote de ferramentas para estudo de ncRNAs que contém dois *pipelines* separados: *Explorer* e *Prophet*. A descrição do funcionamento deles é feita nas sessões seguintes. O *Explorer* faz anotação de ncRNAs num genoma, provendo suas coordenadas genômicas, tipos, famílias e funções conhecidas. Já o *Prophet* usa estatística para prever novas funções de lncRNAs, assinalando termos de Gene Ontology, e pode ser usado para encontrar as funções dos lncRNAs preditos pelo *Explorer*.

3.1.1 *Explorer*

A sequência de passos para caracterizar ncRNAs é resumida na Figura 2. Os seus arquivos de entrada são o genoma da espécie (Figura 2, F1), a montagem do transcriptoma (Figura 2, F2), um banco de dados de sequências de proteínas (Figura 2, F3), bancos de dados de ncRNA (Figura 2, F4), sequências de referência sem coordenadas genômicas (Figura 2, F5), coordenadas genômicas de regiões gênicas anotadas (Figura 2, F6) e um banco de dados de possíveis sequências contaminantes (Figura 2, F8). A utilização de muitos tipos de dados diferentes é um cenário ideal, que nem sempre é possível quando se trabalha com espécies pouco estudadas. Portanto, o *Explorer* foi desenhado para funcionar mesmo quando os dados de entrada citados anteriormente não estiverem disponíveis, com exceção do genoma, que é o único obrigatório. Assim, a ferramenta se torna mais flexível, podendo se adaptar a projetos que tem tipos diferentes de dados disponíveis.

A anotação é feita por 4 estratégias diferentes: (1) Identificação abrangente de ncRNAs no genoma por modelos de covariância (Figura 2, P1-2); (2) Predição de lncRNAs no transcriptoma (Figura 2, P3-7); (3) Anotação por homologia de sequências (Figura 2, P8); e (4) integração de dados de referência (Figura 2, P9-10). Essas estratégias são seguidas de uma fase de redução de redundâncias (Figura 2, P11) e uma de remoção de contaminantes (Figura 2, P12).

Figura 2 – Diagrama de Atividade do *Explorer*



Além do que é definido no padrão UML de Diagrama de Atividades, a figura define (na legenda) marcações específicas para arquivos de entrada/saída e recursos de nuvem. Fonte: os autores.

3.1.1.1 Estratégia 1 - Identificação Abrangente de ncRNAs no Genoma por Modelos de Covariância

Nessa estratégia (Figura 2, processos P1 a P2)), dois *software* baseados em modelos de covariância são usados: *Infernal* (NAWROCKI; EDDY, 2013) e tRNAscan-SE (CHAN; LOWE, 2019). Inicialmente, o módulo *cmscan* (P1) do *Infernal* é usado para predizer ncRNAs no genoma (F1), detectando as 3016 famílias do RFAM (KALVARI et al., 2018). Então, tRNAs são preditos pelo *tRNAscan-SE* (P2), que provê informações específicas sobre tRNA como isótipos e anticódons.

3.1.1.2 Estratégia 2 - Predição de lncRNAs no Transcriptoma

A segunda estratégia (Figura 2, processos P3 a P7) é um *pipeline* de anotação de lncRNA baseada em Al-Tobasei, Paneru e Salem (2016); recebendo como entrada uma montagem de transcriptoma (F2). Ele começa aplicando um filtro de comprimento, descartando transcritos mais curtos que 200nt (P3). Logo depois, a ferramenta *TransDecoder.LongORFs* (HAAS et al., 2013) é aplicada para remover transcritos com ORFs de comprimento superior a 150aa (P4).

Em P5, RNA Samba (CAMARGO et al., 2020) é usado para classificar transcritos como sendo lncRNA ou não, de forma a manter apenas os lncRNA. Transcritos remanescentes são alinhados (P6) a um banco de dados de proteínas provido pelo usuário (F3), usando a implementação do algoritmo BLASTX do software *Diamond* (BUCHFINK; XIE; HUSON, 2015). Transcritos alinhados a uma proteína com um e-value ≤ 0.0001 são descartados.

Todas as sequências restantes são mapeadas ao genoma (F1), usando o alinhador *Minimap2* (LI, 2018), para encontrar suas coordenadas genômicas, com cobertura e identidade $\geq 95\%$. Apenas sequências com coordenadas genômicas encontradas são consideradas como lncRNA e passadas à frente.

3.1.1.3 Estratégia 3 - Anotação por Homologia de Sequências

A próxima estratégia (Figura 2, processo P8) é realizar uma anotação pesquisando por homologia de sequências. Cada sequência nos bancos de dados de ncRNA (F4) é alinhada no genoma, usando *Minimap2*. Sequências alinhadas com cobertura e identidade $\geq 95\%$ são incluídas na anotação como ncRNAs.

3.1.1.4 Estratégia 4 - Integração de Dados de Referência

Sequências de referência que não possuem coordenadas genômicas providenciadas pelo usuário (F5) também são mapeadas contra o genoma usando *Minimap2* (P9), aplicando um filtro de cobertura e identidade $\geq 92\%$. Esse filtro tem um valor mínimo reduzido, comparado aos mapeamentos anteriores, pois essas já são sequências identificadas na

mesma espécie. As anotações destas, que agora tem coordenadas, os resultados da pesquisa por homologia (estratégia 3) e as referencias que já tinham coordenadas genômicas (F6) são submetidas ao processo P10, no qual informações sobre elas são obtidas usando a API REST pública do portal RNACentral (CONSORTIUM, 2019).

3.1.1.5 Remoção de Redundâncias e Contaminantes

Após as diferentes estratégias estarem completas, os resultados delas são integrados através de resolução de redundâncias. O *software gffcompare* (PERTEA et al., 2016) é usado para agrupar as anotações em grupos de redundância (P11). Então, um único representante de cada grupo é escolhido de acordo com uma lista de prioridades. Nessa lista, foi decidido priorizar anotações com mais respaldo em dados conhecidos e dar menos prioridade para aquelas que são apenas predições.

Assim, a primeira opção é escolher os ncRNAs com coordenadas de referencia (F6), seguidos dos ncRNAs conhecidos cujas coordenadas foram determinadas em P9. Depois, as anotações por homologia de sequência (P8), pois estas encontram genes já conhecidos e caracterizados em outras espécies. A próxima prioridade é dada aos resultados das ferramentas que usam modelos de covariância, que predizem ncRNAs que podem ser bastante diferentes no nível de sequência dos já conhecidos, mas que também anotam estes em famílias conhecidas. Dentre as duas ferramentas deste tipo, o *tRNAscan-SE* (P2) tem mais prioridade que o Infernal (P1), pois trás informações específicas sobre tRNA, que não são informadas pelo *Infernal*. A última prioridade é dada às anotações do *pipeline* de lncRNA (P7), que são apenas as predições filtradas da ferramenta RNA Samba.

Se houverem múltiplos ncRNAs no mesmo grupo com a mesma prioridade, aquele com o maior comprimento é escolhido. Após isso, para cada ncRNA com uma família RFAM já determinada, termos GO são associados baseado na anotação funcional da família. O resultado final é uma anotação de ncRNAs (F7), incluindo muitos RNAs de tipos diferentes; todos com coordenadas no genoma.

Caso um banco de dados de contaminantes (F8) esteja disponível, a anotação é limpa buscando por ncRNAs que poderiam vir de contaminantes. As sequências anotadas são alinhadas ao banco de dados (P12), usando *Minimap2*, e qualquer uma tendo um alinhamento com identidade e cobertura $\geq 90\%$ é descartada. O resultado é uma anotação limpa (F9).

3.1.1.6 Implementação

Este *pipeline* foi desenvolvido através da linguagem Python 3.6, combinando o paradigma de Programação Estruturada com Orientação a Objetos. O arquivo do *pipeline*, chamado *explorer.py*, é uma interface de linha de comando que utiliza os parâmetros passados pelo usuário para instanciar a classe *gatherer.Pipeline*, que encapsula a lista de todos os passos descritos na Figura 2. Esses passos são funções definidas em sub-modulos do

módulo *gatherer*. Cada uma dessas funções é responsável por executar os *software* externos necessários, verificar se os resultados desejados foram produzidos e salvar esses resultados num formato adequado para que os próximos passos possam fazer uso deles. Ao longo de todos os passos, as bibliotecas *numpy* (HARRIS et al., 2020) e *pandas* (REBACK et al., 2020) são utilizadas para representar os dados e realizar operações matemáticas diversas. Os endereços de ferramentas, arquivos de dados necessários e APIs são configuradas por um arquivo chamado *config.json*, no formato JSON (ECMA, 2013). Neste arquivo, o usuário pode inserir as localizações dos bancos de ncRNAs, proteínas e de contaminantes.

Os lncRNAs preditos por este *pipeline* podem ter suas funções preditas usando um *pipeline* específico para tentar determinar essas funções, descrito a seguir.

3.1.2 *Prophet*

Assim como foi apontado na introdução, para tentar realizar uma predição eficiente de funções, seria melhor descobrir o quão eficientes os diferentes parâmetros de *threshold* são em identificar pares de genes com similaridade funcional. Seis métricas diferentes foram incluídas, para serem avaliadas e comparadas: Pearson, Spearman, Fisher (EHSANI; DRABLøS, 2018), Sobolev (EHSANI; DRABLøS, 2018), *Maximal Information Coefficient* (RESHEF et al., 2011) e *Distance Correlation* (GUO et al., 2014). Usando a anotação funcional dos genes de uma espécie, a métrica de similaridade semântica SimGIC (PESQUITA et al., 2008), implementada pela ferramenta FastSemSim (MINA, 2019), é calculada para cada par de genes que contenham pelo menos 3 anotações funcionais. Em seguida, para uma métrica de correlação, os Níveis de Confiança são calculados de acordo com as equações 3.1-3.11, separadamente para as ontologias de Função Molecular, Processo Biológico e Componente Celular.

3.1.2.1 Definição dos Níveis de Confiança e seus Respectivos *Thresholds*

Neste trabalho é usada a expressão Nível de Confiança (NC) para se referir a um valor médio de similaridade semântica. Aqui, é definida uma lista de NCs e, para cada métrica de correlação, é verificado em qual valor de *threshold* cada NC é alcançado. Isso é feito para podermos comparar as métricas entre si. Através desses cálculos, é possível saber quais NCs são alcançados pelas diferentes métricas e exatamente em quais *thresholds*, facilitando a comparação entre critérios mais relaxados e mais restringentes.

Nesta seção é descrito o processo de calcular os Níveis de Confiança para uma determinada métrica e ontologia. *GenesAnotados* se refere a todos os genes com associações a termos GO. *AnotaçõesFuncionais* é uma função que retorna a lista de termos GO com os quais um gene está associado. Como mostra a Equação 3.1, a análise só é feita em genes

com três ou mais associações.

$$GenesAnotados = \{a \mid \forall a \in Genes, |AnotaçõesFuncionais(a)| \geq 3\} \tag{3.1}$$

As comparações de métricas de correlação ou similaridade semântica são feitas entre pares de genes, como definido na Equação 3.2.

$$Pares = \{(a, b) \mid \forall a \forall b \in GenesAnotados, a \neq b\} \tag{3.2}$$

Neste passo (Equação 3.3), os coeficientes de correlação são calculados para os pares de genes, usando uma métrica m.

$$Coeficientes(m) = \{(a, b, m(a, b)) | \forall (a, b) \in Pares\} \tag{3.3}$$

A função *Filtrar* (Equação 3.4) verifica se um coeficiente de correlação, calculado por uma métrica de correlação, é o suficiente para passar num filtro. Para as métricas Sobolev e Fisher, o valor vai passar se for menor ou igual ao filtro. Para as outras métricas, o coeficiente precisa ser maior ou igual.

$$Filtrar(valor, filtro, m) = \begin{cases} valor \leq filtro & \text{se m} \in \{Sobolev, Fisher\} \\ valor \geq filtro & \text{caso contrário} \end{cases} \tag{3.4}$$

Em *CoExp* (abreviação de "co-expressões", Equação 3.5), os *Coeficientes* de uma métrica são filtrados, usando um *threshold*. Os que puderem passar pelo filtro são considerados co-expressos.

$$CoExp(filtro, m) = \left\{ (a, b, c) \;\middle|\; \begin{array}{l} \forall (a, b, c) \in Coeficientes(m), \\ Filtrar(c, filtro, m) \end{array} \right\} \tag{3.5}$$

A próxima função, *SimilaridadesSemânticas* (Equação 3.6), toma um conjunto de pares de genes e retorna suas similaridades semânticas, de acordo com a métrica SimGIC, implementada pelo software fastsemsim (MINA, 2019).

$$SimilaridadesSemânticas(coefs) = \{SimGIC(a, b) \mid \forall (a, b, c) \in coefs\} \tag{3.6}$$

Esta função combina as últimas duas. Em *MédiaSS* (Abreviação de Média de Similaridade Semânticas, Equação 3.7), dado um filtro e uma métrica de correlação, as similaridades semânticas são calculadas para todos os pares de genes com coeficientes de correlação suficientes para passar no filtro. Então, a média dessas similaridades semânticas é calculada.

$$MédiaSS(filtro, m) = Média(SimilaridadesSemânticas(CoExp(filtro, m))) \tag{3.7}$$

A função *ThsAltos* (Equação 3.8, abreviação de *Thresholds* Altos) serve o propósito de criar, para uma métrica de correlação, um conjunto de filtros (aqui chamados de *thresholds*), cuja MédiaSS resultante é igual ou superior a um determinado Nível de Confiança. Isso é feito usando cada coeficiente de correlação como um possível *threshold* e então calculando a similaridade semântica média das co-expressões resultantes. Se o NC nunca é alcançado, a função retorna um conjunto vazio.

$$ThsAltos(NC, m) = \left\{ c \;\middle|\; \begin{array}{l} \forall (a, b, c) \in Coeficientes(m) \\ \text{e } MédiaSS(c, m) \geq NC \end{array} \right\} \tag{3.8}$$

Em *Threshold* (Equação 3.9), o conjunto de ThsAltos é calculado e o menos restritivo é selecionado. Isso significa retornar o *threshold* mais alto para Sobolev ou Fisher, e o menor para as outras métricas. Isso é feito pois, se mais de um *threshold* numa mesma métrica resulta em conjuntos de co-expressões com similaridades semânticas equivalentes, é melhor escolher o critério menos restritivo, pois vai trazer mais resultados. Se a função *ThsAltos* retornar um conjunto vazio (a métrica de correlação não foi capaz de alcançar o NC requerido), esta função vai indicar isso com a palavra 'None'.

$$Threshold(NC, m) = \begin{cases} \text{'None'} & \text{se } ThsAltos(NC, m) = \emptyset \\ Max(ThsAltos(NC, m)) & \text{se } m \in \{Sobolev, Fisher\} \\ Min(ThsAltos(NC, m)) & \text{se } m \notin \{Sobolev, Fisher\} \end{cases} \tag{3.9}$$

NCsParaBuscar (Equação 3.10) é a lista de similaridades semânticas médias que serão buscadas nos resultados das métricas. É uma lista de 40 valores, começando em 0.025 e indo até 1.0, com um passo de 0.025.

$$NCsParaBuscar = \{0.025, 0.050, 0.075, ..., 1.0\} \tag{3.10}$$

Finalmente, em *NíveisDeConfiança()* (Equação 3.11), para uma dada métrica, todos os NCsParaBuscar são buscados. Para cada um, o *threshold* que o alcança é calculado, usando a função Threshold, definida anteriormente.

$$NíveisDeConfiança(m) = \{(NC, Threshold(NC, m) \mid \forall NC \in NCsParaBuscar\} \tag{3.11}$$

O processo descrito nesta sessão pode necessitar de muitos recursos computacionais, por isso ele foi otimizado sendo dividido em quatro *scripts* diferentes, todos presentes no repositório público do projeto. Primeiramente todos os coeficientes de correlação e similaridades semânticas entre genes são calculados e armazenados em arquivos. Isso é feito, respectivamente, pelos *scripts calc_all_correlations.py* e *analysis/calc_ss.py*. Depois um *script* chamado *analysis/process_ss_and_corr.py* combina todos esses dados

de correlação e similaridade em uma única base de dados. Finalmente, os dados são usados para calcular os *thresholds* equivalentes a cada Nível de Confiança com o *script analysis/find_best_thresholds.py*.

3.1.2.2 Passos do *Prophet*

Em suma, o *pipeline Prophet* é composto por quatro passos, como ilustrado na Figura 3, isto é: cálculo de coeficientes de correlação (A1), filtragem de correlações usando um Nível de Confiança (A2), listagem de possíveis funções (A3) e enriquecimento usando P-Valor ajustado por FDR (A4). Além dos arquivos de entrada mostrados na figura, há dois parâmetros principais: o Nível de Confiança a ser usado e a lista de métricas de correlação para calcular.

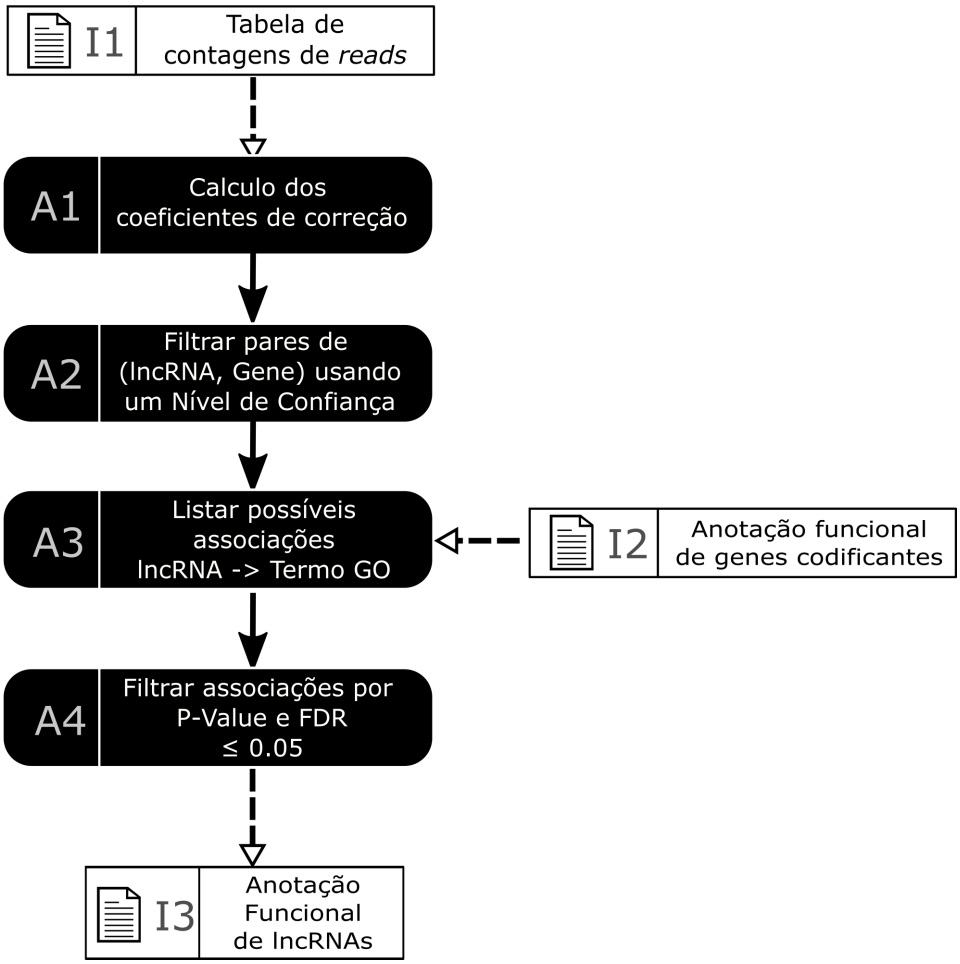
Este *pipeline* pode usar qualquer combinação das 6 métricas listadas na sessão anterior. As correlações são calculadas usando as contagens de expressão gênica (I1), para as métricas selecionadas, de forma paralela e usando a quantidade de núcleos de processamento disponíveis na máquina atual. Um Nível de Confiança é passado como parâmetro e, como os *thresholds* que atingem ele já foram pré-calculados, o *Prophet* obtém os *thresholds* para cada métrica, usando eles para filtrar as correlações (A2). Dessa forma são selecionadas as co-expressões.

O próximo passo é listar as possíveis funções de cada lncRNA (A3). Para cada co-expressão entre um lncRNA e um gene codificante, o lncRNA recebe como possíveis funções as que estão anotadas no gene codificante (I2). Depois, cada uma dessas possíveis associações entre uma função e um lncRNA é testada (A4), usando o teste hipergeométrico (EHSANI; DRABLøS, 2018; JIANG et al., 2015). Os P-Valores são ajustados para FDR usando o teste de Benjamini-Yekutieli (BENJAMINI; YEKUTIELI, 2001). As associações são selecionadas escolhendo aquelas com P-Valor e FDR ≤ 0.05. Esse conjunto estatisticamente relevante de associações é a anotação funcional de lncRNAs (I3).

3.1.2.3 Implementação

Este *pipeline*, assim como o anterior, foi desenvolvido com a linguagem Python. O pacote *scipy* (VIRTANEN et al., 2020) é utilizado para as métricas Pearson e Spearman. Para as métricas DC e MIC são utilizados os pacotes *dcor* (CARREñO, 2020) e *minepy* (ALBANESE et al., 2013), respectivamente. As demais métricas, Fisher e Sobolev, foram implementadas em Python baseando-se na implementação em R feita por Ehsani e Drabløs (2018). Os coeficientes de correlação são calculados com essas métricas de forma concorrente, através da criação de subprocessos para operar em subconjuntos dos dados de entrada. As bibliotecas *obonet* (HIMMELSTEIN et al., 2020) e *networkx* (HAGBERG; SWART; CHULT, 2008) foram utilizadas para navegar o grafo do *Gene Ontology* e as relações entre termos das anotações. Os P-Valores das associações preditas são calculados com *scipy* e então corrigidos para FDR com o pacote *statsmodels* (SEABOLD et al., 2017).

Figura 3 – Diagrama de Atividade do *Prophet*



Fonte: os autores

3.2 Experimentos

3.2.1 Validação do *RNA-Gatherer*

3.2.1.1 Dados Utilizados

O *pipeline Explorer* foi validado na montagem GCRm38 do genoma da espécie *Mus musculus* (usado como F1, Figura 2). Sequencias de transcritos dessa espécie (usadas como F2, Figura 2), incluindo 85490 mRNAs, 28767 lncRNAs, e 7656 ncRNAs de outros tipos foram obtidas do RefSeq (O'LEARY et al., 2016). Sequências de proteínas (usadas como F3, Figura 2) foram obtidas do banco de dados de proteínas Non-Redundant (NR) (NON-REDUNDANT..., 2020). Um conjunto composto por ncRNAs de *M. musculus* conhecidos, mas sem coordenadas genômicas, foi obtido do banco de dados RNACentral (versão 14) para serem localizados no genoma (usados como F5, Figura 2). Como é uma espécie modelo, uma anotação por homologia de sequência (Estratégia 3) acabaria por

trazer todos os genes da espécie que já são anotados, os quais substituiriam as predições. Algo semelhante aconteceria se incluíssemos referencias mapeadas (F6, Figura 2). Por isso, a anotação por homologia de sequência só será aplicada no cenário prático da anotação de *A. Gigas*.

Para testar e avaliar o *Prophet*, também foram usados dados de *M. musculus*, isto porque essa espécie conta com um vasto número de lncRNAs funcionalmente anotados, superior as outras espécies. Para o teste nessa espécie, foram obtidas 26 amostras de RNA-seq a partir da biblioteca ArrayExpress (PARKINSON et al., 2005) E-MTAB-2801. Essas amostras foram limpas usando o software *Sickle* (<https://github.com/najoshi/sickle>), usando as configurações padrão. Depois, as amostras foram usadas para quantificar a expressão gênica dos lncRNA e genes codificantes de *M. musculus*, usando a ferramenta Salmon (PATRO et al., 2017), também sob configurações padrão. Essas contagens foram normalizadas por TPM (Transcritos Por Milhão), gerando uma tabela de contagens (usada como I1, Figura 3). As anotações funcionais de lncRNA foram obtidas pelo portal RNACentral e as dos genes codificantes foram obtidas do banco de dados MGI 6.14 (usadas como I2, Figura 3).

3.2.1.2 Busca por Combinações de Parâmetros com Melhor Performance

Níveis de Confiança foram calculados usando os genes funcionalmente anotados de *M. musculus*. As contagens de expressão dos genes codificantes e lncRNA foram usadas para avaliar a performance do *Prophet* e *LNCRNA2GOA*. Uma busca exaustiva pelas melhores combinações de métricas de correlação foi executada: Para cada NC, todas as possíveis combinações de 1 a 5 métricas que alcançam tal NC foram tentadas; A performance de cada uma dessas predições foi estimada comparando elas com a anotação funcional referência de lncRNAs, obtida do RNACentral.

Este trabalho define duas métricas para avaliar a qualidade de uma predição funcional: Q1 e Q2. Q1 é apenas a porcentagem de associações de referência presentes na predição funcional. Q2 é a porcentagem de associações preditas que são relacionadas a uma associação referência. Essa relação entre associações é definida da seguinte forma: uma associação entre um lncRNA X e um termo GO Y é considerada como "relacionada"a uma referência quando, dentre os termos GO associados a X na referência, existe um com um caminho até Y no grafo de *Gene Ontology*.

Essas duas métricas são usadas para calcular a qualidade da predição, que é uma média ponderada com um peso de 1 para Q1 e 5 para Q2. Após calcular as qualidades, as predições são agrupadas de acordo com o parâmetro NC usado para criar elas. Em cada grupo, a predição com a mais alta qualidade é escolhida como representante de qual a melhor performance que pode resultar de usar um Nível de Confiança.

3.2.2 Estudo dos ncRNAs de *A. gigas*

Tanto o *pipeline Explorer* quando o *Prophet* foram empregados na tarefa de detectar e anotar ncRNAs na espécie *A. Gigas*.

3.2.2.1 Dados Utilizados para Estudar *A. gigas*

A montagem GCA_900497675.1 do genoma de *A. gigas* foi obtida do Genbank. Amostras *paired-end* de RNA-Seq de oito tecidos foram adquiridas do *NCBI BioProject* PRJNA665688 (MARTINS et al., 2020) e dois outros tecidos foram obtidos do *NCBI BioProject* PRJNA353913. A partir do RNACentral, foi obtido um banco de dados de todos os ncRNA atualmente indexados pelo portal, para realizar anotação por homologia de sequências. Destas sequências do RNACentral, aquelas referentes a bactérias e vetores foram separadas como um banco de dados de contaminantes. Além disso, o conjunto dos ncRNAs de *A. gigas* conhecidos, mas sem coordenadas genômicas, foi adquirido para realizar mapeamento de referências. Para executar a anotação funcional usando culpa-por-associação, realizada pelo *Prophet*, as anotações dos genes codificantes da espécie também foram obtidas do estudo de Martins et al. (2020).

3.2.2.2 Anotação de ncRNA

As leituras nas bibliotecas RNA-Seq de *A. gigas* foram limpadas usando *Sickle* sob configurações padrão e então montadas usando *Trinity* v.2.6.5 (HAAS et al., 2013), também com parâmetros padrão. Sequências redundantes foram removidas com *CD-HIT-EST* (FU et al., 2012), com parâmetro de identidade (-c) de 0.9. *Kallisto* (BRAY et al., 2016) foi usado para estimar a expressão gênica dos transcritos e aqueles com menos de 5 contagens foram removidos, deixando um transcriptoma de sequências não-redundantes e expressas. Essa transcriptoma (usado como F2, Figura 2), o genoma de *A. gigas* (usado como F1, Figura 2), o banco de proteínas NR (usado como F3, Figura 2), as sequências de ncRNA do RNACentral (usadas como F4, Figura 2), os ncRNA não mapeados de *A. gigas* (usado como F5, Figura 2) e o banco de contaminantes (usado como F8, Figura 2) foram usados como entrada para o *Explorer*, anotando o genoma de Pirarucu.

Possíveis ncRNAs homólogos, em outras espécies, foram buscados. Para fazer isso, as sequências de ncRNA do Pirarucu foram alinhadas às sequências do RNACentral usando *Minimap2*. Os alinhamentos foram classificados como sendo de similaridade baixa, média ou alta aplicando 3 *thresholds* de identidade e cobertura diferentes: 0.8, 0.85 e 0.9, respectivamente. Para cada ncRNA de Pirarucu, apenas a sequências com o melhor alinhamento (maior identidade) a ele é considerado.

3.2.2.3 Análise de Expressão de lncRNA

Transcritos de lncRNA foram classificados de acordo com a especificidade de suas expressões nos diferentes tecidos. Cada lncRNA encontrado no Pirarucu foi classificado em uma das seguintes categorias:

a) **Não Expresso**: O valor TPM em todas as amostras é menor que 1.0;

b) **Tecido-Específico**: Em uma das amostras de um tecido, a contagem TPM ($\geq$ 1.0) é 5 vezes maior que nas amostras de todos os outros tecidos;

c) ***Housekeeping***: O lncRNA não é tecido-específico, mas as contagens TPM são superiores a 1.0 em todas as amostras de todos os tecidos;

d) **Expressão Mista**: A molécula de lncRNA não se encaixa nas três categorias anteriores;

A expressão diferencial entre amostras de macho e de fêmea dos lncRNA também foi avaliada, usando os pacotes DESeq (ANDERS; HUBER, 2013), DESeq2 (LOVE; HUBBER; ANDERS, 2014), EBSeq (LENG et al., 2013) e edgeR (ROBINSON; MCCARTHY; SMYTH, 2010). Os resultados desses pacotes foram filtrados seguindo os seguintes critérios: O lncRNA não foi classificado como "Não Expresso", Fold-Change ≥ 1.5 e P-Valor < 0.05. Apenas transcritos presentes nos resultados filtrados de 2 ou mais pacotes foram considerados como diferencialmente expressos (DE).

3.2.2.4 Predição e Enriquecimento de Funções de lncRNA

As contagens de expressão gênica estimadas com Salmon e a anotação dos genes codificantes (MARTINS et al., 2020) foram usadas como entradas para executar o *Prophet*, sob parâmetros padrão, e assim prever funções para todos os lncRNA encontrados em *A. gigas*. No novo conjunto de lncRNAs anotados funcionalmente, foram buscados RNAs associados a termos GO de crescimento e maturação sexual (as listas desses termos estão disponíveis no Apêndice C).

O enriquecimento funcional de conjuntos de lncRNAs foi realizado com GOATOOLS (KLOPFENSTEIN et al., 2018). Cada enriquecimento precisa de 3 dados: um conjunto de genes, uma anotação de funções e uma população (lista de todos os genes). Para esses enriquecimentos, a predição funcional foi a anotação e o conjunto de todos os lncRNAs encontrados foi a população. Os conjuntos de lncRNAs foram os diferencialmente expressos, *housekeeping* e os grupos de tecido-específicos. Funções enriquecidas foram filtradas de acordo com alguns critérios: FDR < 0.01; um mínimo de 3 lncRNAs anotados com essa função; e um Fold-Change (como definido na Equação 3.12) ≥ 1.5 entre a frequência da função no conjunto de genes e a frequência na população inteira. Além disso, funções relacionadas a pigmentação e melanina foram buscadas nos resultados do enriquecimento

de lncRNAs diferencialmente expressos entre sexos.

$$|Log2(FoldChange(FrequênciaNoConjunto, FrequênciaNaPopulação))| \quad (3.12)$$

4 Resultados

4.1 Avaliação do *RNA-Gatherer* e das Métricas de Correlação

4.1.1 Performance do *Explorer*

O genoma de *Mus musculus* é aquele com mais anotações funcionais de lncRNA e também inclui anotações para uma variedade de tipos diferentes de ncRNA (CONSORTIUM, 2019), por isso essa espécie foi escolhida para testar o software. A anotação do genoma do camundongo (Tabela 1) resultou em 76343 ncRNA diferentes (incluindo lncRNA, tRNA, miRNA, rRNA, snRNA, sRNA, scRNA e ribozima) e 3468 elementos cis-reguladores (Cis-reg). A maioria dessas anotações eram das ferramentas baseadas em modelos de covariância, *cmscan* e *tRNAscan-SE*. tRNA foi o tipo de ncRNA mais abundante, seguido por miRNA e Cis-reg. Como esperado de uma ferramenta que tenta detectar ncRNAs de famílias conhecidas, o *cmscan* de *Infernal* foi capaz de prever um número maior de ncRNAs do que todas as outras ferramentas.

Tabela 1 – Resumo das anotações de ncRNA em *Mus musculus*.

Tipos de ncRNA	**Número de ncRNAs**	**Famílias RFAM**	***cmscan***	**Mapeamento de Referências**	***RNA Samba***	***tRNAscan-SE***
Gene	76343 (∼95.65%)	638	39179	18	96	37050
tRNA	37060 (∼46.43%)	1	10	0	0	37050
Sem subtipo	25445 (∼31.88%)	8	25445	0	0	0
miRNA	10027 (∼12.56%)	237	10027	0	0	0
snRNA	3192 (∼4.0%)	229	3186	6	0	0
rRNA	278 (∼0.35%)	9	278	0	0	0
lncRNA	267 (∼0.33%)	147	160	11	96	0
sRNA	51 (∼0.06%)	3	51	0	0	0
ribozima	22 (∼0.03%)	4	22	0	0	0
scRNA	1 (∼0.0%)	0	0	1	0	0
Cis-reg	3468 (∼4.35%)	49	3468	0	0	0
Sem subtipo	3433 (∼4.3%)	27	3433	0	0	0
IRES	20 (∼0.03%)	17	20	0	0	0
frameshift element	15 (∼0.02%)	5	15	0	0	0
Outros	1 (∼0.0%)	0	0	1	0	0
Todos	79812	687	42647	19	96	37050

As anotações estão separadas por tipo de ncRNA e fonte de anotação. Fonte: os autores

Dependendo do tipo de ncRNA, a porcentagem deles que são compatíveis com

uma referência (Tabela 2) difere enormemente. A maioria das ribozimas, IRES, snRNA e rRNA estavam presentes na anotação de referência. Apesar dos transcritos testados para lncRNAs terem vindo do banco de dados *RefSeq*, apenas 61% delas estão presentes no *RNACentral*. Isso mostra como as anotações de diferentes bancos de dados podem diferir entre si.

Tabela 2 – Comparação da anotação de *Mus Musculus* com a anotação *RNACentral* de referência

Tipos de ncRNA	**Compatíveis com uma referência (%)**
Gene	14220 (18.626%)
Sem sub-tipo	457 (1.796%)
tRNA	8457 (22.82%)
snRNA	2675 (83.803%)
scRNA	0 (0.0%)
sRNA	0 (0.0%)
ribozima	22 (100.0%)
rRNA	236 (84.892%)
miRNA	2210 (22.04%)
lncRNA	163 (61.049%)
Cis-reg	26 (0.75%)
Sem sub-tipo	7 (0.204%)
frameshift element	0 (0.0%)
IRES	19 (95.0%)
Outros	1 (100.0%)

Um ncRNA compatível é aquele, na nova anotação, que corresponde às coordenadas de um ncRNA na anotação de referência. Para ser compatível, o novo gene ncRNA deve estar localizado dentro das coordenadas de um ncRNA de referência, com um máximo de 5% de seu comprimento fora do ncRNA de referência. Fonte: os autores.

As taxas de falsos positivos e falsos negativos de cada etapa da detecção de lncRNA são mostradas na Tabela 3. Elas foram calculadas apenas para transcritos *RefSeq* cujas informações de tipo de RNA (mRNA, lncRNA, etc.) estavam disponíveis. Esses resultados mostram que a taxa de falsos positivos (FP) diminui a cada etapa, de 75.621% (na etapa 1) para cerca de 4.592% (na etapa 4). Por outro lado, a taxa de falsos negativos (FN) aumenta de 0.003% (na etapa 1) a 59.916% (na etapa 4). Além disso, comparando o FP do *RNA Samba* e do *Explorer* (33.788% e 4.592%, respectivamente), foi observada uma redução considerável do FP quando o *Explorer* foi usado. Apenas sete (0.05%) dos lncRNAs relatados na etapa final estavam codificando sequências de mRNA, sendo o restante dos falsos positivos outros tipos de ncRNA. É uma melhoria quando comparada à previsão feita apenas com o emprego do *RNA Samba*, que incluiu 14% dos RNAs codificantes.

4.1.2 Performance das Métricas de Correlação e do *Prophet*

Para avaliar as métricas de correlação, uma biblioteca de dados de RNA-Seq para amostras de *M. musculus* foi pré-processada, aparando bases de baixa qualidade (Apêndice

Tabela 3 – Taxas de falsos positivos e negativos para RNA Samba e as várias etapas de detecção de lncRNA

Passos	**lncRNA**	**Outros ncRNA**	**mRNA**	**Falsos Positivos (%)**	**Falsos Negativos (%)**
Conjunto original de transcritos	28767	7656	85490	76.4	0.0
Apenas *RNA Samba Explorer* - Estratégia de predição de lncRNAs:	19499	7190	2761	33.788	32.217
1. Remoção de RNAs curtos	28766	3760	85471	75.621	0,003
2. Filtragem de ORFs longos	18773	3304	3679	27.112	34.741
3. Testagem com *RNA Samba*	17285	3015	591	17.261	39.914
4. Filtragem de mRNAs com NR	11531	548	7	4.592	59.916

Fonte: os autores.

A, Tabela 9). Usando essas amostras, as contagens de expressão dos genes de *M. musculus* foram estimadas. Os níveis de similaridade semântica média foram calculados para os diferentes *thresholds* de várias métricas de correlação (Figura 4). A métrica MIC atingiu um nível muito baixo de similaridade, enquanto as co-expressões de DC alcançaram níveis de similaridade mais altos do que outras métricas. Independentemente da métrica utilizada, a similaridade semântica média alcançada na ontologia CC (Componente Celular) foi maior do que nas outras duas ontologias, MF (Função Molecular) e BP (Processo Biológico). As tabelas com os resultados do Nível de Confiança para cada ontologia e métrica são apresentadas no Apêndice B - Níveis de Confiança e Estatísticas Funcionais de Predição. Como exemplo, a Tabela 4 exibe parte dos NCs para o Componente Celular.

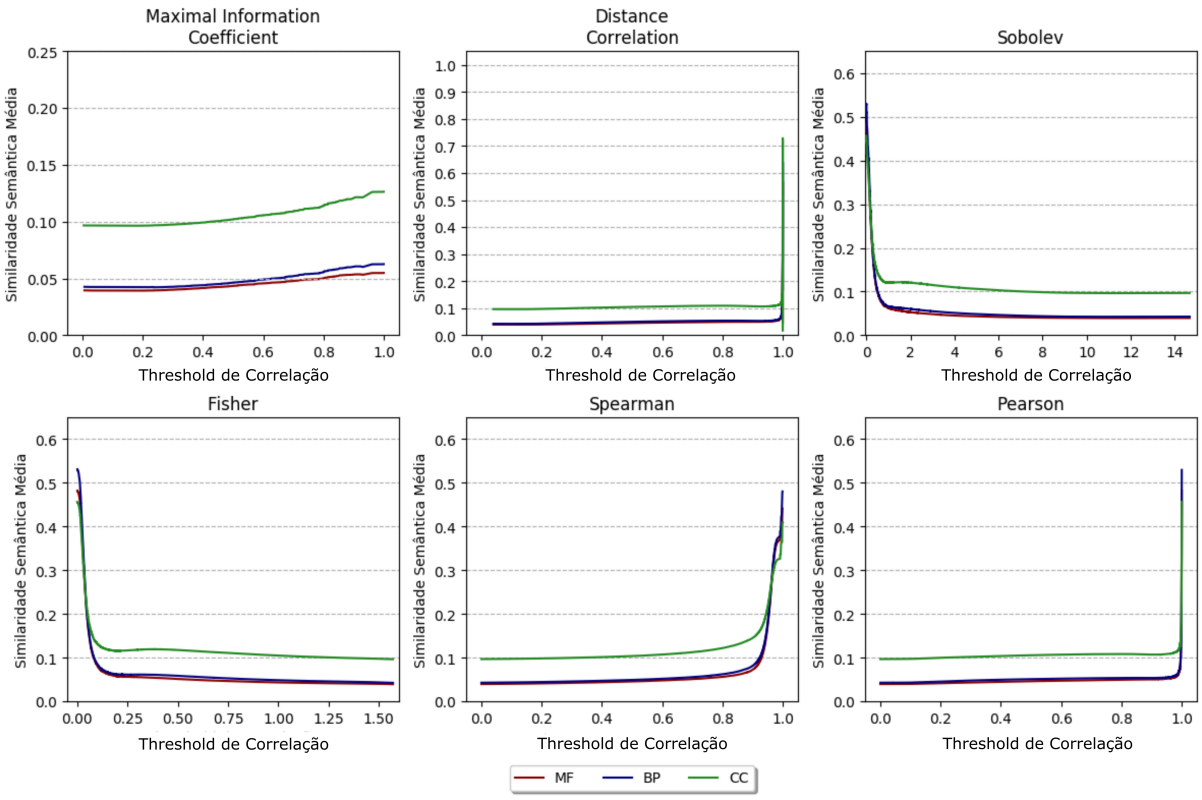
Tabela 4 – Algumas das colunas na tabela de Níveis de Confiança da ontologia Componente Celular

Similaridade Semântica Média	[...]	0.125	0.15	0.175	0.2	0.225	[...]
Nome do Nível de Confiança	[...]	4	5	6	7	8	[...]
MIC	[...]	0.96123	None	None	None	None	[...]
DC	[...]	0.99602	0.99833	0.99893	0.99921	0.99938	[...]
SOB	[...]	0.73897	0.46462	0.3577	0.29845	0.25137	[...]
FSH	[...]	0.11711	0.0755	0.06043	0.05113	0.04456	[...]
SPR	[...]	0.82004	0.90944	0.93436	0.94477	0.95229	[...]
PRS	[...]	0.99318	0.99709	0.99813	0.99864	0.9989	[...]

Cada coluna representa os *thresholds* necessários para as métricas atingirem um determinado Nível de Confiança. O valor 'None' significa que uma métrica não foi capaz de atingir um NC. Fonte: Apêndice B.

Como não conseguia atingir a maioria dos NCs, a métrica MIC não foi usada para testar predições funcionais. Essas predições, feitas em diferentes combinações de métricas

Figura 4 – Similaridade semântica média dos conjuntos de co-expressões



Similaridade semântica média estimada para os conjuntos de co-expressões selecionados por diferentes *thresholds*, em cada uma das 6 métricas de correlação. Os cálculos foram realizados separadamente para cada ontologia, que são representadas por diferentes linhas coloridas no gráfico. Fonte: os autores.

e níveis de confiança, resultaram em uma vasta gama de valores para as métricas de previsão de qualidade Q1 e Q2; conforme definido na subseção "Busca por Combinações de Parâmetros com Melhor Performance"da Metodologia. Na Figura 5 podemos observar como Q1 (eixo X) vai de quase 0% a quase 100%, enquanto Q2 (eixo Y) varia de cerca de 60% a 100%. Predições com valores Q2 muito altos tendem a predizer menos associações de referência e aquelas que predizem a maioria das associações de referência nunca têm um valor Q1 superior a 90%. O software *LCRNA2GOA* foi aplicado para prever a Função Molecular e o Processo Biológico nos mesmos dados de entrada, alcançando uma qualidade semelhante aos NCs mais baixos do *Prophet*, mas com um valor Q2 ligeiramente menor.

O único parâmetro que difere entre as previsões feitas no mesmo Nível de Confiança é a combinação de métricas de correlação usada. Assim, para cada NC, a combinação que resulta na predição de maior qualidade (marcadores de diamante na Figura 5) foi selecionada como o conjunto padrão de métricas de correlação. Para selecionar NCs padrão, o valores de Q2 de cada uma foram dispostos na Figura 6, que permite a visualização de como a qualidade da predição varia conforme os critérios são mais restritivos ou relaxados.

A variação em Q2 tem uma curva parecida em todas as 3 ontologias, uma curva semelhante a uma parábola, com um mínimo global no Nível de Confiança 17, 18 ou 20.

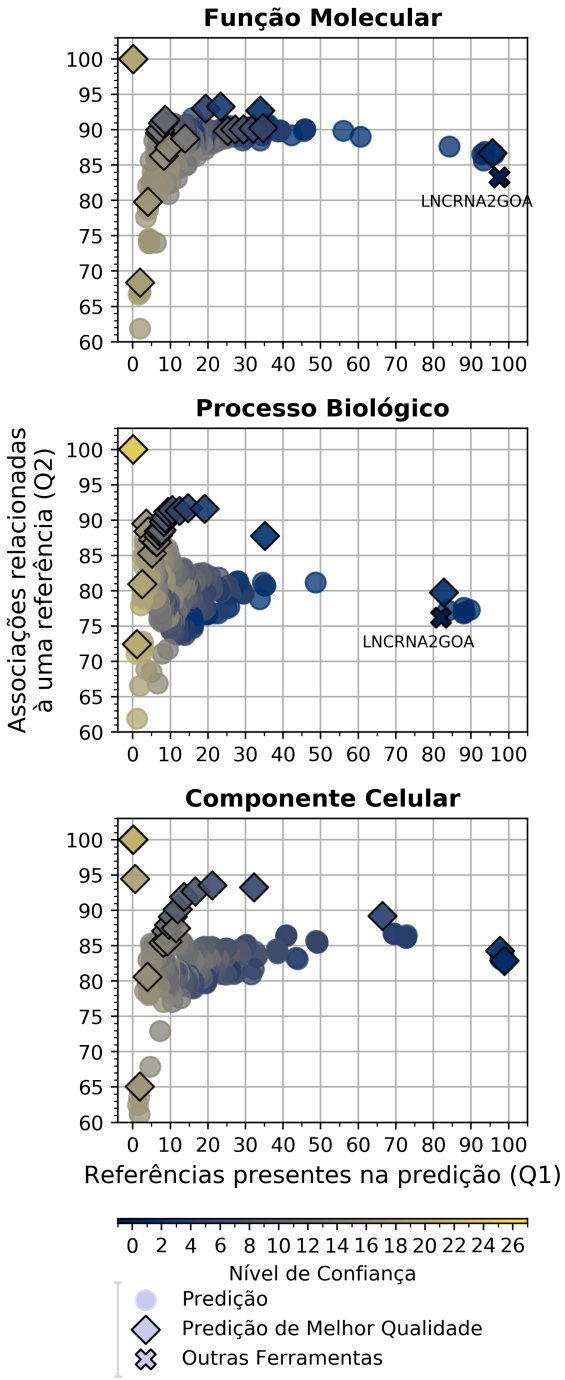
O Nível de Confiança com o valor Q2 mais alto à esquerda do mínimo global foi definido como configuração 'normal' (o padrão) e aquele com o valor mais alto à direita foi definido como a configuração '*high*' (Tabela 5). Dentre as diversas combinações de métricas, a DC apresentou o melhor desempenho nos NC elevados, chegando a um valor de até 100% Q2. Além disso, a maioria das combinações de múltiplas métricas incluiu DC. Para os NC mais baixos, a métrica de Spearman apresentou o melhor desempenho. Para todas as ontologias, a métrica usada como configuração normal foi Spearman e DC foi a configuração alta.

Tabela 5 – Configurações 'Normal' e '*High*' para cada ontologia

Ontologia	**'Normal'**			***'High'***		
	Métrica	**Nível de Confiança**	***Threshold***	**Métrica**	**Nível de Confiança**	***Threshold***
Função Molecular	Spearman	3	0.929	DC	19	1.0
Processo Biológico	Spearman	4	0.936	DC	21	1.0
Componente Celular	Spearman	5	0.909	DC	19	1.0

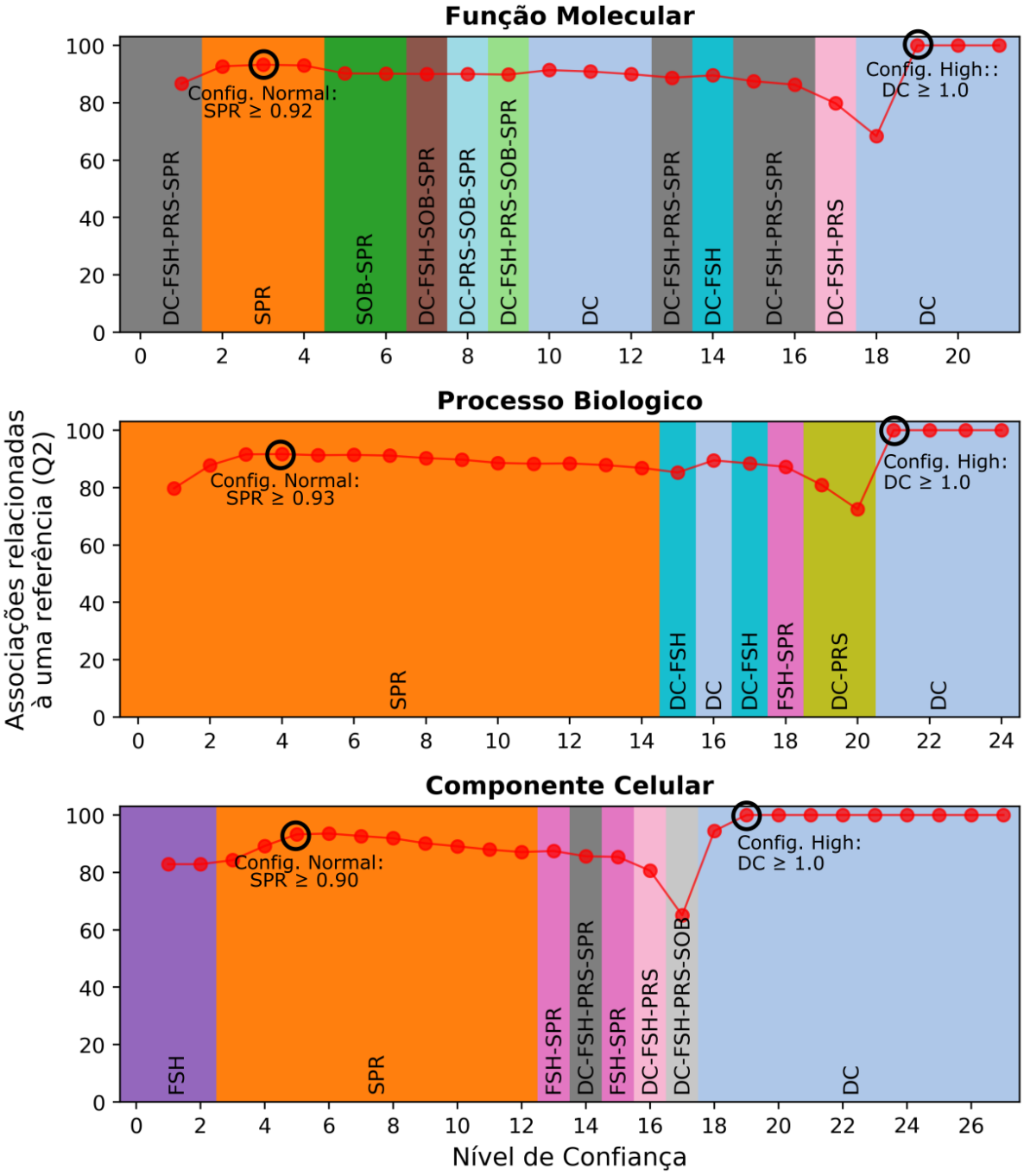
Fonte: os autores.

Figura 5 – Todas as predições funcionais de lncRNA em *Mus Musculus*



A melhor predição de cada NC está marcada como um **diamante**, diferente das outras predições, marcadas como **círculos**. Predições feitas com o mesmo NC estão marcadas com a mesma **cor**. A cor de cada NC está indicada na barra de cores abaixo dos três gráficos. A ferramenta ***LNCRNA2GOA*** não realiza predições para Componente Celular. Fonte: os autores.

Figura 6 – A predição de mais alta qualidade para cada Nível de Confiança



Para cada NC (**eixo X**), é mostrado o valor de Q2 (**eixo Y**) resultante da melhor combinação de métricas. **DC** = *Distance Correlation*; **SOB** = Sobolev; **FSH** = Fisher *Information*; **SPR** = Spearman; **PRS** = Pearson; **Q2** = a porcentagem de associações preditas que são relacionadas a uma associação de referência. A configuração '**normal**' aqui descrita é a padrão para rodar o *Prophet*. A *pipeline* também pode ser executada com a configuração '***high***', para resultados mais precisos (porém limitados). Fonte: os autores.

4.2 Anotação e Análise de ncRNAs em *A. Gigas*

Tabela 6 – Resumo das anotações de ncRNA em *A. gigas*.

Tipos de ncRNA	**Número de ncRNAs**	**Famílias RFAM**	***cmscan***	**Homologia de Sequência**	**Mapeamento de Referências**	***RNA Samba***	***tRNAscan-SE***
Gene	62723 (∼99.08%)	424	2929	221	12	59171	390
lncRNA	59209 (∼93.53%)	27	38	0	0	59171	0
miRNA	1522 (∼2.4%)	171	1468	54	0	0	0
snRNA	1238 (∼1.96%)	168	1231	7	0	0	0
tRNA	571 (∼0.9%)	1	18	153	10	0	390
sRNA	72 (∼0.11%)	25	71	1	0	0	0
rRNA	67 (∼0.11%)	9	60	5	2	0	0
Sem sub-tipo	31 (∼0.05%)	13	30	1	0	0	0
Ribozima	5 (∼0.01%)	4	5	0	0	0	0
Anti-senso	4 (∼0.01%)	4	4	0	0	0	0
Antitoxina	3 (∼0.0%)	1	3	0	0	0	0
CRISPR	1 (∼0.0%)	1	1	0	0	0	0
Cis-reg	565 (∼0.89%)	36	565	0	0	0	0
Sem sub-tipo	544 (∼0.86%)	26	544	0	0	0	0
IRES	7 (∼0.01%)	3	7	0	0	0	0
Frameshift element	5 (∼0.01%)	1	5	0	0	0	0
Leader	5 (∼0.01%)	4	5	0	0	0	0
Thermo-regulator	3 (∼0.0%)	1	3	0	0	0	0
Riboswitch	1 (∼0.0%)	1	1	0	0	0	0
Outros	18 (∼0.03%)	0	0	18	0	0	0
Intron	1 (∼0.0%)	1	1	0	0	0	0
Todos	63307	461	3495	239	12	59171	390

As anotações estão separadas por tipo de ncRNA e fonte de anotação. O tipo 'Sem sub-tipo' se refere a ncRNAs que tem são classificados como 'Gene', mas sem um tipo específico. 'cmscan' e 'tRNAscan-SE' fazem parte da Estratégia 1. 'RNA Samba' é o software do qual vem os resultados da Estratégia 2. As colunas 'Homologia de Sequência' e 'Mapeamento de Referências' se referem às Estratégias 3 e 4, respectivamente. Fonte: os autores

Dois conjuntos de dados RNA-Seq de *A. gigas* foram usados, PRJNA353913 e PRJNA665688. O primeiro possui perfis de expressão da pele e do fígado; enquanto o outro tem perfis de cérebro, pulmão, rim, músculo, coração e gônadas. As amostras de tecido muscular e cardíaco foram coletadas apenas de fêmeas e machos, respectivamente. Todos os outros tecidos possuem uma amostra de cada sexo. As amostras de Pele e Fígado foram coletadas de indivíduos juvenis.

Depois de limpas, as amostras de pele apresentaram conteúdo GC entre 46% a 48%; enquanto as dos demais tecidos apresentaram conteúdo GC entre 51.5% e 58%. Todas as amostras apresentaram índice de qualidade ≥ 25 em 99.38% das bases ou mais (Apêndice

A, Tabela 10). Conforme descrito antes, a biblioteca de amostras foi usada para montar um transcriptoma, que foi usado para prever os lncRNA.

RNAs não-codificantes de *A. gigas* foram anotados usando *Explorer*. Um total de 248 ncRNAs, dos quais 171 eram tRNAs, foram considerados como contaminantes e removidos da anotação final. A anotação final consiste de 62,723 ncRNAs de diversos tipos gênicos, 565 regiões cis-regulatórias, 18 ncRNAs de tipo desconhecido e um ncRNA intrônico (Tabela 6).

Em seguida, para buscar sequências conservadas, os transcritos de ncRNA anotados foram pesquisados em outras espécies. Em total, 706 dos ncRNAs tiveram similaridade com genes de outras espécies (Tabela 7). A espécie com mais sequências similares foi o Aruanã Dourado (298 ncRNAs), seguido por várias outras espécies de peixes ósseos, como Tetra-cego (29 ncRNAs), Peixe-zebra (27 ncRNAs) e Salmão Atlântico (17 ncRNAs). A lista completa de homologias está disponível no Apêndice D.

Tabela 7 – Anotações não-codificantes de *A. gigas* com homologia, agrupadas por espécie.

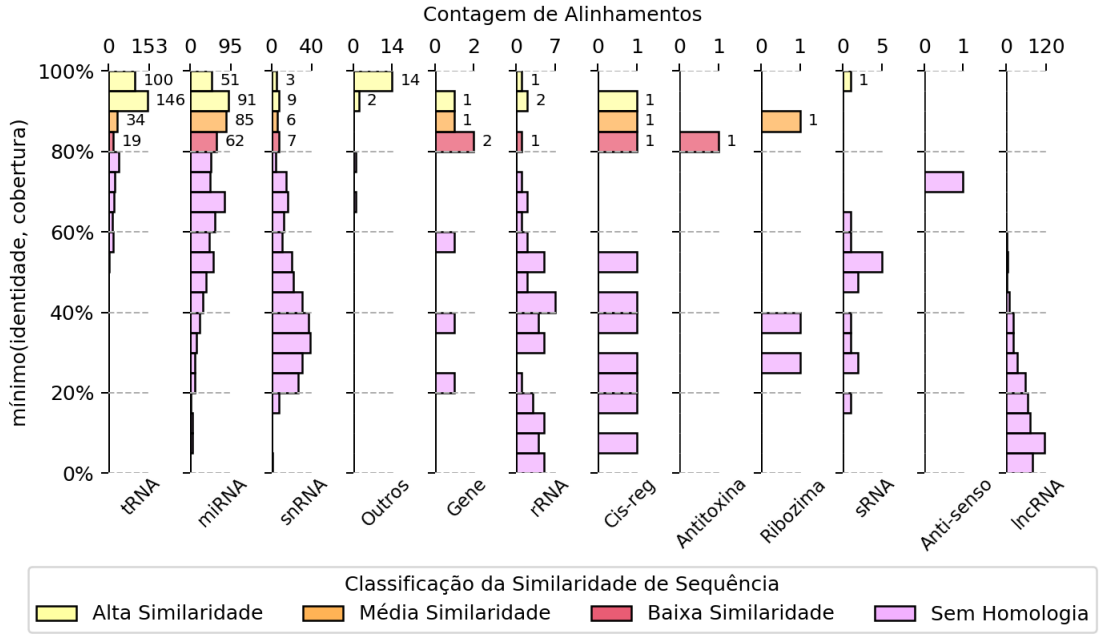
	Baixa Similaridade	**Média Similaridade**	**Alta Similaridade**	**Total**
Todas as homologias	128	140	438	706
Scleropages formosus (Aruanã Dourado)	50	86	162	298
Astyanax mexicanus (Tetra-cego)	2	3	24	29
Danio rerio (Peixe-zebra)	2	3	22	27
Salmo salar (Salmão atlântico)	6	2	9	17
Lepisosteus oculatus (Boca-de-jacaré)	1	6	7	14
Danionella translucida (NCBI:taxid623744)	2	1	11	14
Cyprinus carpio (Carpa)	2	1	10	13
Oreochromis niloticus (Tilapia do Nilo)	3	3	7	13
Outros (espécies com <13 sequências similares):	60	35	186	281
Vários peixes ósseos	21	14	110	145
Vários tetrápodes	23	18	71	112
Taxos menos frequentes (espécies com 10 ou menos sequências similares)	16	3	5	24

Espécies contendo menos que 13 sequências são agrupadas em um táxon mais elevado na sua linhagem. Espécies com 10 ou menos sequências similares foram agrupadas em uma única categoria. Fonte: os autores

Destes 706 RNAs, os tipos mais bem conservados foram os tRNAs e miRNAs, com 58.49% e 20.43% das predições sendo similares a outras espécies, respectivamente

(Apêndice D, tabela *By RNA type*). Apesar dos lncRNAs comporem a maior parte da anotação, nenhum teve uma possível homologia detectada (Figura 7). O mesmo ocorreu com alguns tipos de RNA pouco frequentes na anotação: anti-senso, CRISPR e intron autocatálitico.

Figura 7 – Histogramas das similaridades entre sequências não-codificantes de *A. gigas* e de outras espécies, separadas por tipo da sequência.



Assim como foi definido na metodologia: para ter **Alta Similaridade** (amarelo claro), um alinhamento precisa de no mínimo **0.9 de identidade e cobertura**; Para **Média** (laranja) e **Baixa similaridade** (vermelho), os valores mínimos são de **0.85 e 0.8**, respectivamente. Alinhamentos agrupados nas barras superiores representam uma similaridade quase exata entre uma anotação de *A. gigas* e uma sequência do RNACentral. Alinhamentos colocados nas barras inferiores tem similaridade e/ou cobertura muito baixas. Fonte: os autores.

A fim de prever as funções de RNAs longos não codificantes, através do *Prophet*, os lncRNAs previamente anotados tiveram suas expressões gênicas quantificadas por pseudo-alinhamento usando as amostras RNA-Seq de *A. gigas* (Tabela 10). O número médio de leituras mapeadas por amostra foi de 492,015.41, variando de 8,377 a 1,569,648. Entre os lncRNAs quantificados, 90.727% apresentaram contagens de leitura bruta ≥ 5 em pelo menos uma amostra. O lncRNA mais abundante foi Transcript_1165.0 (responsável por 1.12% de toda a expressão gênica), seguido por Transcript_32.0, Transcript_8043.0, Transcript_9692.0, Transcript_9379.0, Transcript_217908.0, Transcript_185363.0, Transcript_6987.0, Transcript_7136. 0, Transcript_185692.0. Entre os poucos lncRNAs detectados por *cmscan*, o mais expresso foi UFQX01000243.1_HOXA11-AS1_6_0, seguido por UFQX01005174.1_HOXA11-AS1_6_0, UFQX01002149.1_KCNQ1OT1_1_0, UFQX01000676.1_HOTTIP_3_0, UFQX01005169.1_Mico1_0, UFQX01001011.1_WT1-

AS_3_0, UFQX01000399.1_SOX2OT_exon1_0, UFQX01000256. 1_RMST_9_0, UFQX-01001481.1_WT1-AS_7_0 e UFQX01004564.1_TTC28-AS1_2_0.

4.2.1 Análise da Expressão e Funções de lncRNA em *A. Gigas*

Durante a análise dos padrões de expressão dos lncRNAs preditos, foi revelado que cerca de um terço dessas moléculas (19854 de 59209) são tecido-específicas (Tabela 8). Apenas 256 tiveram expressão ao longo de todos os tecidos, sendo classificados como RNAs de *housekeeping*, ou seja, reguladores ativos em todos os tecidos estudados. Destes, 6 estavam relacionados a funções de crescimento. Gônada, pele e cérebro foram os tecidos com mais lncRNAs diferencialmente expressos entre amostras de macho e fêmea (379, 289, e 109 genes, respectivamente), como mostrado na Tabela 8.

Tabela 8 – Análise dos padrões de expressão de lncRNAs identificados em *A. gigas*.

Classificação		**lncRNA na classificação**	**DE por sexo**	**Envolvidos com crescimento**	**Envolvidos com maturação**
Tecido-específico	Cérebro	7486	109	559	0
	Gonada	4807	379	3374	3
	Coração	354	38	24	1
	Rim	964	25	35	0
	Fígado	726	35	2	0
	Pulmão	1413	6	73	0
	Músculo	785	79	13	0
	Pele	3319	289	126	0
	Todos	19854	960	4206	4
Expressão mista		38434	2325	3957	20
Housekeeping		256	3	6	0
Não expressos		655	0	32	0

Fonte: os autores.

Em relação aos lncRNAs de maturação, como pode ser visto na Tabela 8, a maioria deles (20 de 24) apresentam um padrão de expressão misto. Os 4 restantes tecido-específicos: um lncRNA de coração (Transcript_191098.0) com maior expressão em amostras masculinas do que nas femininas; e três lncRNAs de gônada (Transcript_13888.0, Transcript_203363.0, Transcript_228651.0), os quais foram mais expressos em amostras femininas do que em masculinas. Onze dos lncRNAs de maturação (incluindo Transcript_228651) foram anotados com as funções "diferenciação do sexo feminino" (GO:0046660) e "desenvolvimento de características sexuais femininas primárias" (GO:0046545). No entanto, nenhum lncRNA foi associado com funções de maturação específicas de macho. Além disso, um dos três lncRNAs de gônadas (Transcript_13888.0) e um lncRNA de expressão mista (Transcript_284070.0) estavam diferencialmente expressos entre macho e fêmea. A lista completa de RNAs envolvidos na maturação está disponível na seção '*Involved in*

Maturation' do Apêndice C, classificados pelo número de funções de maturação com as quais foram anotados.

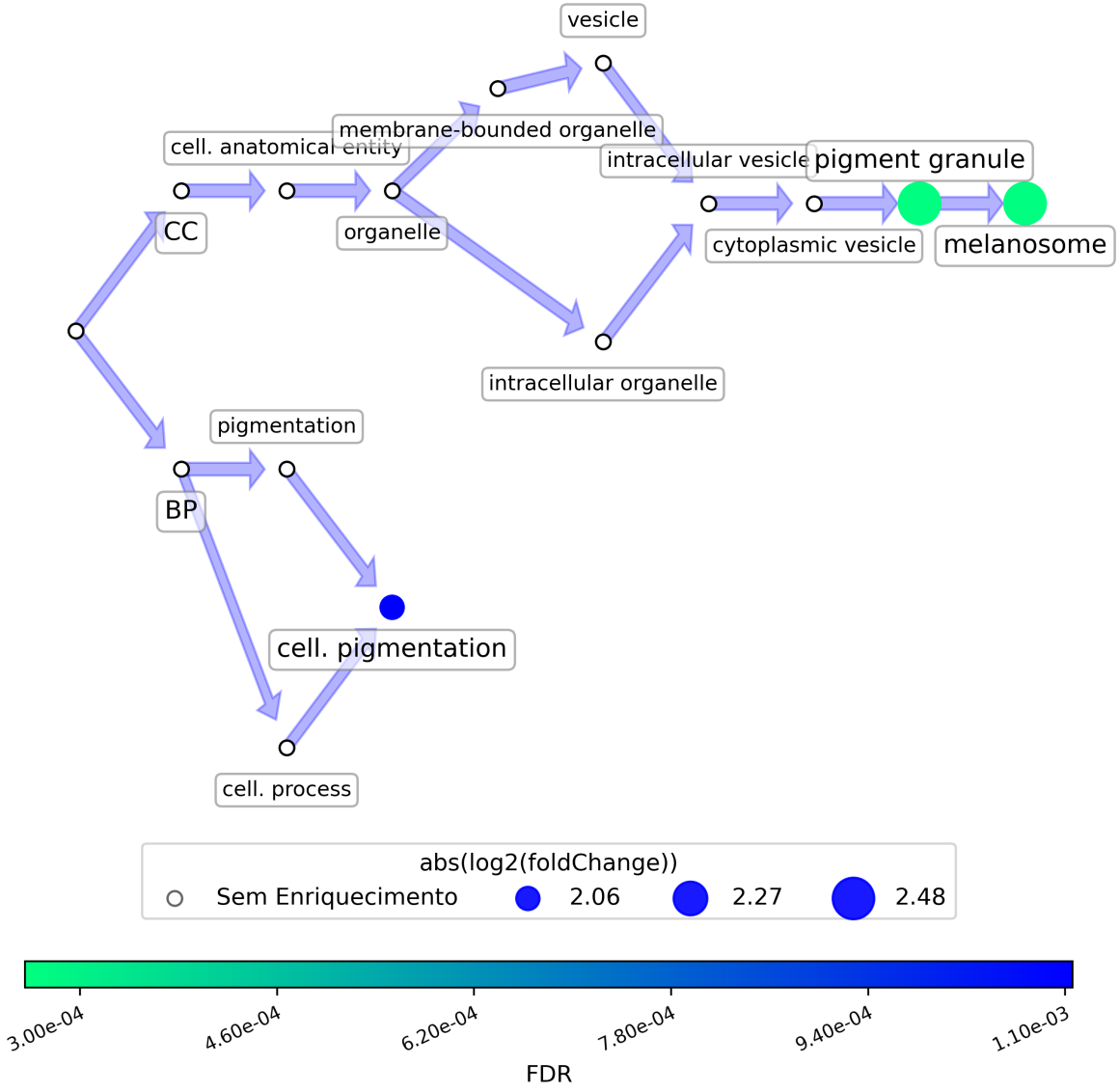
Conforme mostrado na Figura 8, os lncRNAs diferencialmente expressos por sexo estavam funcionalmente enriquecidos para os termos 'pigmentação celular' (GO:0033059), 'grânulo de pigmento' (GO:0048770) e 'melanossoma' (GO:0042470), os quais são diretamente relacionados à pigmentação da pele. Quatro desses transcritos DE (Transcript_116851.0, Transcript_80855.0, Transcript_141417.0 e Transcript_285759.0) foram anotados com tais termos e tiveram uma alta contagem de expressão na pele masculina, tendo também expressão zero na pele feminina. Esses ncRNA podem estar envolvidos em padrões de cores sexo-específicos.

Um total de 8,201 lncRNAs foram anotados com termos de crescimento, sendo 4,206 tecido-específicos, 3,957 com padrão de expressão misto, 32 não expressos o suficiente e 6 de *housekeeping*. Os seis RNAs do conjunto expressos em todos os tecidos foram associados a funções como 'crescimento' (GO:0040007), 'crescimento no desenvolvimento' (GO:0048589), 'regulação do crescimento' (GO:0040008) e 'crescimento celular' (GO:0016049).

Destes, quatro lncRNAs (Transcript_5968.0, Transcript_10586.0, Transcript_186108, Transcript_215159) foram anotados com termos relacionados ao controle do ciclo celular e à proteção contra danos ao DNA, incluindo processos biológicos como: 'envelhecimento' (GO: 0007568), 'célula morte' (GO:0008219), 'regulação do ciclo celular' (GO:0051726), 'reparo de DNA' (GO:0006281) e 'resposta à radiação' (GO:0009314).

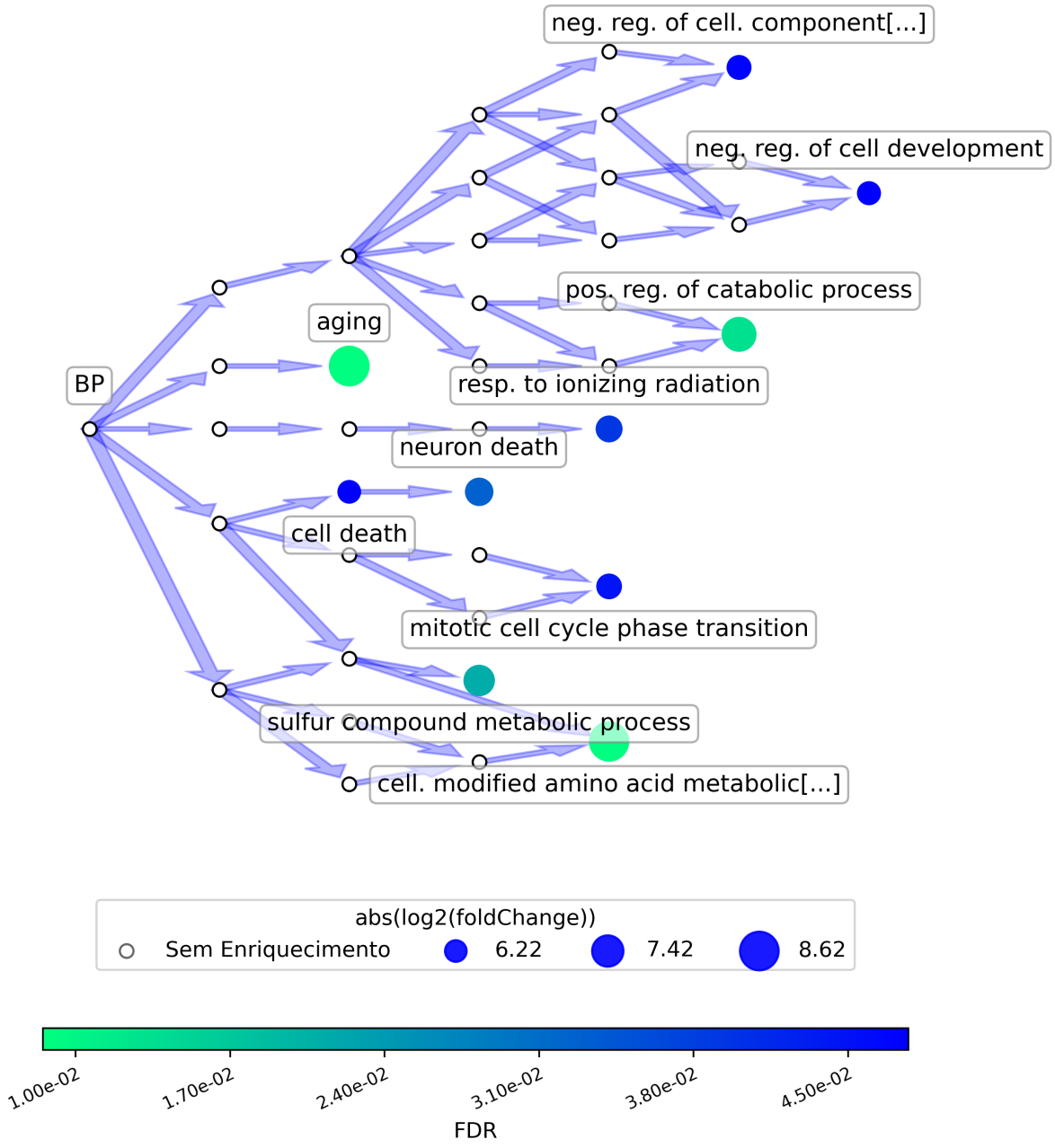
Além disso, os dois *housekeeping*-lncRNAs de crescimento restantes, Transcript_-209775.0 e Transcript_220598.0, estão correlacionados, respectivamente, à Semaforina-3B e RMBS3, proteínas conhecidas por defender as células contra tumores. Este conjunto de seis lncRNA relacionados ao crescimento com super-expressão em todos os tecidos foi submetido à análise de enriquecimento funcional, conforme representado pela Figura 9. A lista completa de RNAs envolvidos no crescimento está disponível na seção '*Involved in Growth*' do Apêndice C, classificados pelo número de funções de crescimento com as quais foram anotados.

Figura 8 – Grafo das relações entre os termos de pigmentação enriquecidos no conjunto de lncRNAs DE de *A. gigas*.



As **cores** dos círculos representam os FDR do enriquecimento funcional (cores mais quentes são melhores) e o **tamanho** do círculo representa o *fold-change*. **Setas** representam uma relação de tipo 'is_a' entre dois termos no grafo do *Gene Ontology*. A cor das setas é irrelevante. Fonte: os autores.

Figura 9 – Grafo das relações entre os termos GO enriquecidos, no conjunto de *housekeeping*-lncRNA associados ao crescimento.



Nomes de termos não-enriquecidos foram removidos da figura para simplificar a visualização. Os nomes de alguns termos enriquecidos foram encurtados, pois eram longos demais para serem mostrados na figura. Fonte: os autores.

5 Discussão

Recentemente, os RNA não-codificantes (miRNA, siRNA, piRNA e lncRNA, dentre outros) foram identificados como moléculas de RNA funcionais (ZHAO; SONG; WANG, 2016). No entanto, embora haja interesses crescentes em estudar essas moléculas, a maioria delas ainda não foi identificada em organismos não-modelo (JANNESAR et al., 2020; LEGEAI; DERRIEN, 2015). Embora espécies intimamente relacionadas a um organismo modelo possam ser anotadas usando homologia de sequência, esse não é o caso de organismos como *A. gigas*.

5.1 As Estratégias de Predição do *Explorer*

Para organismos como o Pirarucu, é preferível aplicar várias estratégias de anotação, a fim de produzir resultados mais completos. A maioria dos softwares para prever ncRNAs detecta apenas um único tipo de RNA. Portanto, quando os pesquisadores precisam anotar vários tipos de sequências não-codificantes, eles precisam desenvolver suas próprias maneiras de executar várias ferramentas e combinar seus resultados, desperdiçando um tempo precioso. Alguns estudos (MOSTAJO et al., 2020; ANTHON et al., 2014) já realizaram esse tipo de anotação de ncRNA em todo o genoma, mas não publicaram um software para realizar a anotação automaticamente. No entanto, há um pipeline disponível, GORAP (github.com/koriege/gorap, Riege & Marz, não publicado), que pode identificar ncRNAs usando modelos de covariância, homologia de sequência e por meio de ferramentas de predição específicas do tipo para tRNA, rRNA e CRISPR. Neste trabalho elaboramos um pipeline que, além de ter funcionalidades semelhantes, também pode prever lncRNAs, mesclar informações de referência fornecidas pelo usuário, obter informações de genes em repositórios públicos e remover possíveis contaminantes da anotação.

Embora o trabalho de referência para a estratégia de previsão de lncRNA (ALTOBASEI; PANERU; SALEM, 2016) do *Explorer* (Estratégia 2) tenha usado o CPC2 (KANG et al., 2017); o presente estudo utilizou a ferramenta RNA Samba, por ter demonstrado desempenho superior, utilizando dados de várias espécies diferentes, quando comparado a outros software (CAMARGO et al., 2020). Outra alteração feita a partir do trabalho de referência foi colocar a etapa de filtragem por BLASTX (6) no final do *pipeline*. Isso foi feito porque o alinhamento de todos os transcritos contra um grande banco de dados de proteínas pode levar muito tempo e recursos computacionais, portanto, é melhor alinhar menos transcritos após mais etapas de filtragem.

A segunda etapa da estratégia de predição de lncRNA do *Explorer*, que filtra ORFs longos, foi responsável por uma grande redução de mRNAs e outros ncRNAs. Porém, ORFs longos foram detectados em grande parte dos lncRNAs (34.7%, Tabela 3), o que fez com

que eles não fossem considerados lncRNAs, aumentando muito o número de falsos positivos. Outra etapa, a filtragem de alinhamentos BLASTX à proteínas, também aumentou o percentual de falsos positivos em 20%. Esses achados podem ser explicados pela biogênese do lncRNA; que pode ser transcrito a partir de regiões intergênicas, exônicas ou regiões distais de codificação de proteínas do genoma, pela enzima RNA-polimerase II (DHANOA et al., 2018).

Em comparação com nosso total final de 59%, testar as transcrições apenas com RNA Samba resultou em uma porcentagem de falsos negativos significativamente menor (32%, Tabela 3). No entanto, fazer isso resultou em uma taxa muito maior de falsos positivos (7.35 vezes mais FPs) do que a abordagem usada neste estudo. Nossa estratégia de predição de lncRNA foi projetada para minimizar, tanto quanto possível, o número de sequências codificantes sendo classificadas como lncRNAs (falsos positivos), mas em troca descarta muitos lncRNAs verdadeiros (falsos negativos).

Outra diferença entre o *Explorer* do *RNA-Gatherer* e outros pipelines de anotação de ncRNA é que a ferramenta desenvolvida por nossa equipe tem a capacidade de incluir anotações de referência na anotação final. Dessa forma, a anotação final é aprimorada pela integração de informações confiáveis. Além disso, como as informações de referência sempre têm prioridade sobre as novas anotações, as anotações de referência substituem automaticamente as previsões que correspondem a elas. Além disso, como nosso software baixa associações funcionais para ncRNAs conhecidos, o usuário não precisa procurá-los na Internet.

5.2 Níveis de Confiança e Predição de Funções para lncRNA

Mesmo dentre os vários tipos de ncRNA, os genes de lncRNA provaram ser sérios obstáculos para identificação e anotação; muitos lncRNAs foram identificados, mas apenas alguns deles já tiveram suas funções identificadas e suas estruturas de transcrição permanecem amplamente incompletas (LAGARDE et al., 2017; GUO et al., 2016). Algumas ferramentas para predição de função de lncRNA foram elaboradas (ZHANG; ZOU; DENG, 2018; GUO et al., 2013; JIANG et al., 2015) para organismos não-modelo, mas, no momento, não se encontram mais disponíveis de forma pública; e também há duas ferramentas que preveem funções em organismos modelo, como *H. sapiens* (ZHANG et al., 2019; YANG et al., 2018). Uma opção para previsão de função em organismos não-modelo é *LNCRNA2GOA* (EHSANI; DRABLøS, 2018), mas o uso desta ferramenta requer conhecimentos de programação.

Portanto, tendo em mente a importância de anotar bem os ncRNA em *A. gigas* (um organismo não modelo sem espécies próximas), neste estudo propomos uma nova ferramenta que inclui uma metodologia sistemática para predizer as funções dos lncRNA,

incluindo métricas para detectar co-expressões não-lineares, que estavam ausentes em trabalhos anteriores. Além disso, também usamos tal ferramenta para caracterizar o perfil dos lncRNA em *A. gigas*, sugerindo seus potenciais processos biológicos.

Em trabalhos anteriores sobre a predição da função de lncRNAs por co-expressão, não havia critérios claros para escolher quais parâmetros de filtragem são relevantes (EHSANI; DRABLøS, 2018; JIANG et al., 2015). Combinando isso com o fato de que os dados usados para avaliar as previsões eram muito limitados, não estava claro quanta relevância esse tipo de previsão tem. É por isso que, antes de proceder a testar a previsão de função em si, este projeto analisou a relação entre várias métricas de correlação com a similaridade funcional. Além das métricas utilizadas anteriormente na predição da função lncRNA, este trabalho também empregou as métricas de Coeficiente Máximo de Informação (MIC) e Correlação de Distância (DC), que visam detectar correlações não-lineares. A medida de similaridade semântica SimGIC (PESQUITA et al., 2008) foi usada para avaliar o quão semelhantes são as funções dos genes selecionados como co-expressos, dependendo do parâmetro de *threshold* usado. Como a Figura 4 mostrou, a semelhança semântica média entre os pares de genes selecionados como co-expressos pode variar muito, dependendo da ontologia e da métrica de correlação empregada.

As métricas cujos resultados mais diferiram entre si foram as para correlações não-lineares, MIC e DC. Em um comentário sobre o artigo do MIC (GORFINE; HELLER; HELLER, 2012), argumentou-se que esta métrica não é poderosa quando o tamanho da amostra não é grande. Neste estudo, um conjunto de 26 amostras de *Mus musculus* foi empregado para *benchmark* do *Prophet*, usando diferentes métricas de correlação e níveis de confiança. A métrica MIC foi a única a atingir apenas níveis baixos de confiança (consulte o Apêndice B), provavelmente devido a esse problema com amostras não muito numerosas. Em contraste, a métrica DC alcançou Níveis de Confiança mais altos do que outras métricas.

Devido ao seu desempenho reduzido, MIC foi a única métrica não usada para testar a predição de funções. Trabalhos anteriores com métodos de previsão de função semelhantes apenas previram processos biológicos e funções moleculares (EHSANI; DRABLøS, 2018; JIANG et al., 2015), mas nosso software provou ser capaz de prever com sucesso componentes celulares associados a lncRNAs. Conforme mostrado na Figura 5, onde cada previsão é representada como um círculo / diamante colorido, as qualidades das previsões feitas usando o mesmo NC (mesma cor do marcador) variam muito. No entanto, apenas as previsões com NC baixo (azuis mais escuros) recuperam 80% ou mais das associações de referência (valor Q1). Da mesma forma, apenas as previsões com NCs muito altos (verdes claras) foram capazes de ter um percentual de associações previstas relacionadas à referência (valor Q2) superior a 95%. Portanto, algumas previsões foram muito precisas, mas recuperaram muito pouca informação, enquanto outras recuperaram quase todas as associações de referência, mas tiveram resultados menos precisos.

A seleção das predições com os melhores desempenhos é baseada principalmente na porcentagem de associações relacionadas à alguma referência (valor Q2). No entanto, quando duas previsões têm valores Q2 muito semelhantes, é escolhida aquela que recupera mais associações de referência (valor Q1). Isso é feito atribuindo a Q2 um peso maior no cálculo da qualidade, conforme descrito no capítulo de metodologia. As melhores previsões de cada Nível de Confiança são destacadas na Figura 5 e seus valores Q2, bem como suas combinações de métricas, são mostradas na Figura 6. Este último gráfico mostra a prevalência das métricas de Spearman e DC. O valor Q2 mais alto foi alcançado pela DC em seu *threshold* máximo (1.0), no qual 100% das associações previstas foram encontradas como relacionadas a uma associação de referência. Essa configuração pode ser usada para fazer predições muito precisas, para quando os pesquisadores quiserem ter certeza de que as novas funções do lncRNA que estão analisando provavelmente refletirão o papel que essas moléculas desempenham no organismo em estudo.

5.3 Os RNAs Não-Codificantes do *A. gigas*

O A. gigas, também chamado de Pirarucu, é uma espécie de peixe muito peculiar. Tendo divergido de seu parente mais próximo com um genoma sequenciado (*Heterotis niloticus*) 59 milhões de anos atrás (HAO et al., 2020), durante o período Paleoceno, e sendo a única no gênero *Arapaima*, nenhuma espécie viva é muito semelhante a esta. Algumas de suas características incluem gigantismo (DU et al., 2019), respiração aérea obrigatória (VIALLE et al., 2018) e nenhum dimorfismo sexual aparente fora da estação reprodutiva (ALMEIDA, 2006). É uma das espécies economicamente mais importantes da bacia do rio Amazonas, fazendo parte da culinária tradicional amazônica (ALMEIDA, 2006). Com capacidade de crescer até 10kg em um ano, o Pirarucu é considerado a espécie mais promissora para o desenvolvimento da piscicultura intensiva na região amazônica (ALMEIDA, 2006; IMBIRIBA, 2001; CARVALHO; NASCIMENTO, 1992).

Atualmente, os estudos sobre o genoma e o transcriptoma dessa espécie têm deixado de lado os ncRNAs e os papéis que essas moléculas podem desempenhar (MARTINS et al., 2020; DU et al., 2019; VIALLE et al., 2018). Por isso, este trabalho desenvolveu uma anotação de ncRNA baseada em múltiplas estratégias, incluindo predição por modelos de covariância, inteligência artificial, homologia de sequência e integração de dados de referência. A maioria dos ncRNA (59209, 93% do total) previstos eram lncRNA, com apenas 38 deles sendo previstos com base em modelos de covariância. Com exceção desses, todos os lncRNA só puderam ser previsto por meio de nossa estratégia de lncRNA, que inclui o uso de um *software* de predição por IA (RNA Samba) e o refinamento de seus resultados.

Uma busca de homologia das sequências de ncRNA anotadas revelou que, com exceção dos tRNAs e de uma pequena parte dos miRNAs, os RNAs não codificantes de *A.*

gigas são muito diferentes dos RNAs de outras espécies, com apenas 1.11% deles tendo semelhança com um RNA de outros organismos. Os lncRNAs previstos, por exemplo, não tinham nenhuma sequência semelhante em outras espécies. Mesmo com poucas semelhanças encontradas (706), a análise de homologia mostrou que a espécie com os ncRNAs mais semelhantes (um total de 298, 42% deles) é um dos parentes sequenciados mais próximos do *A. gigas*, o *S. formosus*. Apesar de ser mais próximo do Pirarucu, nenhuma homologia foi encontrada com RNAs de *H. niloticus*. Isso pode ser uma consequência do fato de que atualmente apenas 39 ncRNAs dessa espécie estão disponíveis, enquanto há um total de 5008 ncRNAs de *S. formosus* anotados (CONSORTIUM, 2019). Cerca de 15% dos ncRNAs semelhantes eram de espécies de tetrápodes e 3% vieram de diversos táxons muito distantes como *Octopodidae* (polvo), *Arthropoda*, *Galeoidea* (tubarões e raias), *Fabidae* (rosídeos), *Poaceae* (gramíneas) e até *Eubacteria*, sugerindo que (embora seja algo raro) às vezes os ncRNAs podem permanecer preservados mesmo em organismos muito distantes. As demais sequências semelhantes eram genes de Tetra-cego, Peixe-zebra, Salmão, Carpa, Tilápia e outros peixes ósseos.

5.3.1 Fatores de crescimento e dimorfismo sexual do Pirarucu

Sendo o rápido crescimento do *A. gigas* uma das principais características que impulsionam o interesse econômico na espécie, é importante entender os mecanismos biológicos que regulam o crescimento deste peixe. Estudos anteriores procuraram apenas por genes codificantes de crescimento (DU et al., 2019), não investigando o papel dos lncRNA na regulação do gigantismo de *A. gigas*. Aqui, foi encontrado um número muito grande de lncRNAs (8201, 13.8% do total) relacionados ao crescimento, incluindo lncRNAs tecido-específicos para todos os tecidos. Um conjunto de seis lncRNAs relacionados ao crescimento, não descritos em trabalhos anteriores, foi classificado como *housekeeping* (Tabela 8), o que significa que são reguladores ativos na maioria dos tecidos. Também foi observado que este mesmo conjunto de *housekeeping*-lncRNAs de crescimento está relacionado ao envelhecimento, morte celular, reparo de DNA e defesa contra tumores. Isso sugere que, nessa espécie de peixe, as vias moleculares para o controle do crescimento também estão profundamente entrelaçadas com a regulação do envelhecimento e defesa contra a formação de câncer.

Atualmente, a principal forma de diferenciar os indivíduos machos de Pirarucu das fêmeas é uma pigmentação vermelha desenvolvida pelos machos durante a época de reprodução, a qual costuma ser mais intensa que a pigmentação feminina (ALMEIDA, 2006). Marcadores moleculares poderiam facilitar a sexagem dos peixes e tornar mais fácil para os criadores de peixes combinar indivíduos machos com fêmeas. Em nossos resultados, o conjunto de lncRNAs diferencialmente expressos entre machos e fêmeas estava enriquecido funcionalmente para termos diretamente relacionados à pigmentação de pele. Quatro dos lncRNAs DE estavam associados com tais termos e expressos principalmente na

amostra de pele masculina juvenil, com expressão zero na amostra de pele feminina. Essas quatro moléculas tem potencial para serem usadas num teste de sexagem de indivíduos jovens. A análise dos lncRNA anotados com funções de maturação também revelou 11 lncRNAs associados a diferenciação sexual feminina e 2 lncRNAs de maturação (um com expressão mista e um específico para gônadas) que tinham expressão diferencial entre os sexos. Esses lncRNAs relacionados à maturação são candidatos potenciais a marcadores biológicos relacionados à diferenciação sexual desse peixe. É importante notar que, assim como os demais lncRNAs, não foi possível encontrar uma possível homologia entre esses potenciais marcadores e genes em outras espécies, nem mesmo em *Scleropages formosus*. Uma vez que estudos anteriores não exploraram ncRNAs e suas funções, esses possíveis biomarcadores não foram descritos anteriormente e representam novas oportunidades para aumentar a produtividade de *A. gigas* em cativeiro.

6 Conclusão

Neste trabalho, descrevemos o processo de descoberta de ncRNAs em *A. gigas* com o *RNA-Gatherer*, que inclui uma *pipeline* para identificação de ncRNA (*Explorer*) e outra para anotação funcional (*Prophet*). Os detalhes dos requisitos e instruções de uso estão documentados em github.com/pentalpha/rna_gatherer.

O *Explorer* foi capaz de identificar 10 tipos de genes não-codificantes (lncRNA, tRNA, miRNA, rRNA, snRNA, sRNA, scRNA, ribozima, *antisense*, antitoxina e CRISPR), 5 tipos de região cis-reguladora (riboswitch, IRES, elemento frameshift, líder e termorregulador) e um ncRNA intrômico. Independentemente de usar nosso pipeline de várias etapas ou apenas RNA Samba, as previsões de lncRNA incluíram muitos ncRNA de outros tipos, portanto, melhorias devem ser feitas na diferenciação de lncRNA de outros ncRNA. As predições funcionais de lncRNA de maior desempenho feitas pelo *Prophet* resultaram do uso da métrica *Distance Correlation*, destacando a importância das relações não-lineares no estudo da correlação entre genes.

Ambos os pipelines, *Explorer* e *Prophet*, foram aplicados para descobrir os ncRNAs de *A. gigas* e suas funções. A predição de lncRNA identificou sequências que foram filtradas por comprimento longo, ORFs curtos, falta de semelhança com proteínas conhecidas e presença no genoma. No entanto, esses lncRNA não tinham semelhança próxima com sequências de ncRNA conhecidas em outras espécies. Isso significa que eles nunca poderiam ser identificados apenas por comparação de sequência. Até mesmo o *Infernal*, que usa modelos de covariância (NAWROCKI; EDDY, 2013), identificou apenas um número muito pequeno de lncRNA (38). A utilização de um software de previsão baseado em inteligência artificial, o *RNA Samba*, foi fundamental para a identificação desses novos lncRNAs.

Em seguida, análises sobre o padrão de expressão de RNAs longos não-codificantes mostraram que um terço deles são tecido-específicos e apenas alguns têm uma função em todos os tecidos. Ao prever funções para todos os lncRNA, 13,8 % pareciam estar relacionado ao crescimento corporal e proliferação celular. Seis desses lncRNAs de crescimento eram genes *housekeeping*, estando correlacionados a várias proteínas conhecidas por terem função na regulação do crescimento animal e na defesa contra tumores.

A predição funcional também revelou 24 lncRNAs de maturação sexual, todos associados com a função de "diferenciação sexual" (GO:0007548), entre os quais há 11 relacionados especificamente ao desenvolvimento das características sexuais femininas e 2 que são diferencialmente expressos (com maior expressão para amostras femininas), o que sugere que os lncRNAs também podem estar envolvidos na diferenciação sexual dessa espécie. O enriquecimento funcional por conjunto de genes indicou que os lncRNA com expressão diferencial estavam enriquecidos para o componente celular "melanossoma" (GO:0042470) e para o processo biológico de "pigmentação celular" (GO:0033059). Quatro

lncRNAs diferencialmente expressos, que mostraram alta expressão na pele masculina juvenil e nenhum na pele feminina, foram anotados com esses termos. Eles são candidatos para reguladores da pigmentação vermelha muitas vezes visível em machos de *A. gigas*. Como o peixe juvenil ainda não está pronto para a reprodução, testar os peixes quanto à expressão desses transcritos pode ser uma maneira não intrusiva e eficaz de identificar o sexo antes que eles precisem ser combinados com outro peixe para reprodução.

O *RNA-Gatherer* é um software livre que visa tornar os estudos sobre ncRNA mais rápidos e completos, acelerando e facilitando a expansão do conhecimento sobre essas moléculas essenciais. Ele se mostrou capaz de anotar os ncRNAs de uma espécie com baixa cobertura de dados, fornecendo preciosos discernimentos sobre seus papéis no organismo.

Referências

AL-TOBASEI, R.; PANERU, B.; SALEM, M. Genome-wide discovery of long non-coding rnas in rainbow trout. *PLoS One*, v. 11, n. 2, p. 1–15, 2016. Citado 7 vezes nas páginas 17, 18, 19, 21, 24, 28 e 53.

ALBANESE, D. et al. Minerva and minepy: a c engine for the mine suite and its r, python and matlab wrappers. *Bioinformatics*, Oxford University Press, v. 29, n. 3, p. 407–408, 2013. Citado na página 33.

ALMEIDA, Y. S. *Análise de parentesco em filhotes de pirarucu (Arapaima gigas Cuvier, 1817), utilizando marcadores microssatélites*. Dissertação (Mestrado em Biotecnologia) — Universidade Federal do Amazonas, 2006. Citado 3 vezes nas páginas 16, 56 e 57.

ANDERS, S.; HUBER, W. *Differential expression of RNA-Seq data at the gene level – the DESeq package*. 2013. Disponível em: <http://bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf>. Citado na página 37.

ANTHON, C. et al. Structured rnas and synteny regions in the pig genomeg. *BMC Genomics*, v. 15, n. 1, p. 459, 2014. Citado 3 vezes nas páginas 20, 21 e 53.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. *Nature genetics*, Nature Publishing Group, v. 25, n. 1, p. 25–29, 2000. Citado 3 vezes nas páginas 15, 18 e 21.

BAEK, J. et al. lncrnanet: Long non-coding rna identification using deep learning. *Bioinformatics*, Maio, p. 1–9, 2018. Citado 3 vezes nas páginas 15, 17 e 19.

BALDI, P. et al. Hidden markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 91, n. 3, p. 1059–1063, 1994. Citado na página 18.

BENJAMINI, Y.; YEKUTIELI, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, JSTOR, p. 1165–1188, 2001. Citado na página 33.

BETANCUR-R, R. et al. Phylogenetic classification of bony fishes. *BMC evolutionary biology*, Springer, v. 17, n. 1, p. 162, 2017. Citado 2 vezes nas páginas 15 e 16.

BJOERTVEDT. *Arapaima gigas in the Siam Centre, Bangkok*. Wikipedia Commons, 2009. Disponível em: <https://commons.wikimedia.org/wiki/File:Arapaima_gigas_01.JPG>. Acesso em: 01 de janeiro de 2020. Citado na página 16.

BRAY, N. L. et al. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, Nature Publishing Group, v. 34, n. 5, p. 525–527, 2016. Citado na página 36.

BRION, P.; WESTHOF, E. Hierarchy and dynamics of rna folding. *Annual review of biophysics and biomolecular structure*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 26, n. 1, p. 113–137, 1997. Citado na página 17.

BRUNEL, H. et al. Miss: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics*, v. 26, n. 15, p. 1811–1818, 2010. Citado na página 22.

BUCHFINK, B.; XIE, C.; HUSON, D. H. Fast and sensitive protein alignment using diamond. *Nature methods*, Nature Publishing Group, v. 12, n. 1, p. 59, 2015. Citado na página 28.

CAMARGO, A. P. et al. Rnasamba: neural network-based assessment of the protein-coding potential of rna sequences. *NAR Genomics and Bioinformatics*, Oxford University Press, v. 2, n. 1, p. lqz024, 2020. Citado 3 vezes nas páginas 19, 28 e 53.

CARREñO, C. R. *vnmabus/dcor: Version 0.5*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3996697>. Citado na página 33.

CARVALHO, L. d. M.; NASCIMENTO, C. do. Engorda de pirarucus (arapaima gigas) em associação com búfalos e suínos. *Embrapa Amazônia Oriental-Circular Técnica (INFOTECA-E)*, Belém, PA: EMBRAPA-CPATU, 1992., 1992. Citado 2 vezes nas páginas 16 e 56.

CHAN, P. P.; LOWE, T. M. trnascan-se: searching for trna genes in genomic sequences. In: *Gene Prediction*. [S.l.]: Springer, 2019. p. 1–14. Citado 2 vezes nas páginas 18 e 28.

CONESA, A. et al. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, Oxford University Press, v. 21, n. 18, p. 3674–3676, 2005. Citado na página 18.

CONSORTIUM, T. R. Rnacentral: a hub of information for non-coding rna sequences. *Nucleic Acids Research*, v. 47, n. D1, p. D221–D229, 2019. Citado 5 vezes nas páginas 15, 21, 29, 39 e 57.

COORDINATORS, N. R. Database resources of the national center for biotechnology information. *Nucleic acids research*, Oxford University Press, v. 46, n. D1, p. D8–D13, 2018. Citado na página 15.

DE-GROOT, S. J. Ecology of teleost fishes. *Aquaculture*, v. 92, p. 290–291, 1991. ISSN 0044-8486. Disponível em: <http://www.sciencedirect.com/science/article/pii/0044848691900312>. Citado na página 16.

DHANOA, J. K. et al. Long non-coding rna: its evolutionary relics and biological implications in mammals: a review. *Journal of animal science and technology*, Springer, v. 60, n. 1, p. 25, 2018. Citado na página 54.

DU, K. et al. The genome of the arapaima (arapaima gigas) provides insights into gigantism, fast growth and chromosomal sex determination system. *Scientific reports*, Nature Publishing Group, v. 9, n. 1, p. 1–11, 2019. Citado 3 vezes nas páginas 17, 56 e 57.

D'AUBENTON, F. Étude de l'appareil branchiospinal et de l'organe suprabranchial d'heterotis niloticus ehrenberg 1827. *Bull. de l'IFAN*, v. 17, p. 1179–1201, 1955. Citado na página 16.

ECMA. Ecma-404: The json data interchange format. 2013. Disponível em: <http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf>. Citado na página 30.

EDDY, S. R.; DURBIN, R. Rna sequence analysis using covariance models. *Nucleic acids research*, Oxford University Press, v. 22, n. 11, p. 2079–2088, 1994. Citado na página 18.

EHSANI, R.; DRABLøS, F. Measures of co-expression for improved function prediction of long non-coding rnas. *BMC Bioinformatics*, v. 19, p. 533, 2018. Citado 7 vezes nas páginas 15, 22, 23, 30, 33, 54 e 55.

EISEN, M. B. et al. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 95, n. 25, p. 14863–14868, 1998. Citado na página 22.

FAGBENRO, O. et al. Haematological profile, food composition and digestive enzyme assay in the gut of the african bony-tongue fish, heterotis (clupisudis) niloticus (cuvier 1829)(osteoglossidae). *Tropical Zoology*, Taylor & Francis, v. 13, n. 1, p. 1–9, 2000. Citado na página 16.

FANG, S. et al. Noncodev5: a comprehensive annotation database for long non-coding rnas. *Nucleic acids research*, Oxford University Press, v. 46, n. D1, p. D308–D314, 2018. Citado na página 21.

FINN, R. D. et al. The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, Oxford University Press, v. 44, n. D1, p. D279–D285, 2016. Citado na página 15.

FLEISCHMANN, R. D. et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, American Association for the Advancement of Science, v. 269, n. 5223, p. 496–512, 1995. Citado na página 15.

FRANCOIS, C. M. et al. Prevalence and implications of contamination in public genomic resources: a case study of 43 reference arthropod assemblies. *G3: Genes, Genomes, Genetics*, G3: Genes, Genomes, Genetics, v. 10, n. 2, p. 721–730, 2020. Citado na página 21.

FREYHULT, E. K.; BOLLBACK, J. P.; GARDNER, P. P. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding rna. *Genome research*, Cold Spring Harbor Lab, v. 17, n. 1, p. 117–125, 2007. Citado na página 18.

FU, L. et al. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, Oxford University Press, v. 28, n. 23, p. 3150–3152, 2012. Citado na página 36.

GORFINE, M.; HELLER, R.; HELLER, Y. Comment on detecting novel associations in large data sets. Não publicado. 2012. Disponível em: <http://www.math.tau.ac.il/~ruheller/Papers/science6.pdf>. Acesso em: 20 nov. 2020. Citado na página 55.

GUO, X. et al. Long non-coding rnas function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res*, v. 41, n. 2, p. e35, 2013. Citado 2 vezes nas páginas 21 e 54.

GUO, X. et al. Advances in long noncoding rnas: identification, structure prediction and function annotation. *Briefings in functional genomics*, Oxford University Press, v. 15, n. 1, p. 38–46, 2016. Citado na página 54.

GUO, X. et al. Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. *PLOS ONE*, v. 9, n. 2, p. e87446, 2014. Citado 2 vezes nas páginas 22 e 30.

HAAS, B. J. et al. De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols*, Nature Publishing Group, v. 8, n. 8, p. 1494, 2013. Citado 2 vezes nas páginas 28 e 36.

HAGBERG, A.; SWART, P.; CHULT, D. S. *Exploring network structure, dynamics, and function using NetworkX*. [S.l.], 2008. Citado na página 33.

HAO, S. et al. African arowana genome provides insights on ancient teleost evolution. *Iscience*, Elsevier, v. 23, n. 11, p. 101662, 2020. Citado 2 vezes nas páginas 16 e 56.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Citado na página 30.

HARRIS, Z. N.; KOVACS, L. G.; LONDO, J. P. Rna-seq-based genome annotation and identification of long-noncoding rnas in the grapevine cultivar. *BMC Genomics*, v. 18, n. 1, p. 1–12, 2017. Citado na página 18.

HIMMELSTEIN, D. et al. *dhimmel/obonet: obonet v0.2.6 release*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.4053204>. Citado na página 33.

HRBEK, T.; CROSSA, M.; FARIAS, I. Conservation strategies for arapaima gigas (schinz, 1822) and the amazonian várzea ecosystem. *Brazilian Journal of Biology*, SciELO Brasil, v. 67, n. 4, p. 909–917, 2007. Citado na página 16.

HUNG, T. et al. Extensive and coordinated transcription of noncoding rnas within cell-cycle promoters. *Nature Genetics*, v. 43, n. 7, p. 621–9, 2011. Citado na página 17.

IMBIRIBA, E. P. Potencial de criação de pirarucu, arapaima gigas, em cativeiro. *Acta Amazonica*, SciELO Brasil, v. 31, n. 2, p. 299–299, 2001. Citado 2 vezes nas páginas 16 e 56.

JANNESAR, M. et al. A genome-wide identification, characterization and functional analysis of salt-related long non-coding rnas in non-model plant pistacia vera l. using transcriptome high throughput sequencing. *Scientific reports*, Nature Publishing Group, v. 10, n. 1, p. 1–23, 2020. Citado na página 53.

JIANG, Q. et al. Lncrna2function: a comprehensive resource for functional investigation of human lncrnas based on rna-seq data. *BMC Genomics*, v. 16, p. S2, 2015. Citado 7 vezes nas páginas 15, 21, 22, 23, 33, 54 e 55.

JOHNSSON, P. et al. Evolutionary conservation of long non-coding rnas; sequence, structure, function. *Biochimica et Biophysica Acta (BBA)-General Subjects*, Elsevier, v. 1840, n. 3, p. 1063–1071, 2014. Citado na página 19.

KALVARI, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding rna families. *Nucleic acids research*, Oxford University Press, v. 46, n. D1, p. D335–D342, 2018. Citado 4 vezes nas páginas 15, 17, 18 e 28.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford University Press, v. 28, n. 1, p. 27–30, 2000. Citado na página 15.

KANG, Y. J. et al. Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research*, Oxford University Press, v. 45, n. W1, p. W12–W16, 2017. Citado 2 vezes nas páginas 19 e 53.

KERN, C. et al. Genome-wide identification of tissue-specific long non-coding rna in three farm animal species. *BMC Med Genomics*, v. 19, n. 1, p. 684, 2018. Citado na página 18.

KLOPFENSTEIN, D. et al. Goatools: A python library for gene ontology analyses. *Scientific reports*, Nature Publishing Group, v. 8, n. 1, p. 1–17, 2018. Citado na página 37.

LAGARDE, J. et al. High-throughput annotation of full-length long noncoding rnas with capture long-read sequencing. *Nature genetics*, Nature Publishing Group, v. 49, n. 12, p. 1731–1740, 2017. Citado na página 54.

LAVOUÉ, S. Was gondwanan breakup the cause of the intercontinental distribution of osteoglossiformes? a time-calibrated phylogenetic test combining molecular, morphological, and paleontological evidence. *Molecular phylogenetics and evolution*, Elsevier, v. 99, p. 34–43, 2016. Citado na página 16.

LEE, H. K. et al. Coexpression analysis of human genes across many microarray data sets. *Genome research*, Cold Spring Harbor Lab, v. 14, n. 6, p. 1085–1094, 2004. Citado na página 22.

LEGEAI, F.; DERRIEN, T. Identification of long non-coding rnas in insects genomes. *Current Opinion in Insect Science*, Elsevier, v. 7, p. 37–44, 2015. Citado na página 53.

LENG, N. et al. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, v. 29, p. 1035–1043, 2013. Disponível em: <https://doi.org/10.1093/bioinformatics/btt087>. Citado na página 37.

LI, G.-Q. Phylogeny of osteoglossomorpha. *Interrelationships of fishes*, Academic Press, p. 163–174, 1996. Citado na página 16.

LI, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, Oxford University Press, v. 34, n. 18, p. 3094–3100, 2018. Citado na página 28.

LOVE, M. I.; HUBBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, v. 15, p. 550, 2014. Citado na página 37.

MA, K. et al. Genome-wide identification and characterization of long non-coding rna in wheat roots in response to ca2+ channel blocker. *Front Plant Sci*, v. 9, n. Março, p. 1–16, 2018. Citado na página 18.

MARTINS, D. L. et al. Characterization and analysis of the transcriptome in arapaima gigas using multi-tissue rna-sequencing. *bioRxiv*, Cold Spring Harbor Laboratory, 2020. Disponível em: <https://www.biorxiv.org/content/early/2020/10/01/2020.09.29.317222>. Citado 4 vezes nas páginas 17, 36, 37 e 56.

MATTICK, J. S. Rna regulation: a new genetics? *Nature Reviews Genetics*, Nature Publishing Group, v. 5, n. 4, p. 316–323, 2004. Citado na página 17.

MATTICK, J. S.; TAFT, R. J.; FAULKNER, G. J. A global view of genomic information–moving beyond the gene and the master regulator. *Trends in genetics*, Elsevier, v. 26, n. 1, p. 21–28, 2010. Citado na página 17.

MINA, M. *FastSemSim*. 2019. Disponível em: <http://sourceforge.net/projects/fastsemsim>. Acesso em: 21 dez. 2020. Citado 3 vezes nas páginas 23, 30 e 31.

MOSTAJO, N. F. et al. A comprehensive annotation and differential expression analysis of short and long non-coding rnas in 16 bat genomes. *NAR Genomics and Bioinformatics*, Oxford University Press, v. 2, n. 1, p. lqz006, 2020. Citado 3 vezes nas páginas 20, 21 e 53.

NAWROCKI, E. P. Annotating functional rnas in genomes using infernal. In: *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*. [S.l.]: Springer, 2014. p. 163–197. Citado na página 18.

NAWROCKI, E. P.; EDDY, S. R. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, Oxford University Press, v. 29, n. 22, p. 2933–2935, 2013. Citado 3 vezes nas páginas 18, 28 e 59.

NELSON, J. S.; GRANDE, T. C.; WILSON, M. V. *Fishes of the World*. [S.l.]: John Wiley & Sons, 2016. Citado 2 vezes nas páginas 15 e 16.

NHGRI. *Human Genome Project Timeline of Events*. National Institutes of Health (NIH), Bethesda, MD, 2020. Disponível em: <https://www.genome.gov/human-genome-project/Timeline-of-Events>. Acesso em: 09 de janeiro de 2021. Citado na página 15.

NIAZI, F.; VALADKHAN, S. Computational analysis of functional long noncoding rnas reveals lack of peptide-coding capacity and parallels with 3 utrs. *Rna*, Cold Spring Harbor Lab, v. 18, n. 4, p. 825–843, 2012. Citado na página 19.

NON-REDUNDANT Proteins. 2020. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Disponível em: <ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz>. Acesso em: 11 jun. 2020. Citado na página 34.

O'LEARY, N. A. et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, Oxford University Press, v. 44, n. D1, p. D733–D745, 2016. Citado na página 34.

PARKINSON, H. et al. Arrayexpress—a public repository for microarray gene expression data at the ebi. *Nucleic acids research*, Oxford University Press, v. 33, n. suppl_1, p. D553–D555, 2005. Citado na página 35.

PATRO, R. et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, Nature Publishing Group, v. 14, n. 4, p. 417, 2017. Citado na página 35.

PEREIRA, G. do V. et al. Characterization of microbiota in arapaima gigas intestine and isolation of potential probiotic bacteria. *Journal of applied microbiology*, Wiley Online Library, v. 123, n. 5, p. 1298–1311, 2017. Citado na página 17.

PERTEA, M. et al. Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nature protocols*, Nature Publishing Group, v. 11, n. 9, p. 1650, 2016. Citado na página 29.

PESQUITA, C. et al. Metrics for go based protein semantic similarity: a systematic evaluation. In: SPRINGER. *BMC bioinformatics*. [S.l.], 2008. v. 9, n. S5, p. S4. Citado 2 vezes nas páginas 30 e 55.

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins expression analysis of short and long non-coding rnas in 16 bat genomes. *Nucleic Acids Research*, v. 35, n. Database issue, p. D61–5, 2007. Citado na página 15.

QUEK, X. C. et al. lncrnadb v2. 0: expanding the reference database for functional long noncoding rnas. *Nucleic acids research*, Oxford University Press, v. 43, n. D1, p. D168–D173, 2015. Citado na página 21.

QUINN, J. J.; CHANG, H. Y. Unique features of long non-coding rna biogenesis and function. *Nat Rev Genet*, v. 17, n. 1, p. 47–62, 2016. Citado 2 vezes nas páginas 17 e 19.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, v. 77, n. 2, p. 257–286, 1989. Citado na página 18.

RAMÍREZ, C. et al. Cetobacterium is a major component of the microbiome of giant amazonian fish (arapaima gigas) in ecuador. *Animals*, Multidisciplinary Digital Publishing Institute, v. 8, n. 11, p. 189, 2018. Citado na página 17.

REBACK, J. et al. *pandas-dev/pandas: Pandas 1.0.3*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3715232>. Citado na página 30.

RESHEF, D. N. et al. Detecting novel associations in large data sets. *Science*, v. 334, n. 6062, p. 1518–1524, 2011. Citado 2 vezes nas páginas 22 e 30.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Biogeosciences*, v. 26, p. 139–140, 2010. Citado na página 37.

SCOTT, E. Y. et al. Identification of long non-coding rna in the horse transcriptome. *BMC Genomics*, v. 18, n. 1, p. 1–11, 2017. Citado na página 18.

SEABOLD, S. et al. *statsmodels/statsmodels: Version 0.8.0 Release*. Zenodo, 2017. Disponível em: <https://doi.org/10.5281/zenodo.275519>. Citado na página 33.

SIGNAL, B.; GLOSS, B. S.; DINGER, M. E. Computational approaches for functional prediction and characterisation of long noncoding rnas. *Trends in Genetics*, Elsevier, v. 32, n. 10, p. 620–637, 2016. Citado 2 vezes nas páginas 17 e 21.

SOURKOV, V. Igloo: Slicing the features space to represent sequences. *arXiv preprint arXiv:1807.03402*, 2018. Citado na página 19.

STEINEGGER, M.; SALZBERG, S. L. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in genbank. *Genome biology*, BioMed Central, v. 21, n. 1, p. 1–12, 2020. Citado na página 21.

SUN, L. et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic acids research*, Oxford University Press, v. 41, n. 17, p. e166–e166, 2013. Citado na página 19.

TAFT, R. J.; MATTICK, J. S. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding dna sequences. *Genome Biology*, BioMed Central, v. 5, n. 1, p. 1–25, 2003. Citado na página 17.

TENNESSEN, J. B. et al. Hidden markov models reveal temporal patterns and sex differences in killer whale behavior. *Scientific reports*, Nature Publishing Group, v. 9, n. 1, p. 1–12, 2019. Citado na página 18.

VIALLE, R. A. et al. Whole genome sequencing of the pirarucu (arapaima gigas) supports independent emergence of major teleost clades. *Genome biology and evolution*, Oxford University Press, v. 10, n. 9, p. 2366–2379, 2018. Citado 3 vezes nas páginas 16, 17 e 56.

VIRTANEN, P. et al. *scipy/scipy: SciPy 1.6.0*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.4406806>. Citado na página 33.

WANG, G. et al. Characterization and identification of long non-coding rnas based on feature relationship. *Bioinformatics*, Oxford University Press, v. 35, n. 17, p. 2949–2956, 2019. Citado na página 19.

WANG, L. et al. Cpat: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic acids research*, Oxford University Press, v. 41, n. 6, p. e74–e74, 2013. Citado na página 19.

WATANABE, L. et al. De novo transcriptome based on next-generation sequencing reveals candidate genes with sex-specific expression in arapaima gigas (schinz, 1822), an ancient amazonian freshwater fish. *PloS one*, Public Library of Science, v. 13, n. 10, 2018. Citado na página 17.

WUCHER, V. et al. Feelnc: a tool for long non-coding rna annotation and its application to the dog transcriptome. *Nucleic acids research*, Oxford University Press, v. 45, n. 8, p. e57–e57, 2017. Citado na página 19.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, Nature Publishing Group, v. 13, n. 5, p. 329–342, 2012. Citado na página 15.

YANG, C. et al. Lncadeep: an ab initio lncrna identification and functional annotation tool based on deep learning. *Bioinformatics*, v. 34, n. 22, p. 3825–3834, 2018. Citado 3 vezes nas páginas 15, 19 e 54.

YATES, A. D. et al. Ensembl 2020. *Nucleic acids research*, Oxford University Press, v. 48, n. D1, p. D682–D688, 2020. Citado na página 17.

ZHANG, J.; ZOU, S.; DENG, L. Gene ontology-based function prediction of long non-coding rnas using bi-random walk. *BMC Med Genomics*, v. 11, n. Suppl 5, p. 99, 2018. Citado 2 vezes nas páginas 21 e 54.

ZHANG, Z. et al. Katzlgo: Large-scale prediction of lncrna functions by using the katz measure based on multiple networks. *IEEE/ACM Trans Comput Biol Bioinform*, v. 16, n. 2, p. 407-416, 2019. Citado 3 vezes nas páginas 21, 22 e 54.

ZHAO, J.; SONG, X.; WANG, K. lncscore: alignment-free identification of long noncoding rna from assembled novel transcripts. *Scientific reports*, Nature Publishing Group, v. 6, p. 34838, 2016. Citado na página 53.

ZHAO, Z. et al. Co-lncrna: investigating the lncrna combinatorial effects in go annotations and kegg pathways based on human rna-seq data. *Database (Oxford)*, v. 2015, p. bav082, 2015. Citado na página 22.

APÊNDICE A – Estatísticas Sobre Limpeza de Amostras de RNA-Seq

O programa Sickle (github.com/najoshi/sickle) foi utilizado para pré-processar as amostras de RNA-Seq utilizadas nesta pesquisa, limpando extremidades de baixa qualidade ou até sequências inteiras, quando estas apresentam baixa qualidade ao longo de todo o seu comprimento.

As duas tabelas abaixo apresentam estatísticas sobre o estado das amostras de *M. musculus* e *A. gigas* após serem pré-processadas. *Base Sum*: bases totais nos *reads* limpos; GC (%): conteúdo GC; N(%): porcentagem de bases não identificadas (N) nos *reads* limpos; Q25 corresponde à porcentagem de bases com qualidade ≥ 25.

Tabela 9 – Estatísticas das amostras limpas de *Mus musculus*.

Nome da Amostra	***Reads* Totais**	***Base Sum***	**GC (%)**	**N (%)**	**Q25(%)**
BRAIN 1	7.6956E+07	7.6948E+09	47.0	0.009	92.575
BRAIN 2	9.9176E+07	1.5957E+10	49.0	0.004	93.787
BRAIN 3	2.8759E+07	2.3006E+09	48.5	0.05	95.442
COLON 1	9.0908E+07	9.0898E+09	48.5	0.047	93.059
COLON 2	5.6601E+07	9.1091E+09	49.5	0.001	91.919
COLON 3	3.0184E+07	2.4146E+09	49.5	0.039	95.238
HEART 1	3.3278E+07	2.3960E+09	45.5	0.016	97.709
HEART 2	1.4134E+07	1.1307E+09	46.5	0.164	94.78
KIDNEY 1	1.0917E+08	1.0916E+10	47.5	0.004	95.256
KIDNEY 2	1.0848E+08	1.6865E+10	47.0	0.007	94.962
KIDNEY 3	2.9049E+07	2.0913E+09	49.0	0.007	98.633
LIVER 1	1.0685E+08	1.0685E+10	47.5	0.002	94.755
LIVER 2	1.1782E+08	1.8961E+10	50.0	0.001	95.547
LIVER 3	3.0650E+07	2.4519E+09	49.0	0.012	95.64
LUNG 1	3.2247E+07	2.3217E+09	48.5	0.015	97.253
LUNG 2	4.3096E+07	6.9350E+09	49.0	0.109	86.492
LUNG 3	1.0369E+08	1.0368E+10	50.5	0.001	95.716
SKELETAL MUSCLE TISSUE 1	1.0268E+08	1.0267E+10	50.0	0.001	94.036
SKELETAL MUSCLE TISSUE 2	9.4474E+07	1.5203E+10	50.0	0.007	92.079
SKELETAL MUSCLE TISSUE 3	2.8710E+07	2.0671E+09	49.5	0.006	98.412
SPLEEN 1	1.0321E+08	1.0320E+10	48.5	0.001	94.186
SPLEEN 2	9.6591E+07	1.5543E+10	50.0	0.001	93.223
SPLEEN 3	1.0242E+08	1.0241E+10	50.0	0.002	94.478
TESTIS 1	9.8979E+07	9.8964E+09	49.5	0.015	94.501
TESTIS 2	9.7233E+07	1.5648E+10	50.5	0.002	92.79
TESTIS 3	3.1488E+07	2.5189E+09	49.5	0.03	95.703

Fonte: os autores

Tabela 10 – Estatísticas das amostras limpas de *Arapaima gigas*.

Nome da Amostra	***Reads* Totais**	***Base Sum***	**GC (%)**	**N (%)**	**Q25(%)**
BRAIN FEMALE	1.3611E+07	1.9861E+09	51.5	0.001	99.739
BRAIN MALE	1.1900E+07	1.7660E+09	55.0	0.001	99.883
GONAD FEMALE	2.9612E+07	4.3751E+09	52.0	0.001	99.861
GONAD MALE	6.3277E+06	9.0724E+08	58.0	0.001	99.64
HEART MALE	7.0586E+06	1.0143E+09	58.0	0.001	99.642
KIDNEY FEMALE	8.5317E+06	1.2286E+09	54.5	0.001	99.675
KIDNEY MALE	8.8805E+06	1.2680E+09	57.5	0.001	99.611
LIVER FEMALE	1.6562E+06	1.4282E+09	54.0	0.651	99.76
LIVER MALE	1.8951E+06	1.5844E+09	56.0	0.577	99.773
LUNG FEMALE	6.8423E+06	9.8969E+08	55.5	0.001	99.703
LUNG MALE	1.6229E+07	2.3631E+09	55.5	0.001	99.745
MUSCLE FEMALE	1.1560E+07	1.6545E+09	55.0	0.001	99.63
SKIN FEMALE	5.7701E+05	3.4935E+08	46.0	0.756	99.382
SKIN MALE	1.3106E+06	9.4121E+08	48.0	0.62	99.629

Fonte: os autores

APÊNDICE B – Níveis de Confiança e Estatísticas Funcionais de Predição

Este apêndice contém os resultados dos cálculos de Níveis de Confiança e as estatísticas sobre as predições funcionais de teste realizadas. As tabelas e listas deste apêndice estão armazenadas externamente através do repositório público do projeto, no endereço:

<www.github.com/pentalpha/rna_gatherer/dissertation/Confidence Levels and Functional Prediction Stats.xlsx>

O arquivo contém as seguintes sessões:

- **confidence_levels-MF**, **confidence_levels-BP** e **confidence_levels-CC**: Estas são as tabelas de Níveis de Confiança das métricas de correlação. Há uma sessão para cada uma das ontologias GO: função molecular, processo biológico e componente celular, respectivamente;
- **MF-prediction_stats**, **BP-prediction_stats** e **CC-prediction_stats**: Parâmetros e estatísticas para todas as predições de teste feitas das funções de lncRNA de *M. musculus*, separadas pela ontologia dos termos preditos. Contém as seguintes colunas:
  - ***confidence level*** - O parâmetro NC da predição;
  - ***metrics*** - O parâmetro de combinação de métricas;
  - ***number of metrics*** - O número de métricas diferentes;
  - ***size compared to reference*** - O número de associações na predição dividido pelo número de associações na anotação de referência;
  - ***% in reference*** - Valor de Q1;
  - ***% with path to reference*** - Valor de Q2;
  - ***quality*** - Qualidade, calculada baseado em Q1 e Q2;

APÊNDICE C – Análise da Expressão e Funções dos lncRNAs de Pirarucu

Neste apêndice estão contidos os resultados brutos da análise de padrão de expressão dos lncRNAs, assim como da busca por funções de maturação sexual e crescimento. As tabelas e listas deste apêndice estão armazenadas externamente através do repositório público do projeto, no endereço:

<www.github.com/pentalpha/rna_gatherer/dissertation/Analysis of Expression and Functions of Pirarucu lncRNAs.xlsx>

O arquivo contém as seguintes sessões:

- ***lncRNA classification***: O mesmo que a Tabela 8;
- ***analysis per transcript***: Classificação de cada lncRNA. Inclui o tipo de padrão de expressão, média de expressão em macho/fêmea, *fold-change*, expressão diferencial, se está envolvido em crescimento e se está envolvido em maturação;
- ***growth functions***: Termos do *Gene Ontology* usados para determinar se um ncRNA está associado a crescimento;
- ***involved in growth***: Versão reduzida da segunda sessão, só com os lncRNAs relacionados a crescimento;
- ***growth housekeeping lncrna***: Funções de crescimento e proteínas co-expressas com os *housekeeping*-lncRNA associados a crescimento;
- ***maturation functions***: Termos do *Gene Ontology* usados para determinar se um ncRNA está associado a maturação sexual;
- ***involved in maturation***: Versão reduzida da segunda sessão, só com os lncRNAs relacionados a maturação;
- ***sex differential***: Lista de lncRNAs diferencialmente expressos entre macho e fêmea;
- ***involved in male pigmentation***: Lista de lncRNAs diferencialmente expressos e envolvidos com funções de pigmentação;

APÊNDICE D – Homologias com Outras Espécies dos ncRNAs de *A. Gigas*

Neste apêndice estão os resultados brutos da busca por homologia de ncRNAs. As tabelas e listas deste apêndice estão armazenadas externamente através do repositório público do projeto, no endereço:

<www.github.com/pentalpha/rna_gatherer/dissertation/A. Gigas ncRNA homologs in other species.xlsx>

O arquivo contém as seguintes sessões:

- ***Homologs***: Tabela com as homologias encontradas entre ncRNAs de Pirarucu e de outras espécies. Contém o nível de similaridade, a espécie da homologia, o tipo de RNA, a identidade e a cobertura;
- ***By RNA type***: Anotações com similaridade à sequências do banco de dados, agrupadas por tipo da anotação;
- ***By Taxon (resumed)***: Igual a Tabela 7;
- ***By Taxon Full List***: Versão da Tabela 7 onde os táxons menos frequentes também são especificados;