UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA
TESE DE DOUTORADO

Juliana Assis Geraldo

Integração de Dados para Avaliação da Qualidade da Anotação dos Genes Codificadores de Proteínas em Eucariotos

Orientador: Dr. Gabriel Fernandes

Coorientadora: Dra. Jessica Kissinger

Belo Horizonte
Junho de 2019

Juliana Assis Geraldo

Integração de Dados para Avaliação da Qualidade da Anotação dos Genes Codificadores de Proteínas em Eucariotos

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial a obtenção do título de Doutora em Bioinformática.

ÁREA DE CONCENTRAÇÃO: BIOINFORMÁTICA GENÔMICA

Orientador: Dr. Gabriel Fernandes

Coorientadora: Dra. Jessica Kissinger

Belo Horizonte
Junho de 2019

04 Geraldo, Juliana Assis.
Integração de dados para avaliação da qualidade da anotação dos genes codificadores de proteinas em eucariotos [manuscrito] / Juliana Assis Geraldo. – 2018.

116 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Gabriel Fernandes. Coorientadora: Profª. Drª. Jessica Kissinger.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Bioinformática - Teses. 2. Genomas. 3. DNA. 4. Sintemia. 5. Gene. I. Fernandes, Gabriel. II. Kissinger, Jessica. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU:

Ficha catalográfica elaborada pela Biblioteca do Instituto de Ciências Biológicas

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG
Avenida Presidente Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
Endereço eletrônico: bioinfo@icb.ufmg.br 55 31 3409-2554

"Integração de Dados para Avaliação da Qualidade da Anotação dos Genes Codificadores de Proteínas em Eucariotos"

Juliana Assis Geraldo

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Gabriel da Rocha Fernandes - Orientador
FIOCRUZ

Prof. José Miguel Ortega
UFMG

Prof. João Luís Reis Cunha
UFMG

Profa. Glória Regina Franco
UFMG

Prof. Arthur Gruber
USP

Belo Horizonte, 10 de junho de 2019.

Este trabalho foi desenvolvido no Instituto de Pesquisas René Rachou – IRR – FIOCRUZ, sob orientação do Dr. Gabriel Fernandes; Kissinger's group na *University of Georgia* – UGA, Athens – USA, sob orientação da Dra. Jéssica Kissinger (período sanduíche Janeiro 2016 a janeiro de 2017, janeiro a abril de 2019); *Parasite Genomics Group* no *Wellcome Sanger Institute*, Hinxton UK, sob supervisão do Dr. Matthew Berriman (Agosto-Setembro 2016). A bolsa de estudo foi proveniente da agência de fomento Capes Proex e, durante o período sanduíche, pela agência de fomento do CNPq, por meio do programa Ciência sem Fronteiras. Todo o trabalho foi produzido utilizando dados públicos e/ou fornecidos por colaboradores.

Dedico esse trabalho á minha mãe, Maria de Fátima de Assis,
Se eu nunca desisti, o motivo sempre foi você

AGRADECIMENTOS

Agradeço ao meu orientador Dr. Gabriel da Rocha Fernandes, pela oportunidade concedida e por contribuir com o meu aperfeiçoamento como pesquisadora, fornecendo múltiplas oportunidades para interagir com uma ampla rede de pesquisadores nacionais e internacionais. Obrigada por todos ensinamentos, principalmente como deixar de ser tola. Mesmo que eu não tenha aprendido por completo o principal ensinamento, aprendi muito sobre a vida. Agradeço também por compartilhar o seu conhecimento comigo e pela amizade. Sem dúvidas, você é o melhor orientador que eu poderia ter! Muito Obrigada.

Meus sinceros agradecimentos a minha coorientadora, Dra. Jessica Kissinger. Quem me orienta desde o mestrado, responsável por abrir inúmeras portas em minha carreira. Obrigada por praticamente me adotar! Agradeço pela amizade e confiança. Também agradeço a todos os membros do laboratório.

Agradeço ao Dr. Matthew Berriman pela oportunidade de estágio em seu laboratório no Wellcome Sanger Institute e ao Dr. Alan Tracey por todos os ensinamentos sobre curadoria manual de genes.

A peça chave deste trabalho, o futuro Dr. Francislon Silva de Oliveira. Agradeço por todos ensinamentos e ajuda na programação da ferramenta desenvolvida. Agradeço também pela amizade, compartilhamento do sofrimento e das alegrias em terras tupiniquins e estrangeiras!

Maria Júlia, minha aluna de vocação científica por ter ajudado na curadoria manual dos genes, e por ter me permitido o aprendizado da orientação.

Agradeço todos os colaboradores, os quais me disponibilizaram todos os dados para o desenvolvimento desta tese: Emory University: Dr. Stacey A Lapp, Dra. Mary R Galinski; University of Georgia: Dr. Rodrigo Baptista, Dr. Jeremy Debarry; Wellcome Institute Sanger: Dr. Alan Tracey; Pontifícia Universidade Católica do Rio Grande do Sul: Dr. Henrique Figueiró, Dr. Eduardo Eizirik; Fiocruz Minas: Plataforma de Bioinformática.

Meu muito obrigada ao grupo de trabalho da Fiocruz: Dr. Fabiano Pais. Aos estudantes da bioinfo: Joicy, Pâmela, João, Amanda e todos os que passaram pelo laboratório. Como sempre, agradeço imensamente ao Fausto Santos, por todo suporte e eficiência! Agradeço a todos pelos anos de amizade! Sentirei saudade.

Agradeço aos amigos que fiz durante essa longa jornada. Aos amigos em Belo Horizonte, Cambridge e Athens. Especialmente meu muito obrigada a família Baptista: Rodrigo, Camila, Ana Clara, Francisco, Joseph, Creusa Maria (Naná). Agradeço minha família pelo incentivo nos estudos e pelos momentos de privação dos estudos para viver além do lattes.

Agradeço ainda ao ISCB e RSG por terem me proporcionado momentos de aprendizagens além da tese.

A secretaria da Pós-Graduação em Bioinformática! Muitíssimo obrigada por todo o suporte que venho recebendo desde o mestrado. O trabalho de vocês é essencial na vida de cada um que passa pelo programa. Vocês fazem a diferença.

As agências de fomento Capes e CNPq, por me contemplarem com a bolsa de estudos durante esses 48 meses de minha formação.

Ao médico psiquiatra Dr. Gustavo Coutinho de Faria, por ter cuidado da minha saúde mental durante os tempos sombrios da vida.

Deixo o meu profundo agradecimento a todos aqueles que contribuíram de alguma maneira para a realização deste trabalho. Sejam aqueles que participaram do meu processo de formação e aprendizado ao longo da minha vida, bem como aqueles envolvidos de alguma forma no presente estudo.

"Eu posso não ter ido para onde eu pretendia ir, mas eu acho que acabei terminando onde eu pretendia estar".

Douglas Adams

RESUMO

Estudos de sequenciamento completo de genomas estão se tornando comuns, principalmente devido ao baixo custo, rapidez e precisão das tecnologias de sequenciamento atualmente disponíveis. Em consequência, o volume de dados está aumentando rapidamente e genomas completos e incompletos estão agora disponíveis para uma grande variedade de espécies. No entanto, a montagem e anotação desses dados de sequenciamento em genomas anotados de alta qualidade, continua sendo um grande desafio. As anotações dos genomas estão melhorando constantemente, todavia numerosos erros de anotações continuam presentes nos dados depositados em bancos de dados públicos, sejam estes erros na estrutura ou na função do gene. O processo de avaliação da qualidade da anotação, por muitas vezes, ainda é realizado manualmente o que é bastante custoso, principalmente para grandes e complexos genomas. Deste modo, o presente estudo teve como objetivo geral compreender os desafios da anotação estrutural dos genes codificadores de proteínas de genomas completos de organismos eucariotos, bem como, propôs desenvolver um novo método baseado em sintenia de ortólogos e integração de dados para avaliar de maneira automática a qualidade das anotações geradas, reduzindo, assim, o tempo de curadoria manual dos genes codificadores de proteínas. Para alcançar o objetivo, anotações dos genes codificadores de proteínas em genomas completos de diferentes eucariotos foram realizadas para os seguintes organismos: *Panthera onca* (mamífero)*, Plasmodium coatneyi* e *Plasmodium knowlesi* (parasitos apicomplexas com genomas pequenos), *Schistosoma mansoni* (parasito com genoma médio e alta complexidade de estrutura). Os genomas abrangem diferentes características para representar a diversidade entre os processos de anotação. Durante o processo da anotação, foi possível levantar os casos dos erros de anotações passíveis de detecção automática. Diante do exposto, foi desenvolvida uma plataforma para avaliação automática da qualidade dos genes codificadores de proteínas. A plataforma permite realizar a detecção de erros, utilizando a integração de dados multi-ômicos, com informações da sintenia de genes ortólogos de espécies intimamente relacionadas e informações da estrutura da anotação do gene. No total, o programa contém três módulos: 1-Sintenia de Ortólogos, 2- Estrutural e 3-Transcricional. Os genes com possíveis erros detectados recebem uma baixa pontuação, enquanto aos genes confiáveis é atribuída uma pontuação mais alta. Assim, o novo arquivo de saída gerado pode ser carregado diretamente em programas como WebApollo e Artemis, para executar uma curadoria manual naqueles genes com baixa pontuação, reduzindo o tempo de curadoria manual da anotação. Com a ferramenta foi possível reduzir em 58% a necessidade de curadoria manual dos genes codificadores de proteínas dos genomas estudados.

Palavras-chave: Genoma, Anotação, Avaliação-Qualidade, RNA-Seq, Ortólogos, Sintenia.

ABSTRACT

Whole genome sequencing studies are becoming common in view of the low cost of the sequencing technologies currently available. In consequence, the volume of genome projects is rapidly increasing, and complete genomes are now available for a wide variety of species. Due to the amount of new whole genome sequencing several software and strategies has been developed to evaluate the genome assembly quality. Even in the face of a high-quality genome assembled, the challenge of obtaining a good genome annotation remains. One of the biggest claims is to evaluate the quality of the whole genome annotation. The process of evaluating annotation quality, for many times, is still performed manually which is costly, especially for large and complex genomes. The present study aimed to comprehend the challenges of structural annotation of genes encoding proteins from complete genomes of eukaryotic organisms, as well as, proposed to develop a new method based on synteny of orthologs and integration of multi-omics data, to evaluate automatically the quality of the annotations generated, thus reducing the time of manual curation of the genes encoding proteins. To obtain the result, genes encoding proteins in whole genomes of different eukaryotic organisms were required for the following organisms: *Panthera onca* (mammal), *Plasmodium coatneyi* and *Plasmodium knowlesi* (small genome parasites), *Schistosoma mansoni* (medium genome parasite and high complexity of structure). The genomes cover different characteristics to represent the diversity between the annotation processes. During the annotation process, was possible to raise the cases of annotation errors that can be detected automatically. In this context, a platform was developed for automatic evaluation of the quality of the genes encoding proteins. The platform allows to detect the errors using multi-omic data integration, with synteny information from orthologous genes of closely related species and information on the structure of the gene annotation. In total, the program contains three modules: 1- Synteny of Orthologous, 2- Structural and 3- Transcriptional. The genes with possible errors detected receive a low score, while the reliable genes are assigned with a higher score. Thus, the new generated output file can be loaded directly into programs such as WebApollo an Artemis to perform a manual curation on those genes with low scoring, reducing manual annotation curation time. It was possible to reduce by 58% the need for manual curation of the genes encoding proteins of the studied genomes.

Keywords: Genome, Annotation, Evaluation-Quality, RNA-Seq, Orthologous, Synteny.

LISTA DE FIGURAS

LISTA DE QUADROS E TABELAS

LISTA DE SIGLAS E ABREVIATURAS

BLAST Basic local alignment search tool
CD-HIT Cluster Database at High Identity with Tolerance
CDD Conserved Domain Database
CDS coding sequence
COG Cluster of Orthologous Groups
Chr cromossomo
DNA Ácido Desoxirribonucleico
EST Expressed sequence tag
HMM Modelos ocultos de Markov
ISO-Seq Isoform Sequencing
KEGG Kyoto Encyclopedia of Genes and Genomes
KO KEGG Orthology
mRNA RNA mensageiro
Multi-omics multiple omes: genome, proteome, transcriptome, epigenome
NCBI National Center for Biotechnology Information
OrthoMCL Orthology Markov Cluster Algorithm
OrthoFinder Orthology Markov Cluster Algorithm
PacBio Pacific Biosciences
pb pares de bases
RNA Ácido ribonucleico
RNA-Seq RNA Sequencing
UEKO UniRef Enriched KO
UniProtKB UniProt Knowledge base
UTR Untranslated Region

SUMÁRIO

1- INTRODUÇÃO

Estudos de sequenciamento completo de genomas estão se tornando comuns, principalmente devido ao baixo custo das tecnologias de sequenciamento atualmente disponíveis. Em consequência, o volume de dados está aumentando rapidamente e genomas completos e incompletos estão agora disponíveis para uma grande variedade de espécies (Alhakami, Mirebrahim, and Lonardi 2017; Goodwin, McPherson, and McCombie 2016; Phillippy 2017; Shapland et al. 2015; Yandell and Ence 2012).

Apesar do crescimento dos estudos genômicos por sequenciamento, limitações tecnológicas dos sequenciadores ainda estão presentes. Atualmente vivemos em um cenário de duas vertentes: 1-) Sequências Curtas (*short reads*) com alta qualidade, mas não capazes de resolver o problema da montagem para regiões repetitivas do DNA, e; 2-) Sequências Longas (*long reads*) com qualidade questionável e dificuldade no sequenciamento de homopolímeros (Goodwin, McPherson, and McCombie 2016; Sohn and Nam 2018).

A tecnologia de sequenciamento Illumina, atualmente a mais bem-sucedida plataforma de sequenciamento de sequências curtas, funciona em três etapas básicas: amplificação, sequenciamento e análise. O processo começa com o DNA purificado, passando por fragmentação e ligação aos adaptadores, sendo o sequenciamento realizado de maneira massiva/paralela. Os pontos-chave do sequenciamento Illumina, como citado anteriormente, são: geração de leituras curtas e com alta qualidade (Illumina 2017).

Por outro lado, duas plataformas vêm sendo amplamente utilizadas para a geração de sequências longas: Pacifc Biosciences - PacBio (J. et al. 2009) e Oxford Nanopore (Clarke et al. 2009). Ambas funcionam por método de síntese em tempo real. A maior vantagem destas plataformas é a obtenção de longas sequências, facilitando, assim, os estudos de montagem de genomas. A maior desvantagem é a baixa qualidade das sequências geradas, quando comparada com a plataforma de sequenciamento Illumina (Ar 2016).

1.1 Montagem de Genomas Eucariotos

Independentemente do tipo de sequenciador escolhido para o estudo da obtenção da sequência completa do DNA de um organismo, dois tipos de abordagens podem ser realizados para montar os genomas:

1. *De Novo* ou *ab initio*: a montagem do genoma é feita exclusivamente a partir das sequências geradas, sem usar como base um genoma de referência.
2. Mapeamento: é aplicado quando já existe um genoma montado e pode-se usá-lo como referência na tomada de decisão ao encontrar ambiguidades e erros de sequenciamento.

A montagem *de novo* possibilita a identificação de pontos de divergência fundamentais entre espécies/linhagens oferecendo oportunidade para investigar possíveis inovações biológicas (Vaattovaara et al. 2019). Já o mapeamento é utilizado no estudo de populações, em que o ressequenciamento da espécie é realizado.

Assim como as tecnologias de sequenciamento, os programas para montagem dos genomas possuem limitações. Fatores biológicos como complexidade do genoma, bem como erros humanos de desenho experimental são considerados agravantes dos possíveis erros dos programas de montagem dos genomas (Alhakami, Mirebrahim, and Lonardi 2017).

Diversos programas estão atualmente disponíveis para a montagem *de novo,* como pode ser visto na Tabela 1.

Tabela 1: Programas para Montagem de Genomas

Programa	Leituras Curas	Leituras Longas	Híbridas	Referência
ABySS	x			SIMPSON *et al.*, 2009
VELVET	x			ZERBINO; BIRNEY, 2008
SGA	x			SIMPSON, 2014
PERGA	x			ZHU *et al.*, 2014
SPADES	x	x	x	BANKEVICH *et al.*, 2012
ALLPATHS-LG	x	x	x	BUTLER *et al.*, 2008
CANU		x	x	KOREN et al., 2017
SMARTdenovo	x	x		https:// github.com/ruanjue/smartdenovo

Estes diferem muito em termos de desempenho (velocidade, escalabilidade, requisitos de hardware, aceitação de novas tecnologias de sequenciamento) e no resultado (composição da sequência montada). A escolha da ferramenta vai variar de acordo com o organismo de estudo (tamanho e complexidade do genoma), com o tipo de sequenciador (sequências longas ou curtas, ou ambas) com cobertura do sequenciamento (quantidade de dado gerado) e com os parâmetros selecionados.

Outro grande desafio para os bioinformatas responsáveis pela montagem dos genomas é o acompanhamento da literatura dos programas existentes. Muitas ferramentas depois de publicadas não recebem mais suporte pelos autores, ficando, assim, desatualizados por não

conseguirem acompanhar a evolução do tipo de sequência gerada pelas novas tecnologias de sequenciamento (Alhakami, Mirebrahim, and Lonardi 2017).

A avaliação da qualidade da montagem do genoma é considerada um quesito de extrema importância nos estudos genômicos. Inúmeras métricas já foram e vêm sendo propostas por diversos autores para avaliar a qualidade da montagem. Algumas das métricas mais reconhecidas atualmente são as propostas pela competição Assemblathon (Bradnam et al. 2013) e Genome Assembly Gold-Standard Evaluation (GAGE) (Steven L Salzberg et al. 2012), estas destinam-se a avaliar os métodos atuais das montagens de genomas. O projeto Assemblathon está em sua terceira edição (ainda não publicada) e vem acompanhando a evolução das tecnologias de sequenciamento. Algumas das métricas propostas pelo Assemblathon podem ser acessadas no site do projeto, bem como nos artigos já publicados (Bradnam et al. 2013).

1.1.1 Reconciliação de Montagem

A maioria dos genomas eucariotos está inacabada devido, principalmente, às limitações dos sequenciadores e dos algoritmos de montagens dos genomas. Como citado anteriormente, diversas ferramentas para montagem de genomas estão disponíveis, mas nem sempre é óbvio qual ferramenta e combinações de parâmetros utilizar para a obtenção do melhor resultado. Logo, múltiplas montagens com diferentes montadores e parâmetros são realizadas; a melhor montagem obtida destes é a selecionada para continuação do trabalho (Alhakami, Mirebrahim, and Lonardi 2017; Baptista et al. 2018).

A reconciliação de montagem surge como uma abordagem atraente a qual possibilita a fusão de várias montagens com a intenção de produzir um consenso de maior qualidade. Várias ferramentas de reconciliação foram propostas na literatura (ALHAKAMI; MIREBRAHIM; LONARDI, 2017; ZIMIN *et al.*, 2008).

1.1.2 Genomas Rascunhos

Devido principalmente ao baixo custo de sequenciamento completo de DNA, diversos laboratórios têm acesso agora a este tipo de dado. Por muitas vezes, o objetivo do trabalho de alguns laboratórios é estudar determinados genes, consequentemente, o delineamento experimental do estudo é realizado para gerar um genoma rascunho, do inglês *draft genomes*, sendo estes de alta ou baixa qualidade.

Entende-se por *draft genome*, ou genoma rascunho, aquele genoma o qual apresenta sequência incompleta, podendo ter baixa qualidade das informações, e/ou as informações, principalmente de ordenação, erradas (Blake et al. 2016). Consequentemente, estes genomas necessitam de grandes melhorias. Esse tipo de sequência incompleta, bem como os genomas chamados de completos, vêm recebendo incentivos para serem publicados como relatórios ("*reports*" em inglês) e não mais como artigos científicos do tipo "*genome paper*" (Smith 2016). Alguns autores ainda preferem depositar diretamente os dados genômicos no GenBank (Benson, Dennis A Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi and N 2017) (ou outras bases de dados como o ENA: *European Nucleotide Archive*), sem submeter manuscritos. A última estratégia citada ajuda a enriquecer as bases de dados, entretanto não é tão benéfica para quem o faz, visto que o pesquisador necessita publicar para se manter no mundo acadêmico científico (Smith 2016).

Vários autores têm apontado o cuidado a ser tomado ao trabalhar com genomas incompletos (DENTON *et al.*, 2014; NAUMANN, 2003; SCHNOES *et al.*, 2009; SMITH, 2016). DENTON e colaboradores (2014), em especial, estudaram a magnitude do problema dos genomas incompletos, principalmente no que diz respeito à informação gênica, como por exemplo, a quantidade de genes e famílias gênicas presentes no genoma.

Por outro lado, outros autores acreditam que não podemos subestimar o ganho da ciência com a publicação dos *drafts* e novos genomas (Gabaldón and Huynen 2004; Wit and Gilleard 2017). GABALDÓN e colaboradores (2004), mostraram que o crescente número de genomas sequenciados vem adicionando novas dimensões às ferramentas de análise de sequência e predição da função proteica. GILLEARD (2017) também evidencia a importância do número de projetos de conjuntos de genomas especialmente para helmintos e apresenta o melhoramento dos genomas de referência conforme a quantidade de dados depositados aumentam.

1.2 Anotação dos Genomas Eucariotos

Com o genoma sequenciado e montado, a próxima tarefa desafiadora é identificar os elementos do genoma. A predição da estrutura e atribuição da função dos genes é realizada por meio da anotação dos genomas.

Basicamente, a anotação do genoma é dividida em duas etapas: Estrutural e Funcional. A anotação Estrutural é aquela responsável por definir as características genômicas, como por exemplo, os genes, regiões repetitivas, diferentes tipos de RNA, bem como a composição de sequência e a localização destas registradas no genoma (regiões de início e fim). Por outro lado,

a anotação Funcional é responsável por inferir e atribuir função a um gene ou elemento do genoma (Yandell and Ence 2012).

Embora anotar uma montagem de genoma eucarioto está agora ao alcance de não especialistas, esta continua a ser uma tarefa desafiadora. Isso porque um genoma eucarioto apresenta um nível de complexidade elevado, quando comparado aos procariotos (Yandell and Ence 2012).

Diversos programas para anotação dos genomas estão disponíveis (Borodovsky 2015; Brendel, Xing, and Zhu 2004; Curwen et al. 2004; Grabherr et al. 2011; Holt and Yandell 2011; Kent 2002; Korf 2004; Lee et al. 2013; Pertea et al. 2015; Robinson 2012; Rutherford et al. 2000a; Skinner et al. 2009; Slater and Birney 2005; Stanke et al. 2008; Steinbiss et al. 2016; Thibaud-Nissen et al. 2013; Trapnell et al. 2012), alguns dos programas estão representados na Tabela 2. Assim como as ferramentas para montagem dos genomas, os programas para anotação apresentam vantagens e desvantagens, a melhor anotação dependerá de vários fatores relacionados ao organismo de estudo e aos parâmetros selecionados.

Tabela 2: Programas para Anotação de Genomas

	Programa	**Referência**
Programas Ab initio e com Evidências Guiadas	SNAP	KORF, 2004
	Augustus	STANKE *et al.*, 2008
	Genemark	BORODOVSKY, 2015
	GeneSeqer	BRENDEL; XING; ZHU, 2004
Alinhadores e Montadores de ESTs, RNASeqs	BLAT	KENT, 2002
	Exonerate	SLATER; BIRNEY, 2005
	Cufflinks	TRAPNELL *et al.*, 2012
	Trinity	GRABHERR *et al.*, 2011
	Stringtie	PERTEA *et al.*, 2015
Pipelines de Anotação	MAKER 2	HOLT; YANDELL, 2011
	NCBI	THIBAUD-NISSEN *et al.*, 2013
	Ensembl	CURWEN *et al.*, 2004
	Companion	STEINBISS *et al.*, 2016
Genome browsers para Curadoria	Artemis	RUTHERFORD *et al.*, 2000a
	WebApollo	LEE *et al.*, 2013
	JBROWSE	SKINNER *et al.*, 2009
	IGV	ROBINSON, 2012

Uma das ferramentas de maior acurácia é o AUGUSTUS. Este não requer necessariamente entrada experimental adicional, uma vez que pode ser aplicado modo *ab initio* de predição. No entanto, evidências extrínsecas de várias fontes, como o sequenciamento do

transcritoma ou as anotações de genomas intimamente relacionados, podem ser integradas para melhorar a exatidão e a completude da anotação.

Existem, ainda, as chamadas plataformas de anotação, as quais agregam combinações de diferentes programas, permitindo, assim, a obtenção de um resultado mais completo, exemplos deste tipo de plataforma estão representados na Tabela 2. A plataforma de anotação do MAKER2, uma das ferramentas amplamente utilizadas, permite acrescentar evidências de expressão gênica (dados de RNA-Seq e/ou ESTs) quando disponíveis para a espécie de estudo, e também dados de aminoácidos, podendo ser de espécies intimamente relacionadas. A utilização dessas informações otimiza a anotação podendo levar a melhor acurácia dos resultados.

1.2.1 Métricas de Desempenho dos Preditores Gênicos

Os preditores gênicos compartilham as duas principais medidas de acurácia do gene predito: Sensibilidade e Especificidade. Cada uma dessas, são medidas em relação a algum padrão, geralmente a uma anotação referência de alta confiabilidade (Burset and Guigo 1998).

A Sensibilidade (SN) calcula a previsão do preditor, contabilizando os Falsos Negativos (FN) e Verdadeiros Positivos (VP). SN = VP / (VP +FN). Por outro lado, a Especificidade (SP) sobrepõe a referência, contabilizando os Falsos Positivos (FP) e Verdadeiros Positivos (VP). Por conseguinte, SP = VP / (VP + FP).

Com algumas modificações, SN e SP também podem ser usados para comparar duas anotações entre si. Esta é a abordagem adotada pelo Sequence Ontology Project [1] para calcular a distância de edição da anotação (AED), que pode ser usada para medir a congruência entre uma anotação e a evidência de apoio (Yandell and Ence 2012).

AED é calculada da mesma maneira que SN e SP, mas no lugar de um modelo de gene de referência, as coordenadas da união da evidência alinhada são usadas: AED = 1 - AC, em que AC = (SN + SP) / 2. Um AED de 0 indica que a anotação está em perfeita concordância com as evidências fornecidas, enquanto um AED de 1 indica uma falta completa de suporte à evidência para a anotação (Holt and Yandell 2011; Yandell and Ence 2012).

Mesmo diante deste sistema de acurácia dos genes pelos preditores, diversos erros continuam presentes nos atuais métodos de anotação da estrutura dos genes.

1.2.2 Tipos de Erros de Anotação

Nem todos os genomas montados possuem anotação gênica. Dos genomas e genes anotados não existem ainda estudos aprofundados para detecção de erros de anotação. SCHNOES e colaboradores (2014) fizeram uma estimativa baseados em algumas famílias gênicas em diferentes bancos de dados, em que a taxa de erro estimada chegou a 74%. Os tipos de erros analisados foram tanto os estruturais como os funcionais (Denton et al. 2014; Schnoes et al. 2009; Toker, Feng, and Pavlidis 2016a).

Em geral, os erros na predição estrutural podem ser de diversos tipos, como por exemplo: genes incompletos, duplicados, divididos, ausência de exons e/ou introns, genes não preditos, pseudogenes.

A anotação Estrutural é essencial para que a Funcional esteja correta. Em ambos os casos, a anotação do gene em alta qualidade, como a identificação de regiões codificadoras de proteínas, promotores de genes e regiões não traduzidas (UTRs), é crítica para a investigação da função gênica. Os erros da parte Funcional são devidos, basicamente, a falhas na Estrutural (como perda de domínio-exon), como também a propagação de erros de bases de dados atribuindo a função errada à proteína (Vaattovaara et al. 2019).

Erros de sequenciamento podem levar a erros na montagem do genoma, bem como na anotação dos genes. Erros na montagem do genoma também induzem a erros de anotação, consequentemente, levam a erros nas etapas subsequentes (Heydari et al. 2017).

1.2.3 Avaliação da Qualidade da Anotação dos Genomas

Diversos pesquisadores preocupam-se com a qualidade dos dados, como a predição correta da estrutura dos genes e a atribuição correta da função (Crellen et al. 2016; Denton et al. 2014; Goble et al. 2008; Heiko Müller, Naumann 2003; José Domingos Coutinho, Regina Franco, and Pereira Lobo 2015; Steinbiss et al. 2016; Toker, Feng, and Pavlidis 2016b). Desde 2003, alguns anos após o projeto de sequenciamento do genoma humano ter sido concluído em sua primeira etapa, já era possível encontrar relatos da inquietude de alguns autores em relação à qualidade dos genomas anotados publicados (Heiko Müller, Naumann 2003) e alguns anos mais tarde, o desassossego e o alerta dos riscos de trabalhar com genomas incompletos (Denton et al. 2014).

As anotações dos genomas estão melhorando constantemente, entretanto as anotações automáticas exigem curadoria manual e muitas vezes a validação experimental. Este tipo de

curadoria é particularmente importante para genes com introns grandes e ou muito pequenos, bem como para genes localizados em regiões repetitivas, amplas famílias de genes entre outros.

Grandes esforços vêm sendo realizados para melhorar a anotação dos novos e também dos já publicados genomas (M. A. Dragan et al. 2016). Uma das maneiras de melhorar a qualidade dos dados consiste no processo de avaliação da informação gerada pelos programas de anotação.

O processo de avaliação por muitas vezes ainda é realizado manualmente o que é bastante custoso, principalmente para grandes e complexos genomas. O processo manual, por ser bastante custoso, pode levar ao erro de quem o faz, porém essa também pode ser a maior vantagem deste método: ter o controle humano para avaliação (M. Dragan et al. 2016). Incorporar as informações precisas da curadoria manual e de anotações corrigidas em bancos de dados tem melhorado drasticamente o desempenho dos preditores automatizados de genes (Ederveen, Overmars, and van Hijum 2013).

Após a predição da estrutura das sequências dos genes codificadores de proteínas é necessário atribuir a função a estes genes identificados. Diferentes classificações podem ser empregadas, sejam estas através de uma ferramenta de alinhamento como: BLAST (Camacho et al. 2009), DIAMOND (Buchfink, Xie, and Huson 2014), MMseqs2 (Steinegger and Söding 2017), entre outros, ou por busca de padrões em bancos de dados secundários contendo a informação funcional, como por exemplo: HMMER (Potter et al. 2018), RPS-BLAST (Marchler-Bauer et al. 2017). Por vezes, os bancos de dados agrupam genes ortólogos pela conservação da função.

1.2.4 Genômica Comparativa como Aliada na Detecção de Erros de Anotação

Com o crescente número de genomas completos disponíveis, a análise de novos genomas, como por exemplo a anotação, vem sendo cada vez mais baseada em estudos comparativos. A genômica comparativa apresenta uma importância crucial nos estudos, mas não podemos subestimar os pontos negativos, como a propagação de erros dos genomas iniciais.

Como parte do estudo comparativo, podemos citar os estudos evolutivos. A identificação de relações de ortologias de genes codificadores de proteínas é fundamental para todos os aspectos da genômica comparativa. Durante o curso da evolução, a especiação e a duplicação geram pares de genes homólogos, podendo ser classificados em duas categorias: ortólogos e parálogos. Os ortólogos evoluem por descendência vertical de um único gene ancestral, ou seja por especiação e os parálogos resultam por meio da duplicação (Koonin 2005).

Os genes ortólogos tendem a desempenhar funções similares, em contraste com os parálogos, os quais tendem a divergir no papel funcional. O raciocínio subjacente a esta conjectura é que, quando um gene cria uma cópia no mesmo genoma, então este genoma passa a ter dois genes capazes de realizar a mesma função. Assim, uma dessas cópias pode tornar-se livre da pressão seletiva, resultando em uma maior taxa de mutações e facilitando uma possível mudança na funcionalidade.

Várias ferramentas estão atualmente disponíveis para a identificação de Ortólogos e Parálogos, estas se diferenciam nos algoritmos propostos para detecção. Entre estes podemos citar dois grupos principais: um grupo de métodos aborda o problema, inferindo relações par-a-par entre genes em duas espécies e, em seguida, estendendo a ortologia a várias espécies, identificando conjuntos de genes que abrangem essas espécies, em que cada par de genes é um ortólogo. Os métodos populares que adotam essa abordagem incluem o MultiParanoid (Alexeyenko et al. 2006) e o OMA (Altenhoff et al. 2011) entre outros. Os métodos responsáveis por adotarem essas abordagens de pares têm altos níveis de precisão na recuperação de ortólogos, no entanto estes sofrem com baixas taxas de detecção de grupos de ortólogos, devido a complicações decorrentes de duplicações gênicas.

O segundo grupo de métodos tenta identificar os grupos de ortólogos completos. Os grupos de ortólogos contêm tanto ortólogos quanto parálogos, e neste contexto é frequentemente usado como uma unidade de comparação para genômica comparativa. Podemos citar os programas OrthoMCL (Li, Stoeckert, and Roos 2003) e OrthoFinder (Emms and Kelly 2015) nessa categoria. Ambos utilizam o BLAST (Kent 2002) para calcular pontuações de similaridade de sequências entre sequências em múltiplas espécies e então usma o algoritmo de agrupamento MCL (https://dl.acm.org/citation.cfm?id=868986) para identificar clusters altamente conectados (grupos de sequências similares) dentro deste conjunto de dados.

A identificação de ortólogos é relativamente simples em espécies intimamente relacionadas, as quais contêm poucos rearranjos genômicos. Apesar de ser uma técnica simples, atribuir corretamente a ortologia, especialmente quando determinada por meio de métodos computacionais automatizados, requer informações completas dos organismos. Como citado anteriormente, a quantidade de genomas incompletos (*drafts*) sendo publicados está aumentando, logo, estes genomas passam pelo processo de anotação gênica e inferência de ortólogos, mas continuam incompletos. Nesse tipo de genoma existe uma maior probabilidade de estar faltando genes, tornando a identificação de ortólogos incompleta, ou até mesmo a atribuição de falsos ortólogos (Vallender 2010).

Um dos problemas mais comuns da atribuição incorreta de ortólogos surge como resultado da anotação incompleta (ou errônea) do genoma (Vallender 2010). Isoformas gênicas podem ser grandes problemas para os atuais preditores de homologia, um dos programas mais citados: OrthoMCL não é capaz de trabalhar com isoformas de um mesmo gene, esquivando-se assim da possível inferência de pseudo-paralólogos, e, ao mesmo tempo, jogando fora informações preciosas da evolução dos genes e genomas.

Organismos eucariotos, principalmente grandes mamíferos, apresentam vantagens na detecção de ortólogos, devido a alguns motivos, como por exemplo: questões de xenologia causadas pela transferência horizontal de genes não estão presentes, a evolução convergente no nível de sequência simplesmente não teve o tempo necessário para ocorrer. O ganho e a perda de genes são incomuns, com exceção de grandes famílias gênicas, como os receptores olfativos (grandes felinos e primatas não humanos (Figueiró et al. 2017)). A evolução entre espécies próximas está principalmente na diferença de expressão dos genes em tecidos e condições diferentes, sexo dentre outras.

Outra característica nos estudos de ortólogos é que se espera um alto nível de sintenia de ortólogos em espécies intimamente relacionadas. Teoricamente, dois genes ortólogos compartilham genes vizinhos ortólogos, no entanto, há uma pequena chance de ocorrerem coincidências de homologia ao acaso (Jun, Mandoiu, and Nelson 2009).

Sendo assim, espera-se que a utilização da genômica comparativa, principalmente do estudo de sintenia de ortólogos, possa nos fornecer informações valiosas sobre os genes codificadores de proteínas na anotação dos genomas de eucariotos. Essas informações seriam capazes de nos dizer se a ausência/presença de determinado gene, com base no ortólogo, poderia ser um possível erro de anotação.

As informações aqui levantadas reforçam a importância da curadoria manual da anotação automática dos genomas. Diante do exposto, o presente estudo visa compreender os desafios da anotação estrutural dos genes codificadores de proteínas de genomas completos de organismos eucariotos, bem como, propõe desenvolver novo método baseado em sintenia de ortólogos e integração de dados multi-ômicos, para avaliar de maneira automática a qualidade das anotações geradas.

2 OBJETIVO GERAL

Integrar dados multi-ômicos e estudos de genômica comparativa para anotar e avaliar de maneira automatizada a qualidade das anotações dos genes codificadores de proteínas em genomas completos de organismos eucariotos.

2.1 Objetivos Específicos

- Anotar genomas de organismos complexos com características genômicas distintas que configuram desafios para a bioinformática;
- Integrar dados multi-ômicos para anotação dos genes codificadores de proteínas em genomas completos de organismos eucariotos;
- Avaliar a importância de cada um dos dados multi-ômicos para a qualidade da anotação;
- Implementar um procedimento automático utilizando os dados multi-ômicos e o peso atribuído a cada um dos dados para avaliação da qualidade da anotação;
- Propor uma nova ferramenta de avaliação da qualidade da anotação estrutural dos genes codificadores de proteínas em genomas completos de eucariotos, por meio de integração de dados.

O presente estudo será divido em dois capítulos:

- **Capítulo 1**: Apresentará a anotação dos genes codificadores de proteínas em genomas completos de diferentes organismos eucariotos: *Panthera onca, Plasmodium coatneyi*, *Plasmodium knowlesi*, *Schistosoma mansoni*. Diferentes estratégias foram adotadas de acordo com o organismo de estudo e com as características de cada um dos genomas, bem como, os dados gerados para cada um dos organismos.

    - *Panthera onca* (Jaguar), representa o genoma de mamífero. É um genoma grande (2.4Gb), com uma montagem fragmentada (~5000 *scaffolds*).

    - *Plasmodium coatneyi* e *Plasmodium knowlesi*. Parasitos apicomplexas. Apesar de representarem genomas pequenos (~24Mb), são organismos que apresentam regiões repetitivas e grandes famílias gênicas.

    - *Schistosoma mansoni*. Parasito multicelular, também considerado com genoma pequeno (380Mb). É um genoma que apresenta alto nível de dificuldade para anotação, isso, porque é característico desse parasito apresentar micro exons (3pb) e grandes introns, bem como inúmeras isoformas gênicas.

- **Capítulo 2**: Durante o processo de anotação dos genomas acima citados, foi possível levantar os casos dos erros de anotações passíveis de detecção automática. Deste modo, o presente capítulo teve como objetivo central propor um novo método de avaliação da qualidade da anotação estrutural dos genes codificadores de proteínas em genomas completos de organismos eucariotos, por meio da integração de dados.

    - Desenvolvimento da Ferramenta: será apresentado o desenvolvimento da ferramenta de avaliação da qualidade da anotação dos genes codificadores de proteínas, bem como a validação da mesma;

    - Estudos de Casos: em seguida serão exibidos alguns estudos de casos da aplicação da ferramenta desenvolvida.

CAPÍTULO 1
ESTUDO DA ANOTAÇÃO DOS GENES CODIFICADORES DE PROTEÍNAS EM GENOMAS COMPLETOS DE ORGANISMOS EUCARIOTOS

A primeira parte do capítulo 1 apresenta o estudo do genoma do felino *Panthera onca*, Jaguar brasileiro ou também conhecido como onça pintada. O desenvolvimento deste trabalho tem um valor primordial para o desenvolvimento desta tese. Foi por meio deste projeto que conseguimos traçar o propósito do presente estudo. Em 2015, antes de iniciar o meu doutorado, pude participar da fase inicial do delineamento experimental do sequenciamento do genoma do Jaguar brasileiro.

Desde o delineamento para o sequenciamento do genoma, já começamos a traçar os desafios da montagem e anotação. Com uma montagem do genoma considerada aceitável para iniciar o processo de anotação, nos deparamos com um novo desafio: como avaliar a qualidade da anotação a ser gerada? Diante do exposto, e devido a minha experiência prévia em montagem de genoma de grandes mamíferos, foi possível propor um projeto de tese, em que eu poderia desenvolver uma metodologia reprodutível para avaliar a qualidade da anotação dos genomas de interesse.

Esse projeto, bem como os outros que serão apresentados no decorrer da escrita, serviram como estudos de caso, gerando uma base de dados para aprimorar o modelo proposto no segundo capítulo, onde será apresentada a ferramenta desenvolvida para alcançar o propósito da tese.

Para todos os projetos aqui descritos, abordarei apenas a parte da anotação dos genomas, mesmo quando fui responsável pela montagem do genoma e/ou outra parte do estudo. Será apresentada apenas uma breve descrição do genoma do organismo. As informações completas dos dados gerados, montagem dos genomas e propósitos dos trabalhos podem ser acessadas no anexo da tese, onde estão disponíveis todos os trabalhos publicados em artigos científicos.

3 O PROJETO JAGUAR

O Projeto Genoma Jaguar é um esforço multinacional liderado pelo Grupo de Pesquisa em Genética de Carnívoros da Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Brasil. O objetivo do projeto é caracterizar o genoma da onça-pintada (*Panthera onca*), utilizando diferentes abordagens, empregando recursos genômicos para realizar análises comparativas com outros mamíferos, bem como, estudos aprofundados em nível populacional.

O projeto tem como objetivo fornecer informações para melhorar as estratégias de conservação e manejo em favor das populações de onça-pintada ameaçadas de extinção.

3.1 O genoma da *Panthera onca*

O tamanho do genoma nuclear do jaguar é de 2.4Gb, contendo 25 pares de cromossomos além dos cromossomos sexuais X,Y. Foi sequenciado o genoma de um único indivíduo macho do Pantanal brasileiro com a cobertura de ~94x, usando três tipos de bibliotecas: 180 pares de bases-pb paired-end; 3-kb mate pair e 8-kb mate pair. Com a montagem do genoma foi possível atingir as métricas apresentadas na Tabela 3.

Tabela 3: Métricas de Qualidade da Montagem do Genoma

Metric	Contigs	Scaffolds
CEGMA partial	86.29%	93.15%
CEGMA full	49.60%	56.05%
#contigs	158329	7521
#contigs >1kb	156436	7442
Total length	2,284,631,488	2,405,344,986
Total length >1kb	2,282,949,125	2,405,268,288
Largest contig	338,209	8,985,697
GC (%)	41.51	41.51%
N50	28.53 kb	1.52 Mb
L50	23,26	474
%gaps	0%	5%

3.2 Metodologia

A seguir será apresentada a metodologia utilizada para anotar a montagem do genoma da onça, bem como as métricas de avaliação da qualidade da anotação dos genes codificadores de proteínas.

3.2.1 Anotação dos Genes Codificadores de Proteínas

A plataforma de anotação do MAKER2 (Holt and Yandell 2011) foi escolhida para anotar os genes codificadores de proteínas. As predições gênicas *ab initio* e guiadas por evidências foram produzidas pelo programa SNAP versão 2013-02-16 e pelo programa

Augustus versão 3.0.3 (STANKE *et al.*, 2008). Para permitir uma boa qualidade à anotação, três diferentes estratégias foram empregadas na plataforma do MAKER:

Como primeira estratégia, o programa SNAP foi rodado dentro da plataforma, gerando um arquivo do tipo GFF3. Esse arquivo GFF3, bem como as proteínas do Tigre (Tabela 4) foram fornecidas como evidências para aprimorar o modelo. O Tigre não é o felino mais próximo evolutivamente da onça, porém, na época, era o único genoma disponível para treinamento do modelo. Por conseguinte, foi construído um Modelo Oculto de Markov (HMM) personalizado para a predição dos genes do Jaguar. HMM é um modelo com parâmetros desconhecidos, com o desafio de determinar os parâmetros ocultos a partir dos parâmetros observáveis. Os parâmetros extraídos do modelo podem então ser usados, no caso da anotação, para reconhecer os padrões gênicos do jaguar e aplicar o modelo treinado ao genoma.

Para a segunda etapa, foram acrescentados os transcritos montados do jaguar para fornecer maiores evidências para a plataforma de anotação (ver FIGUEIRÓ *et al.*, 2017 para descrição completa da montagem dos transcritos). Os transcritos não foram diretamente utilizados para treinamento no Augustus devido à baixa cobertura do sequenciamento.

Na última etapa, o modelo gerado: Jaguar.hmm foi utilizado para finalizar a anotação dos genes.

Tabela 4: Genoma do Tigre (*Panthera tigres*)

Common name	Scientific Name	Type of Data	Accession Number	Reference
Tiger	*Panthera tigris*	High-coverage genome	ATCQ01	Cho et al. 2013

Anotações dos genes baseadas em evidências no MAKER2 foram produzidas usando configurações padrão, com exceção de dois parâmetros: foi permitida a predição de isoformas e a acurácia aumentada para 0,5 (AED). Para isso o seguinte comando foi utilizado: maker -CTL, gerando três arquivos: maker_bopts.ctl, maker_exe.ctl, maker_opts.ctl. O arquivo maker_opts.ctl foi editado com os dois parâmetros mencionados.

3.2.2 Avaliação da Qualidade da Anotação

Para avaliar a qualidade da anotação gerada diferentes métricas foram empregadas, entre estas podemos citar: características de bancos de dados - UniProt (Wasmuth and Lima 2016), InterProDatabase (Mitchell et al. 2019), KEGG (Minoru Kanehisa et al. 2016), estratégias de

propagação de ortólogos e cobertura (*completeness*) por dados de RNA-Seq também foram integrados.

3.2.2.1 Anotação Funcional com InterProScan5

A ferramenta do InterProScan5 (Jones et al. 2014) foi utilizada, pois a mesma permite combinar diferentes métodos de reconhecimento de assinatura de proteína em um único recurso. Sendo esta capaz de fornecer uma visão geral das famílias às quais uma proteína pertence, identificando a presença e a organização de domínios de sequência proteica, bem como resíduos críticos.

A sequência do proteoma predito (aminoácidos) no formato FASTA foi submetida ao programa localmente com acesso ao banco já processado, indicado pelo comando `-iprlookup`. As correspondências foram então calculadas em relação a todas as assinaturas de banco de dados de membros necessárias (BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, HAMAP, PatternScan, SuperFamily, SignalPHMM, TMHMM, HMMPanther, Gene3D, Phobius, and Coils) os resultados foram configurados para saída no formato TSV (um arquivo simples delimitado por tabulações). Linha de comando utilizada:

```
interproscan-5.7-48.0/interproscan.sh -i Jaguar.all.maker.proteins.fasta -goterms
-iprlookup -pa -f xml -f tsv -b output
```

3.2.2.2 Avaliação da Anotação com CDD

Nesta etapa, trabalhamos com os genes que não mostraram uma assinatura proteica por meio da ferramenta do InterProScan5. Para isso, foi escolhido o CDD: banco de dados de domínio conservado do NCBI (Marchler-Bauer et al. 2017), o qual fornece uma ferramenta on-line e local para anotar domínios proteicos. Para acessar esse banco localmente é necessário utilizar o RPS-BLAST, com o seguinte comando, onde cog refere-se ao banco de dados:

```
rpsblast -query Jaguar.all.maker.proteins.fasta -db Cog -out rps-blast.out -
evalue 1e-2 -outfmt 6
```

3.2.2.3 Anotação Funcional e Avaliação da Anotação com Uniprot e KEGG

Os genes previstos foram caracterizados por meio de uma pesquisa BLAST contra a Base de dados do UniProt (e-value 1e-10), localmente:

```
'blastall -p blastp -d uniprot.fasta -i Jaguar.all.maker.proteins.fasta -m8 -b1 -
a 40 -e 1e-10 -o JaguarUniprot'
```

Os genes foram designados para grupos KEGG Orthology usando um banco de dados UniRef Enriched (UEKO) (M Kanehisa and Goto 2000) http://maxixe.icb.ufmg.br/ueko/.

Os genes codificadores de proteínas foram anotados por BLAST recíproco contra banco de dados UniProt.

3.2.2.4 Predição de Ortólogos com OrthoMCL

O OrthoMCL é uma ferramenta capaz de fornecer um método escalonável para a construção de grupos ortólogos por meio de vários táxons. Para a busca de ortólogos do jaguar, apenas o proteoma predito do tigre foi utilizado. O programa foi rodado sem alterar os parâmetros.

3.2.2.5 Avaliação da Anotação com Dados de RNA-seq

Os conjuntos de dados de RNA-seq foram gerados neste estudo, a partir do mesmo indivíduo cujo genoma foi sequenciado. Este conjunto de dados foi utilizado para avaliar os genes anotados. Para tanto, a sequência de cada gene foi tratada como referência e as leituras de RNA-seq foram mapeadas contra o genoma. A ferramenta do Bedtools (Quinlan and Hall 2010) foi utilizada. Neste caso, o conjunto de dados foi o resultado do mapeamento dos transcritos contra a referência no formato de arquivo BAM, característico de mapeamento (detalhes do formato: http://samtools.github.io/hts- specs/SAMv1.pdf).

```
bedtools coverage -s -a file.gff3 -b file.bam > newfile.gff3
```

3.2.2.6 Construção do Banco de Dados

Para aprofundar a análise deste grande conjunto de dados, foi construído o jaguar_SQL, um banco de dados relacional usando o MySQL “Structured Query Language” como sistema

de gerenciamento de banco de dados.

Para recuperar potenciais genes no banco de dados acima mencionado, os identificadores dos genes foram utilizados (Tabela 5) em consultas SQL. A coluna 1 da tabela corresponde aos identificadores dos 25.451 genes codificadores de proteínas. As demais colunas são: id_jaguar_KO, id_jaguar_InterPro, id_jaguar_CDD, id_jaguar_OrthoMCL, id_jaguar_UniProt, respectivamente. Para estas colunas foram acrescentados os valores 0 ou 1, sendo o valor 0 correspondente a ausência de evidência para determinado identificador do gene e 1 presença de evidência para o gene. Evidência neste caso é definido, como por exemplo: determinado gene teve um domínio assinalado utilizando InterProScan5, portanto, o gene recebe o valor de 1, o gene não assinalado recebe o valor 0.

As duas outras colunas presentes correspondem à porcentagem de cobertura em extensão (*completeness*) por RNA-seq e a outra ao tamanho do gene (apenas a parte codificadora: exons). A nona coluna foi acrescentada com a soma da presença das evidências.

Para avaliação da anotação dos genes codificadores de proteínas, os seguintes critérios foram seguidos: todos os genes com soma maior ou igual a 1 foram considerados com alguma evidência e todos os genes com cobertura por RNA-seq maior que 75% (com soma nula ou não).

Tabela 5: Jaguar_SQL

id_jaguar_completo	id_jaguar_KO	id_jaguar_InterPro	id_jaguar_CDD	id_jaguar_OrthoMCL	id_jaguar_uniprot	RNAseq	id_jaguar_length	Soma
Jaguar_00000019-RA	0	1	0	1	1	0.0455696	1185	3
Jaguar_00000021-RA	0	0	0	0	0	0	9	0
Jaguar_00000024-RA	0	1	0	0	1	0.1201814	1323	2
Jaguar_00000025-RA	0	0	0	0	0	0	438	0
Jaguar_00000026-RA	0	1	0	1	1	0.1818182	1452	3
Jaguar_00000037-RA	0	1	0	1	1	0.1359223	309	3
Jaguar_00000039-RA	0	1	0	0	0	0.5111111	315	1
Jaguar_00000042-RA	0	1	0	1	1	0.5241228	1368	3
Jaguar_00000059-RA	0	1	0	1	1	0.6852518	1668	3
Jaguar_00000083-RA	0	0	0	0	0	0.5644444	225	0
Jaguar_00000092-RA	0	0	0	0	0	0.3166667	60	0
Jaguar_00000097-RA	0	1	0	0	1	0.6345382	249	2
Jaguar_00000108-RA	0	0	0	0	0	0.6810898	624	0
Jaguar_00000115-RA	0	0	0	1	1	0.24487	2193	2
Jaguar_00000116-RA	0	1	0	1	1	0.3893387	1482	3
Jaguar_00000117-RA	0	1	0	1	1	0.1596587	1641	3
Jaguar_00000118-RA	0	1	0	1	1	0	783	3

3.3 Resultados

No total, foram identificados 25.451 genes codificadores de proteínas, cujo as características estão descritas na Tabela 6.

Tabela 6: Genes Codificadores de Proteínas

Feature	Number
Base pairs	32,384,308
N50	1,827
Longest gene	26,310
Median gene size	927
% genes > 100 b	99.56%
% genes > 200 b	97.27%
% genes > 500 b	73.65%
% genes > 1 kb	46.56%

A abordagem do InterProScan5 foi capaz de atribuir assinaturas a 22.191 genes codificadores de proteínas da onça-pintada, já o banco de dados do CDD foi capaz de anotar 197 desses genes codificadores de proteínas, não atribuídos com a ferramenta do InterProScan5. Ao todo, 21.279 genes do jaguar (83,6%) mostraram correspondências únicas para UniProt (best-hit) e 4.020 genes apresentaram resultados com KO atribuídos. A Figura 1 ilustra os quatro passos adotados para inferir alguma evidência da veracidade do gene predito.

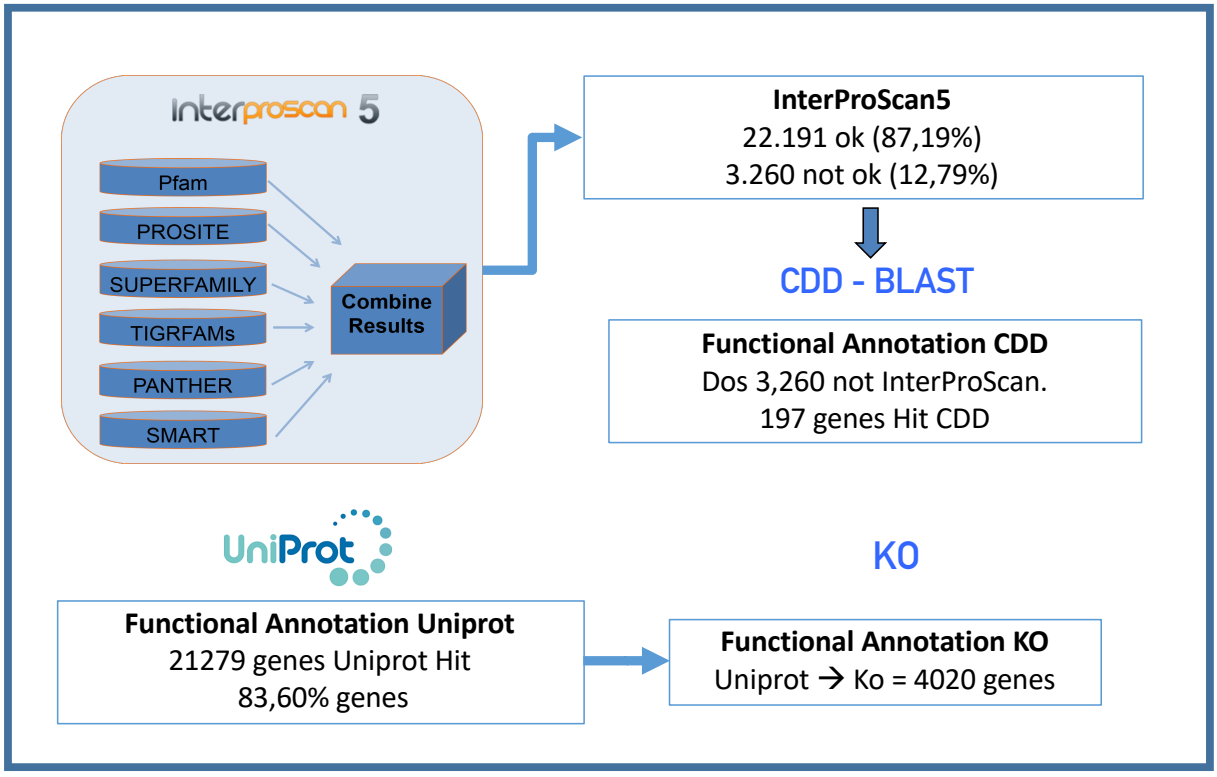

Figura 1: Esquema Representando os Passos Adotados para Fornecer Evidências dos Genes Preditos. InterproScan5 foi rodado localmente, sendo capaz de assinalar pelo menos um domínio proteico a 87% dos genes preditos, os genes que não receberam assinaturas de domínios foram submetidos à estratégia CDD. Blast foi realizado contra o banco de dados complete do UniProt e em seguida a abordagem para recuperar KO do KEGG.

A Figura 2, apresenta em um esquema os 21.575 genes atribuídos a algum grupo de ortólogo, quando comparado ao proteoma predito do Tigre.

Figura 2: Genes Compartilhados entre o Jaguar e o Tigre. Resultados obtidos por meio da ferramenta OrthoMCL.

Dos aproximadamente 25 mil genes, 16.586 genes (65%) alcançaram cobertura de RNA-seq superior a 75% da sequência predita.

No total 24.411 apresentaram algum tipo de evidência para fornecer suporte de confiabilidade ao gene. A Figura 3 representa o banco de dados completo, onde foram plotados os genes Ids, a Soma das evidências, a cobertura por RNA-Seq e o tamanho dos genes.

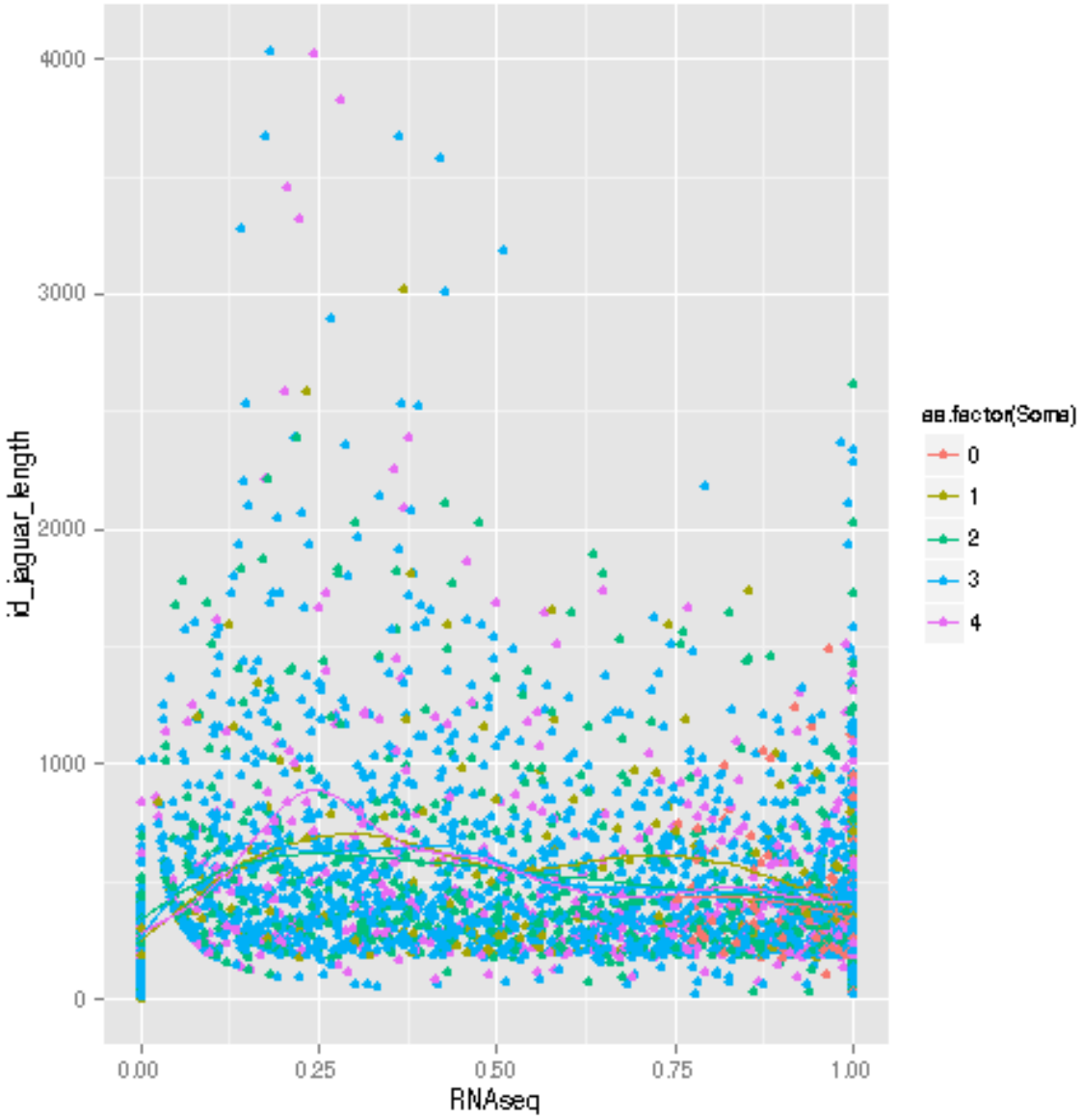

Figura 3: Jaguar_SQL. A figura representa o banco de dados gerado, em que foram plotados: identificador do gene, a cobertura por RNA-seq e os valores de score=0-4. O tamanho não é um fator limitante do gene, genes com tamanhos grandes e pequenos apresentaram algum tipo de evidência.

3.4 Considerações

O genoma do Jaguar, apresentou grande dificuldade no processo de anotação dos genes codificadores de proteínas. Devido as estratégias de anotações empregadas no presente estudo, foi possível alcançar uma anotação considerada de boa qualidade.

Recurso de Bioinformática: Plataforma de Bioinformática Fiocruz Minas

4 PARASITOS APICOMPLEXAS

A segunda parte do capítulo 1, abordará as anotações de dois parasitos apicomplexas, um deles, nunca estudado em sua sequência completa do genoma: *Plasmodium coatneyi*, e o segundo: *Plasmodium knowlesi*, em que foi proposto uma melhoria na montagem e anotação da sequência do genoma.

Os dados de sequenciamento de ambos os parasitos foram gerados pelo projeto MAPHIC (Malaria Host-Pathogen Interact Center), do qual pude fazer parte durante o meu período sanduíche na Universidade da Geórgia (UGA). As características dos dois parasitos, bem como os dados disponíveis e montagens serão abordados juntos, já as estratégias de anotação dos genomas serão abordadas de maneira separada, devido à peculiaridade de cada estratégia. Os trabalhos foram realizados em tempos diferentes, sendo que os dados gerados para o *P. coatneyi* foram desenvolvidos primeiro (primeiro semestre de 2016), e só tardiamente o trabalho para *P. knowlesi* começou a ser desenvolvido (segundo semestre de 2016). As publicações de ambos os estudos estão disponíveis no anexo deste trabalho.

4.1 *Plasmodium knowlesi* e *Plasmodium coatneyi*

Plasmodium knowlesi e *Plasmodium coatneyi* são parasitas intracelulares tendo como hospedeiro natural *Macaca fascicularis* e *Macaca mulatta* e, como hospedeiro intermediário, mosquitos do gênero *Anopheles spp*. . *Plasmodium knowlesi* também é responsável por causar malária em humanos. Mesmo não sendo capaz de infectar humanos, *P. coatneyi* vem se mostrando de grande importância nos estudos de malária humana, servindo como um excelente modelo para estudo dessa doença parasitária.

O tamanho do genoma nuclear de ambos os parasitas é de aproximadamente 24Mb, contendo 14 cromossomos, aproximadamente 5 mil genes codificadores de proteínas. Uma das principais características desses genomas é o conteúdo repetitivo, o que dificulta a montagem do genoma. Ambos apresentam famílias gênicas de antígeno variantes (do inglês: *variant antigen gene families*, SICAvar e kir genes), codificando os antígenos variantes conhecidos como as proteínas de Aglutinina de Células Infeccionadas por esquizontes (da sigla em inglês SICA) que está relacionada à virulência do parasito. Esses genes apresentam exons alternativos (geralmente os últimos), em relação aos quais os programas de anotação de genomas possuem grande dificuldade em anotar (Chien et al. 2016; S. a Lapp, Korir, and Galinski 2009; Stacey A. Lapp et al. 2013; Pain et al. 2008). Outra importante característica do genoma do parasita

causador da malária é o alto conteúdo de A+T. Em *Plasmodium falciparum*, por exemplo, o teor de A+T é de 82%, maior do que qualquer outro organismo cujo DNA tenha sido caracterizado (Weber 1987).

4.1.1 Sequenciamento e Montagem dos Genomas

O genoma nuclear de *Plasmodium coatneyi* foi montado utilizando apenas dados da Pacific Biosciences RS II (PacBio), cobertura superior a 50X (ver CHIEN *et al.*, 2016 em anexo para detalhamento dos dados gerados). A montagem apresenta aproximadamente 500 gaps distribuídos ao longo das sequências, sendo considerada de alta qualidade por ter conseguido montar os 14 cromossomos. O tamanho do genoma é de 26.4 Mb.

O genoma do parasito *Plasmodium knowlesi* foi montado por meio de dados de PacBio III e da técnica de *chromatin conformation* (Hi-C), cobertura superior a 279X. Apenas 25 gaps estão presentes na montagem. A montagem contém os 14 *scaffolds* representando os 14 cromossomos e 15 pequenos *unitigs*. O tamanho do genoma é de 24.6Mb (LAPP, S. A., Geraldo, J. A *et al.*, 2018) .

4.2 Metodologia

As anotações geradas nesse estudo serão chamadas de PKNOH quando referente ao *Plasmodium knowlesi*, e PCOAH quando referente ao *Plasmodium coatneyi*.

4.2.1 Estratégia da Anotação do Genoma do Parasito *Plasmodium coatneyi*

O primeiro passo do trabalho foi estabelecer uma estratégia para a anotação do genoma do parasito. Para alcançar esse resultado, a plataforma do MAKER2 versão 2.31.8 (Holt and Yandell 2011) foi escolhida. Os programas Augustus versão 3.0.3 (Stanke et al. 2008) e SNAP versão 2006-07-28 (Korf 2004) presentes na plataforma, foram selecionados para predições *Ab initio* e guiadas por evidência dos genes.

A plataforma de anotação do MAKER2 permite acrescentar evidências de expressão gênica (dados de RNA-Seq e/ou ESTs) quando disponíveis para a espécie de estudo e dados de proteínas, podendo ser de espécies relacionadas. A utilização dessas informações otimiza a anotação podendo levar a melhor acurácia dos resultados.

Foram selecionados no banco de dados PlasmoDB (Bahl et al. 2003) (https://plasmodb.org) a anotação do organismo de interesse. Como este foi o primeiro estudo

a decifrar a sequência completa do genoma do parasito *P. coatneyi*, o organismo selecionado para treinamento do modelo de anotação foi o organismos mais próximo evolutivamente disponível na época do estudo: *Plasmodium knowlesi* (Pain et al. 2008).

Do conjunto de genes codificadores de proteínas adquiridos, apenas os seguintes genes foram utilizados: genes codificadores de proteínas considerados completos (códon de iniciação e códon de parada) por meio do *script* desenvolvido: coding_report.pl, genes não redundantes, sendo a redundância removida por meio do programa CD-HIT (Fu et al. 2012) `cd-hit -i db -o db90 -c 0.9`, em que o db corresponde ao banco de dados dos genes codificadores de proteínas, db90 ao arquivo de saída e 0.9 corresponde a 90% de redundância, ver http://www.bioinformatics.org/cd-hit/cd-hit-user-guide.pdf , para referências completas.

As sequências proteicas (aminoácidos) também foram obtidas no PlasmoDB e passaram pelo mesmo filtro dos transcritos. Estavam presentes na base de dados do UniProt (Wasmuth and Lima 2016), 432 genes pertencentes ao *Plasmodium coatneyi* os quais foram acrescentados para predição dos genes codificadores de proteínas. A Tabela 7 mostra os dados utilizados para anotação guiada por evidência. PKNH (versão 18/06/2015) foi adquirido no PlasmoDB, entretanto, o Wellcome Sanger Institute foi a fonte original que sequenciou, montou e anotou o genoma (Pain et al. 2008) o genoma está também disponível no GeneDB (Logan-klumpler et al. 2012), ; b) PKNA1-H.1 adquirido no GenBank CWHQ00000000.2 (Moon et al. 2016) e c) *Plasmodium coatneyi* (UniProt 15/02/2016).

Tabela 7: Genes e Genomas Utilizados para Anotação e Comparação

Nome	Organismo	BD	Genes	Proteinas
PKNH	*P. knowlesi*	PlasmoDB	5483	5390
PKNA1.H.1	*P. knowlesi*	GenBank	5430	5172
P. coatneyi	*P. coatneyi*	UniProt	432	432

As informações previamente fornecidas (genes/proteínas) foram adicionadas na predição dos genes, fazendo um tipo de predição guiada por evidências. Essa anotação guiada por evidências, juntamente com a predição *Ab initio* do programa SNAP, incorporado à plataforma do MAKER2 geraram um arquivo em formato GFF o qual foi utilizado para treinar o preditor gênico Augustus. O mesmo filtro para remoção de redundância e de genes incompletos foi aplicado. Apenas os genes considerados completos e únicos foram fornecidos para treinamento do modelo.

O preditor Augustus, assim como o SNAP, é capaz de fazer a predição completamente *Ab initio*, ou seja, utilizando apenas a sequência do genoma, porém a estratégia de predição guiada foi a escolhida por acrescentar mais informações à anotação final. Em seguida, a sequência do genoma foi novamente fornecida como arquivo de entrada no MAKER2, mas da segunda vez acrescentando os modelos gênicos treinados pelos preditores.

As configurações utilizadas no MAKER2 foram produzidas utilizando o comando maker -CTL, gerando três arquivos: maker_bopts.ctl, maker_exe.ctl, maker_opts.ctl. O arquivo maker_opts.ctl foi alterado para melhorar a predição dos genes. A Figura 4 exibe os principais parâmetros alterados: EST: foram fornecidas as sequências de nucleotídeos de *P. knowelesi*, Protein Homology: proteínas de *P. knowlesi* mais as proteínas que estavam disponíveis no UniProt para *P. coatneyi*. Repeat Masking: foi utilizado um arquivo fornecido pelo grupo da Dra. Kissinger, em que as regiões repetitivas de 8 plasmódios já haviam sido caracterizadas. Gene Prediction: os modelos do SNAP e Augustus foram fornecidos na segunda etapa. AED_treshold foi alterado de 1 para 0.5 com o propósito de aumentar a acurácia da predição.

No final, a plataforma do MAKER2 forneceu um consenso de todos os métodos de predição abordados: guiado por evidência e *Ab initio*.

```
#-----EST Evidence (for best results provide a file for at least one)
altest=PlasmoDB-29_PknowlesiH_AnnotatedESTs.fasta #EST/cDNA sequence file in fasta format from an alternate organism

#-----Protein Homology Evidence (for best results provide a file for at least one)
protein=ProteinsPlasmoCoatneyknowlesi  #protein sequence file in fasta format (i.e. from mutiple oransisms)

#-----Repeat Masking (leave values blank to skip repeat masking)
repeat_protein=proteins_masked.fasta #provide a fasta file of transposable element proteins for RepeatRunner

#-----Gene Prediction
snaphmm=PKNOH #SNAP HMM file
augustus_species=PKNOH #Augustus gene prediction species model
est2genome=1 #infer gene predictions directly from ESTs, 1 = yes, 0 = no
protein2genome=1 #infer predictions from protein homology, 1 = yes, 0 = no

#-----External Application Behavior Options
cpus=26 #max number of cpus to use in BLAST and RepeatMasker (not for MPI, leave 1 when using MPI)

#-----MAKER Behavior Options
min_contig=200 #skip genome contigs below this length (under 10kb are often useless)

AED_threshold=0.5 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)
min_protein=20 #require at least this many amino acids in predicted proteins
alt_splice=1 #Take extra steps to try and find alternative splicing, 1 = yes, 0 = no
always_complete=1 #extra steps to force start and stop codons, 1 = yes, 0 = no
tries=3 #number of times to try a contigs if there is a failure for some reason
```

Figura 4: Principais Parâmetros Modificados no Arquivo maker_opts.ctl do MAKER2: Parâmetros referentes à anotação de PCOAH.

4.2.2 Estratégia da Anotação do Genoma do Parasito *Plasmodium knowlesi*

Para o parasito *P. knowlesi* duas diferentes abordagens de anotação foram empregadas. Primeiramente, as predições de genes foram realizadas pela plataforma do MAKER2, como

descrito anteriormente. Em seguida, a sequência do genoma montada foi submetida ao servidor Web Companion[1] (Steinbiss et al. 2016) para anotação e análise de genomas de parasitas utilizando uma abordagem baseada em referência. Em ambos os casos, a sequência referência utilizada foi pertencente à mesma cepa: Pk1 (A+) *Malayan strain clone* (dados apresentados na Tabela 7). A anotação do genoma no Companion foi produzida usando as configurações padrão, com duas exceções descritas a seguir: não modificar a sequência de entrada e a acurácia do preditor Augustus foi aumentada, tendo sido definido para 0,5 o valor de corte.

Para avaliar a qualidade dos genes codificadores de proteínas preditos pelas abordagens do MAKER2 e Companion, uma comparação entre as duas anotações foi realizada por meio da ferramenta ParsEval (Standage and Brendel 2012). Os genes únicos identificados por cada abordagem foram selecionados para curadoria manual utilizando o programa Artemis versão 16.0 (Rutherford et al. 2000b). Para atribuir uma anotação funcional às proteínas previstas, as informações geradas pelo Companion foram mantidas e para as proteínas anotadas que não foram previstas pelo Companion, foram feitas buscas de ortólogos usando OrthoFinder versão 1.1.5 (Emms and Kelly 2015). As informações funcionais obtidas foram transferidas para a anotação final. InterProScan5 (Jones et al. 2014) também foi utilizado para ajudar a atribuir função às proteínas.

4.2.3 Comparação das Anotações existentes de *P. knowlesi* com a Nova Anotação Gerada

Essa etapa só foi realizada para PKNOH devido à inexistência de genoma completamente anotado para *P. coatneyi*.

Os genes codificadores de proteínas preditos e corrigidos manualmente (a correção manual fora comparada com as anotações existentes de PKNH e PKNA1-H.1 (ver Tabela 7). As comparações foram realizadas por meio do programa OrthoFinder com configurações de parâmetros padrão. Uma comparação dos comprimentos das proteínas também foi realizada usando scripts PERL personalizados, disponíveis no material suplementar online no presente estudo: https://github.com/Juassis/Material_Suplementar_Tese.

[1] https://companion.sanger.ac.uk

4.3 Resultados

Os genomas dos parasitos *Plasmodium coatneyi* e *Plasmodium knowlesi* foram anotados e avaliados. No total 5,516 genes codificadores de proteínas foram preditos, 45 tRNAs, 11 rRNAs para o *Plasmodium coatneyi* – PCOAH (Tabela 8).

Para o parasito *Plasmodium knowlesi* (PKNOH), duas estratégias de anotação foram empregadas: MAKER2 e Companion e os resultados foram comparados. A correspondência entre as regiões codificadoras (CDS), da posição inicial à final, entre as estratégias foi de 73,5%. 201 *locus* preditos utilizando as duas estratégias apresentaram divergências totais, sendo 76 *locus* preditos unicamente pela plataforma do MAKER2 e 25 unicamente preditos pelo Companion. Desta forma, o total de 1620 genes foram selecionados para curadoria manual.

Tabela 8: Métricas da Anotação dos Genomas

Nome	Genes	Proteinas	tRNAs	rRNA	Pseudogenes
PKNH	5483	5290	45	11	NA
PKNA1-H.1	5430	5138	45	14	NA
PKNOH-Maker2	5356	5300	45	11	NA
PKNOH- Companion	5315	5253	45	11	152
PKNOH-Manually curated	5321	5258	45	12	18
PCOAH	5572	5516	45	11	NA

NA= Not Available

4.3.1 Correção manual em PKNOH Utilizando o Resultado da Anotação Combinada: SICAvar Família Multigênica e Pseudogenes

A família gênica SICAvar além de ser de grande interesse nos estudos de variabilidade em *Plasmodium spp*., mostrou-se interessante para avaliação de uma boa anotação e montagem do genoma, isso, porque os genes pertencentes a esta família apresentam estrutura complexa e de difícil predição por parte dos programas automáticos. Os genes pertencentes ao Tipo I, podem apresentar mais de 16 exons e o Tipo II de 3 a 4 exons. Genes com grande quantidade de exons não são comuns em apicomplexas.

Um resumo dos genes SICAvar identificados nesse estudo está representado na Tabela 9 e na Figura 5. Vários genes incompletos ainda estão presentes, entretanto 23 genes, que antes estavam anotados como fragmentos, agora são considerados completos. No total, 129 genes completos foram preditos na anotação deste estudo, sendo 110 do Tipo I e 19 do Tipo II.

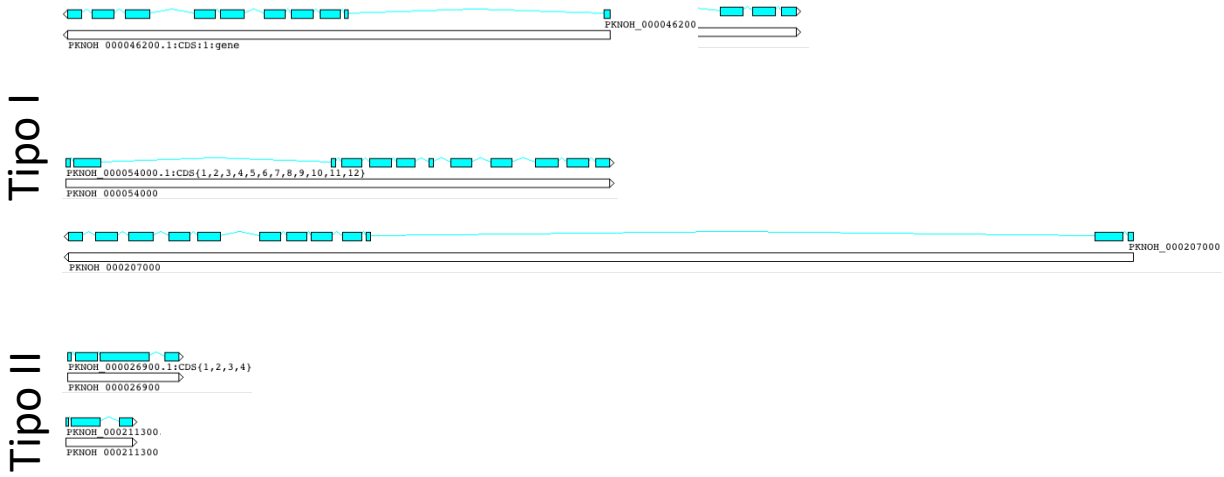

Figura 5: Genes SICAvar: Exemplos representativos de genes SICAvar de Tipo I e Tipo II com exons marcados em azul, e a sua direccionalidade indicada com setas colocadas no final do exon (3'). Os genes SICAvar de Tipo I são caracterizados por múltiplos exons (5-16), muitas vezes com introns extremamente grandes, particularmente entre os exons 2 e 3. Os genes SICAvar de Tipo II têm 3 ou 4 exons e são mais compactos com introns menores. Em 5 dos 6 exemplos mostrados, os dois exons iniciais mostrados são típicos dessa família gênica.

Tabela 9: SICAvar genes do Tipo I e II

	PKNOH		**PKNH**	
Tipo I	**Tipo II**	**Tipo I**	**Tipo II**	
Full *SICAvar* genes	117	19	87	14
SICAvar fragments	22	0	128	5
Exon #	11 (5–16)	3 (3–4)	11 (3–16)	3 (3–4)
Gene length (nt)	14 159 (6312–32 900)	3473 (2051–4269)	ND	ND
Coding length	5652 (834–6777)	1488 (963–3051)	5940 (1401–8295)	2630 (1302–3309)
Protein length (aa)	1892 (278–2783)	496 (321–1017)	1979 (466–2764)	876 (433–1102)
Alpha domains	133	138		
Beta domains	763	765		
C-terminal domains	133	143		

É importante ressaltar que a maioria dos genes SICAvar completos, quer sejam de Tipo I ou Tipo II, têm uma característica inicial dos primeiros dois exons em que se observou um motivo PEXEL (S. a Lapp, Korir, and Galinski 2009) e dois exons finais que codificam um

domínio transmembrana e Domínio citoplasmático conservado (Al-khedery, Barnwell, and Galinski 1999). Por estudar essas características previamente descritas, fomos capazes de conectar os exons por meio da anotação manual. É válido ressaltar que a alta qualidade da montagem do genoma levou a uma melhor anotação.

A partir da anotação combinada: automática e manual, foi predito que o genoma de PKNOH contém, 5258 genes codificadores de proteínas, 45 tRNAs, 12 rRNA. 4792 dos genes codificadores de proteínas são ortólogos de PKNH e PKNA1-H (Tabela 8, Figura 6).

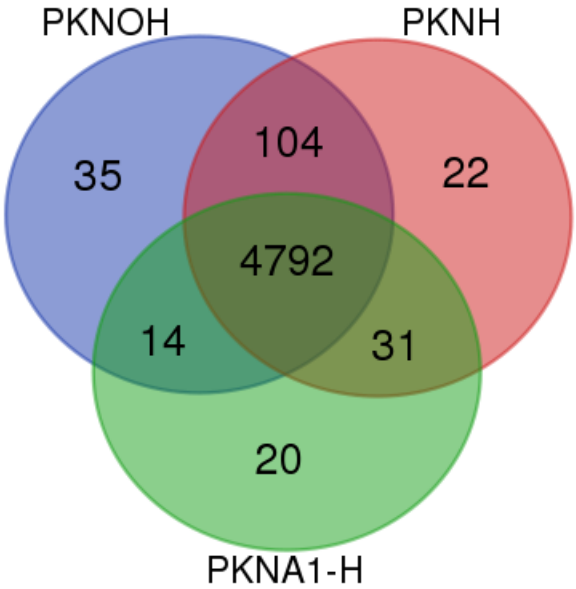

Figura 6: Diagrama de Venn Representando os Genes Codificadores de Proteínas dos Genomas Anotados de *P. knowlesi*. PKNOH = genoma montado e anotado (automático e manual) neste estudo. PKNH e PKNA1-H descritos na Tabela 1.

4.4 Considerações

A sequência completa do genoma de *P. coatneyi* foi disponibilizado pela primeira vez em bancos de dados públicos. Considera-se que a montagem e anotação do parasito alcançaram uma boa qualidade, mas ainda é preciso refinamento de ambas.

A nova montagem e anotação da sequência do genoma do parasito *P. knowlesi*, por sua vez, é considerado um estudo de alta qualidade, tendo sido gerado um genoma/anotação referência para a espécie.

Recursos de Bioinformática: Plataforma de Bioinformática Fiocruz Minas e
The Georgia Advanced Computing Resource Center (GACRC) University of Georgia

5 O PARASITO SCHISTOSOMA MANSONI

A última parte do Capítulo 1 abordará a anotação do parasita multicelular *Schistosoma mansoni*. Esse trabalho foi realizado no Wellcome Institute Sanger, UK, sendo todos os dados gerados e fornecidos pelo Instituto. O estudo do genoma deste parasito trouxe novos desafios ao desenvolvimento da ferramenta de avaliação da qualidade da anotação, mas também nos apresentou como os avanços tecnológicos dos sequenciamentos irão afetar o futuro não tão distante de todas as atuais análises genômicas.

5.1 O Genoma do Parasito

O parasito *Schistosoma mansoni* apresenta um ciclo de vida complexo com hospedeiro intermediário: caramujos do gênero *Biomphalaria spp*. e hospedeiro definitivo: humano. O genoma nuclear do parasito tem 380Mb, 7 cromossomos autossômicos mais os cromossomos sexuais ZW, aproximadamente 12 mil genes codificadores de proteínas (Berriman et al. 2009; Protasio et al. 2012). Exibe uma organização gênica complexa como, por exemplo, os Micro-Exons (do inglês polymorphic mucin genes - MEGs), Mucinas Polimórficas (polymorphic mucin genes - SmPoMucs) e Proteínas do tipo alérgeno (venom allergen-like - SmVALs) com estrutura genética incerta, orientada aparentemente para a geração de variabilidade proteica por meio de *splicing* alternativo e a presença de múltiplas cópias desses genes. Todas essas características indicam que o parasita desenvolveu um sistema sofisticado para interagir com seus hospedeiros (Crellen et al. 2016; Farias et al. 2011; Lu et al. 2016).

5.1.1 Montagem da Nova Versão do Genoma

Em julho de 2016, o Instituto Sanger compartilhou com os colaborados nos EUA, a chamada versão 6 (v6) do genoma de *Schistosoma mansoni* (até então só estava disponível publicamente a v5.2). Essa montagem foi gerada por meio de dados de sequenciamento PacBio e mapeamento óptico e com os dados antigos da plataforma Illumina. A montagem na v6, foi enviada para um refinamento por meio do cruzamento de metodologias e integração de mais mapas genéticos, no intuito de corrigir os *contigs*/*scaffolds* montados em locais errados na montagem. Toda a metodologia e a lista de marcadores podem ser acessadas no material suplementar da presente tese.

Em agosto de 2016, mudei o local do meu doutorado sanduíche, indo trabalhar no Wellcome Institute Sanger (Hinxton, UK). O objetivo da minha mudança temporária de grupo

foi para aprender mais sobre o processo de anotação estrutural dos genes codificadores de proteínas, utilizando dados de Sequenciamento de RNA-Seq PacBio. Ao mesmo tempo, pude coletar dados e ajudar a gerar um novo modelo para anotação da nova versão do genoma.

Apenas no primeiro semestre de 2017 o genoma foi finalizado, porém a qualidade alcançada foi superior à esperada. Os dados de PacBio agregados com o mapeamento óptico melhoraram muito a montagem do genoma. As informações de ligação do mapeamento genético permitiram identificar erros na montagem e atribuir muitos *scaffolds* desalinhados aos devidos cromossomos. Por consequência, a versão do genoma passou a ser chamada de v7.1 e não mais v6. Devido a este fator esse trabalho exibirá a comparação da v5.2 com a v7.1. No tempo em que estive no Instituto trabalhando na v5.2, para quando a v7.1 estivesse pronta, pudéssemos transferir os dados com mais facilidade.

5.2 Metodologia

Diante do exposto, essa parte do trabalho irá abordar o melhoramento da anotação da v5.2, a geração do modelo para treinamento dos programas e o trabalho com os dados de terceira geração de RNA-Seq, bem como a transferência da anotação para a nova versão do genoma.

5.2.1 ISO-Seq – Sequenciamento de RNA por PacBio

Os dados de RNA-Seq provenientes da plataforma de sequenciamento de terceira geração PacBio são chamados de ISO-Seq data[2]. Esse tipo de dado vem sendo recentemente muito utilizado para detecção de isoformas gênicas (*splicing* alternativo). Devido ao grande comprimento da *read* do sequenciador e da capacidade de gerar pelo menos uma *read* completa por transcrito, essa técnica tem se mostrado extremamente valiosa para melhoramento da anotação de genomas.

Diferentemente dos dados oriundos de sequenciamento de segunda geração, os dados de ISO-Seq não precisam ser montados, as *leituras* representam o comprimento completo do transcrito. Portanto, apenas o mapeamento das *leituras* é realizado contra o genoma. O resultado do mapeamento é em formato binário do tipo BAM. As duas Figuras (7,8) ilustram as vantagens do ISO-Seq em relação às tecnologias que geram pequenas *leituras* (Figura 7) e processo simplificado de geração dos dados, bem como o mapeamento contra o genoma de interesse (Figura 8).

[2] http://www.pacb.com/applications/rna-sequencing/

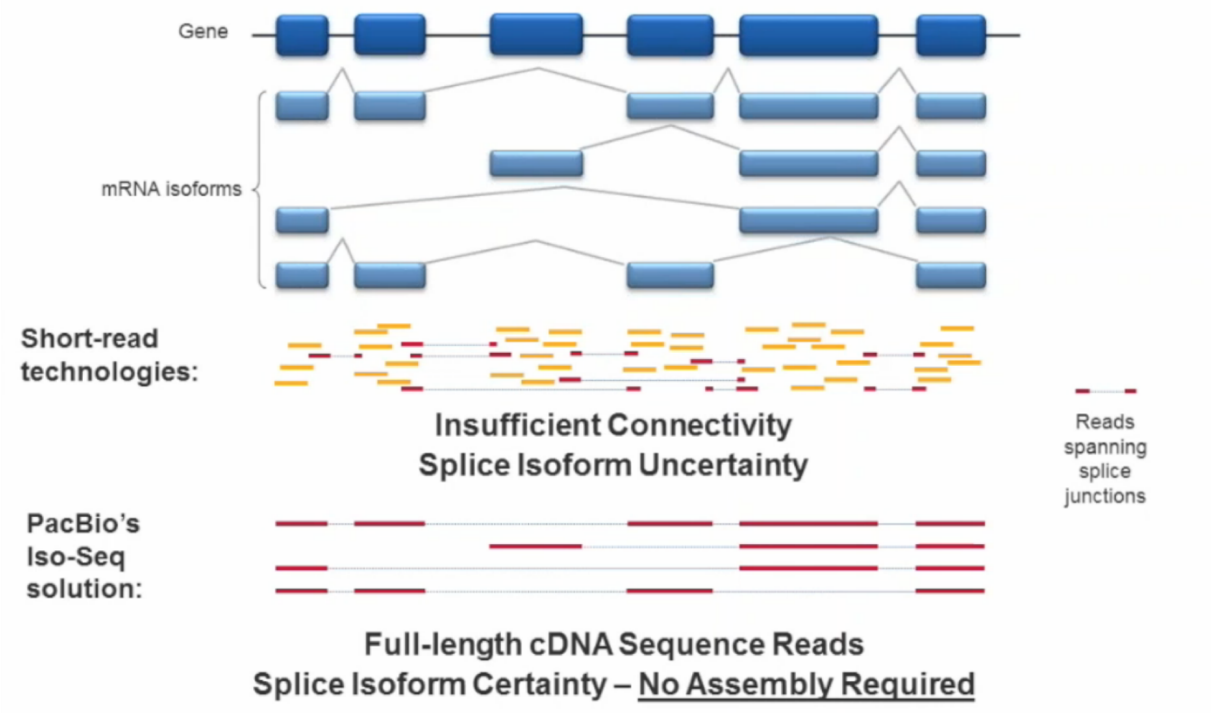

Figura 7: ISO-Seq: A aplicação de sequenciamento de isoformas (Iso-Seq) gera sequências de cDNA completas - do extremo 5 ' dos transcritos à cauda poli-A - eliminando a necessidade de reconstrução do transcritoma usando algoritmos de inferência de isoformas. O método também gera informações precisas sobre as junções de exons (splice junctions) e os locais de início da transcrição. Fonte: https://www.pacb.com/applications/rna-sequencing/.

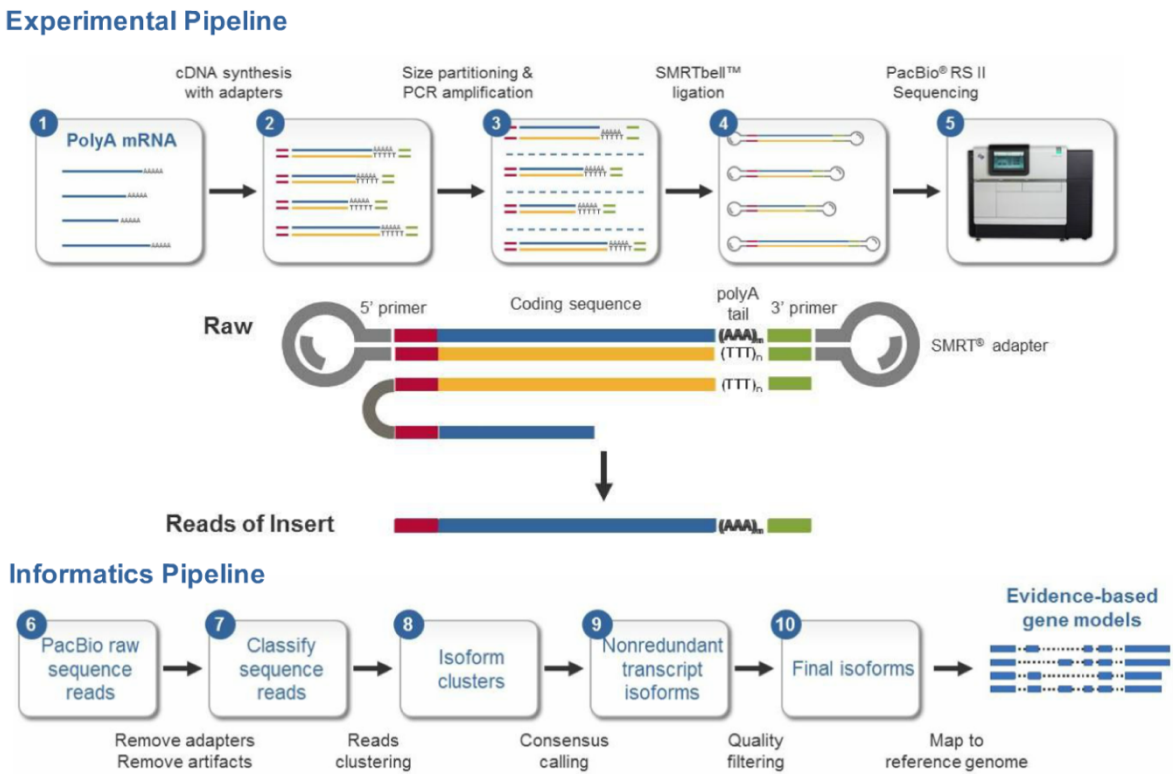

Figura 8: Pipeline de Geração das Leituras de ISO-Seq: O processo experimental passa pela adição de poli-A e amplificação do transcrito. Adaptadores são ligados nas extremidades e sequenciados. As sequências brutas do PacBio, passam por remoção dos adaptadores e de artefatos, em seguida pela classificação das sequências (completos e não completos – *full-length* e não *full-length*). Passam pelo Cluster de isoformas (número de isoformas de alta qualidade polida (HQ) --- Esse é o número de clusters HQ, cada cluster representando uma isoforma exclusiva. O processo da filtragem por qualidade é realizado e por fim, as leituras são mapeadas contra o genoma de referência. Fonte: https://www.pacb.com/applications/rna-sequencing/.

Um dos grandes desafios do método, é que por se tratar de nova metodologia não existem muitas referências e ferramentas disponíveis para trabalhar com esse tipo de dado e nem uma metodologia bem estabelecida. Recentemente, algumas ferramentas foram disponibilizadas: ToFu2[3], MatchAnnot[4], Angel[5] e SQANTI[6], entretanto nenhuma das ferramentas foi capaz de responder as nossas perguntas biológicas.

Em relação à taxa de erro do sequenciamento PacBio, esta é bem menor para o ISO-Seq, visto que cada posição em uma *read* é um consenso de várias rodadas do sequenciamento do DNA (circular). Estudos mostram essa taxa de erro perto de 2,34%, sendo 0,64% de *mismatches*, 1,07% de inserções, 0,63% de deleções (Abdel-Ghany et al. 2016).

[3] https://github.com/PacificBiosciences/IsoSeq_SA3nUP/wiki
[4] https://github.com/TomSkelly/MatchAnnot
[5] https://github.com/PacificBiosciences/ANGEL
[6] https://bitbucket.org/ConesaLab/sqanti

Com o intuito de melhor aproveitar estes dados em nossos estudos, o presente trabalho se propôs a desenvolver uma estratégia para avaliar e melhorar a anotação do genoma utilizando apenas os dados de ISO-Seq. Essa estratégia deve ser fácil de reproduzir, para poder ser reaplicada na versão 7 do genoma.

5.2.1.1 Mapeamento das Leituras:

Foram geradas 3 bibliotecas de ISO-Seq sendo 1 para Esquistossômulo, 1 Verme Adulto Fêmea, 1 Verme Adulto Macho.

Os adaptadores foram removidos das sequências brutas, em seguida foram classificadas em *Full length* e não *Full length* (completas, 5' -3' e incompletas), clusters foram formados entre os transcritos iguais (removendo redundância), as *leituras* foram filtradas por qualidade de 95%, e por as sequências foram mapeadas contra o genoma referência de *S. mansoni* v5.2 e posteriormente na v7.1.

Todas as etapas foram realizadas por meio da plataforma fornecida pela PacBio: SMRT-Analysis, utilizando o protocolo RS_IsoSeq.v2.3 seguindo as configurações padrões, com exceção do filtro de qualidade que foi aumentado para 95%. O mapeador utilizado por este protocolo é o GMAP (Wu and Watanabe 2005). Todas as configurações podem ser encontradas no protocolo RS_IsoSeq.v2.3 [7], representado no Quadro 1. Protocolo completo está disponível no material suplementar 1.

[7] https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-(v2.3)-Tutorial-%231.-Getting-full-length-reads

Quadro 1: Protocolo RS_IsoSeq.v2.3

```
<?xml version="2.3" encoding="utf-8"?><smrtpipeSettings>
  <protocol version="2.3.0" id="RS_IsoSeq.1" editable="false">
  <application>De novo assembly</application>
    <param name="name" label="Protocol Name">
      <value>RS_IsoSeq</value>
      <input type="text"/>
      <rule required="true"/>
    </param>
    <param name="description">
      <value>Classify, de novo cluster, and map cDNA sequences.</value>
      <textarea />
    </param>
    <param name="version" hidden="true">
      <value>1</value>
      <input type="text"/>
      <rule type="digits" required="true" min="1.0"/>
    </param>
    <param name="state" hidden="true">
      <value>active</value>
      <input value="active" type="radio"/>
      <input value="inactive" type="radio"/>
    </param>
    <param name="reference" hidden="true">
      <value></value>
      <rule required="false"/>
    </param>
    <param name="control" hidden="true">
        <value></value>
    </param>
    <param name="fetch" hidden="true">
      <value>common/protocols/preprocessing/Fetch.1.xml</value>
    </param>
    <param name="filtering">
      <value>common/protocols/consensus/IsoSeq_ReadsOfInsert.1.xml</value>
      <select multiple="true">
        <import                                contentType="text/directory"
extension="xml">common/protocols/consensus</import>
      </select>
    </param>
    <param name="barcode" editableInJob="true" hidden="true"/>
    <param name="isoseq_classify">
      <value>common/protocols/isoseq/Classify.1.xml</value>
    </param>
    <param name="isoseq_cluster">
      <value>common/protocols/isoseq/Cluster.1.xml</value>
    </param>
    <param name="isoseq_report" hidden="true">
      <value>common/protocols/isoseq/IsoSeqReports.1.xml</value>
    </param>
  </protocol>
  <moduleStage name="fetch" editable="true"/>
  <moduleStage name="filtering"/>
  <moduleStage name="barcode"/>
  <moduleStage name="isoseq_classify"/>
  <moduleStage name="isoseq_cluster"/>
  <moduleStage name="isoseq_report"/>
</smrtpipeSettings>
```

5.2.2 Avaliação da anotação com os dados de ISO-Seq

Para entender melhor a problemática da utilização dos dados do ISO-Seq na avaliação da anotação do genoma, um levantamento das informações foi realizado por meio da curadoria manual na versão pública (ano de 2016) do genoma (*Schistosoma mansoni* versão 5.2).

Fazendo a curadoria manual dos genes codificadores de proteínas do genoma[8], foi possível levantar casos em que o ISO-Seq forneceu evidências suficientes da estrutura do gene. De maneira geral, os genes presentes no arquivo de anotação (GFF3) da v5.2 foram comparados com as sequências já mapeadas no genoma, primeiramente por meio de visualização e edição no Programa Artemis (Rutherford et al. 2000b), e posteriormente de maneira mais automatizada, como será descrito no próximo tópico.

A Figura 9 ilustra um esquema representando três possíveis casos encontrados na metodologia para avaliação da qualidade da anotação do genoma utilizando apenas dados de ISO-Seq. Na linha gene está representado o comprimento do gene e as três linhas abaixo representam a cobertura pelas *leituras* do ISO-Seq:

- Caso 1: Correspondência perfeita (*Perfect match*), em que as *leituras* ISO-Seq correspondem exatamente as coordenadas de início e fim do gene, bem como de todos os exons que constituem esse gene.
- Caso 2: Gene maior que ISO-Seq: em que as leituras do ISO-Seq são menores que o gene predito.
- Caso 3: Gene menor que ISO-Seq: as leituras ISO-Seq cobrem regiões não preditas pela anotação.

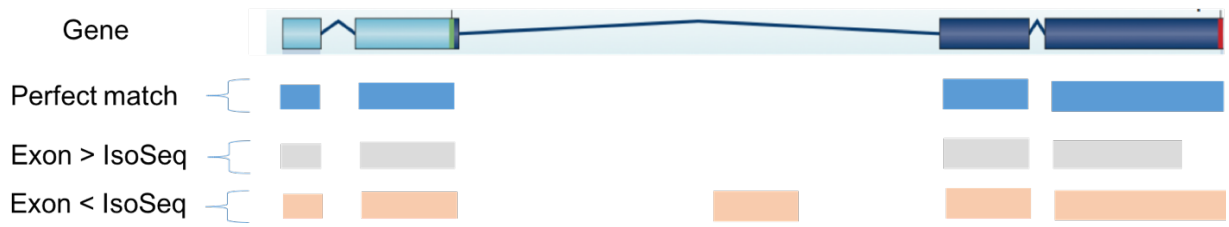

Figura 9: Avaliação da Qualidade da Anotação por ISO-Seq. A Primeira linha representa o gene presente no arquivo anotado (GFF). Na segunda linha está presente o caso de perfeita correspondência entre os arquivos GFF e BED indicando que a sequência do ISO-Seq tem o mesmo comprimento do gene predito. As últimas duas linhas representam casos em que o ISO-Seq é menor e casos onde este é maior do que o gene predito.

[8] ver http://www.genedb.org/Homepage/Smansoni Annotation statistics

5.2.2.1 Automatização da Avaliação da Anotação com os Dados de ISO-Seq

Por meio da curadoria manual foi possível fazer o levantamento de casos das possíveis divergências entre o proteoma predito e o ISO-seq.

O script desenvolvido para detectar de maneira automática os casos levantados pode ser acessado em: https://github.com/Juassis/SharingCodes/tree/master/ISOseqPacBio. Basicamente o *script* funciona da seguinte forma: dois arquivos de entrada devem ser fornecidos: 1-) o arquivo do proteoma predito (GFF3) e 2-) o arquivo do mapeamento em formato BED. O arquivo BAM, gerado pelo mapeamento do PacBio deve ser convertido em formato BED por meio da ferramenta do BedTools (Quinlan and Hall 2010).

- É realizada a interseção das coordenadas da *feature* mRNA (com todos os exons) do arquivo GFF e das *leituras* ISO-Seq mapeadas em formato BED, para verificar o comprimento total do mRNA coberto por uma ou mais reads.
- A sobreposição entre as coordenadas é verificada;
    - Haja vista que o sequenciamento por PacBio ainda apresenta dificuldades em trabalhar com homopolímeros, podendo, portanto, ter acrescentado ou excluído alguma base, um pequeno desvio foi permitido para detectar as regiões de sobreposições entre as coordenadas (desvio *1.0001).

Especificamente neste caso do genoma de *S. mansoni*, as regiões UTRs foram excluídas. Essas regiões são inúmeras vezes compartilhadas por mais de um gene. A informação das regiões de UTR fica a critério do usuário do *script*, podendo incluir ou não essa informação.

Essa estratégia foi aplicada tanto para a versão v5.2 quanto para a versão v7.1 do genoma. O intuito da aplicação na v5.2 do genoma, foi poder gerar modelos gênicos confiáveis para serem transferidos para a nova anotação, bem como para treinar novos modelos gênicos.

5.2.3 Anotação da v7.1 do Genoma de *Schistosoma mansoni*

A anotação dos genes codificadores de proteínas foi realizada de duas maneiras:

a) transferência do modelo gênico da v5.2 para a v7.2 (após correção manual); e
b) predição gênica baseada em evidências (ISO-Seq + illumina)

Após a correção manual dos genes que apresentaram divergência entre o proteoma predito e os dados de ISO-Seq, a transferência dos modelos gênicos curados foi realizada por meio do programa RATT (Otto et al. 2011).

Para criar um conjunto de treinamento para o preditor Augustus, a estratégia de incorpoeração de dados do ISO-Seq e Illumina (RNA-Seq) foi adotada. Os dados de ilumina utilizados foram os dados já publicados (todas as fases de vida do parasito, excesso ovo), todo o protocolo de mapeamento foi realizado seguindo os passos do artigo de PROTASIO; DUNNE; BERRIMAN, (2013).

O nome dado à estratégia de integração desse tipo de dados é a incorporação de *hints*. Um "*hint*" refere-se a uma região, podendo conter informações sobre a fase de leitura (frame) fita (*strand*), sítios de *splicing* e uma pontuação de confiabilidade (score), bem como informações sobre a fonte de evidência para cada transcrito. O conjunto de *hints* foi gerado a partir dos dados de mapeamento do RNA-Seq contra o genoma.

5.2.3.1 Avaliação da Qualidade da nova Anotação e Curadoria Manual

Com o resultado da anotação, o *script* de cobertura: ISOSeqPacBio.pl foi rodado. A Curadoria manual foi executada nos transcritos os quais apresentaram alguma divergência dos dados do ISO-Seq. O processo de curadoria manual ocorreu por meio do programa WebApollo, pela equipe de curadores do GeneDB responsáveis pelo parasito (a qual pude fazer parte).

A Figura 10 apresenta um esquema do processo de anotação da montagem v7.1.

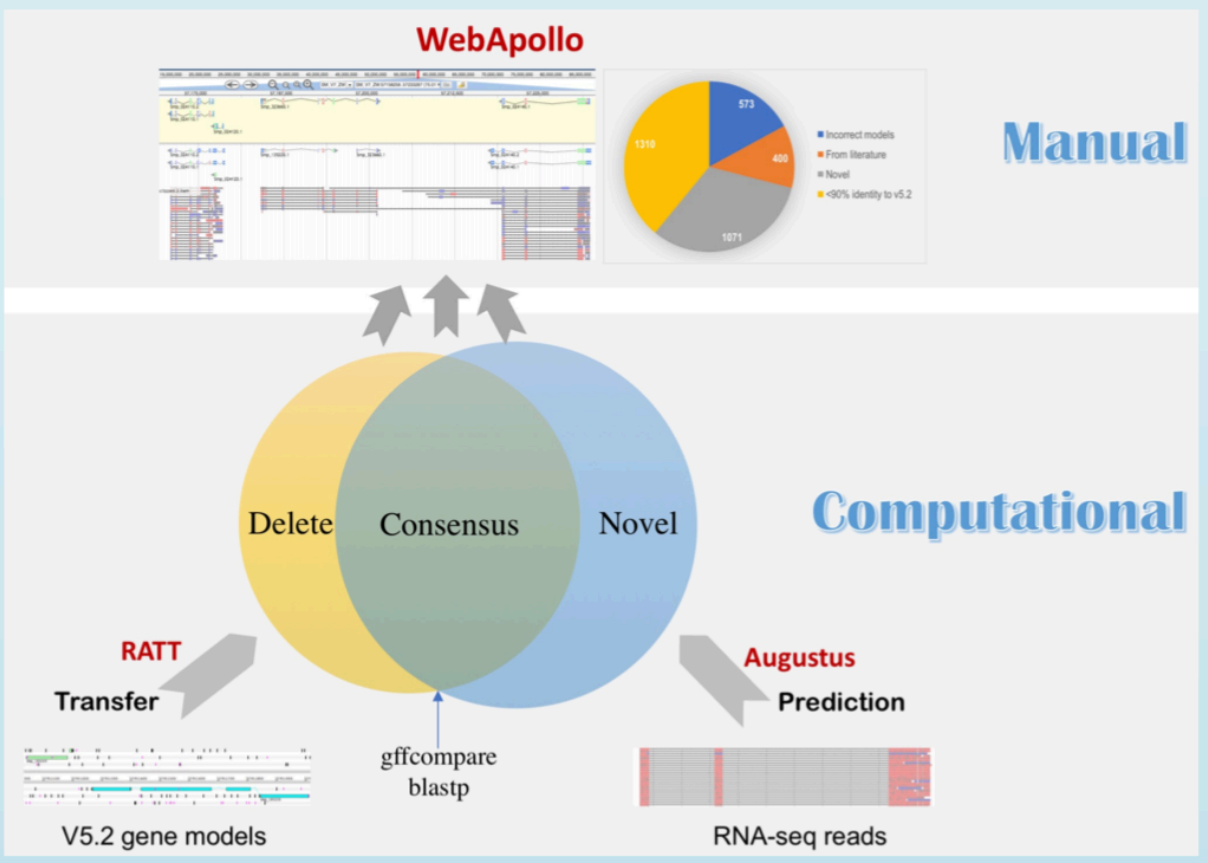

Figura 10: Pipeline de Anotação Estrutural usando a Ferramenta de Transferência de Anotação RATT, o preditor de Genes Augustus e a Ferramenta de curadoria WebApollo. Evidências usadas para Augustus incluem RNA-Seq de Illumina de todas as fases da vida (exceto ovo), e PacBio ISO-seq para três fases de vida do parasito.

5.3 Resultados

O mapeamento foi realizado contra a montagem do genoma na v5.2 de agosto de 2016. O total de 2.836,659 *leituras* foram classificadas como completas e sem quimeras (aproximadamente 18% do total de *leituras* geradas). A média dessas *leituras* foi de 2.518pb, o que se aproxima do tamanho médio dos transcritos de *S. manosni* que é de 2.480pb (o tamanho da *leitura* gerada pelo ISO-Seq deve ser o tamanho real do transcrito).

A Figura 11 mostra a quantidade de genes codificadores de proteínas com 100% de cobertura em extensão (*Completeness Coverage*) para cada uma das 3 bibliotecas geradas e para a união das três. Do total de 11.832 genes codificadores de proteínas, 8.450 apresentaram cobertura de 100%, sendo que pelo menos uma *leitura* ISO-Seq cobriu todo esse gene.

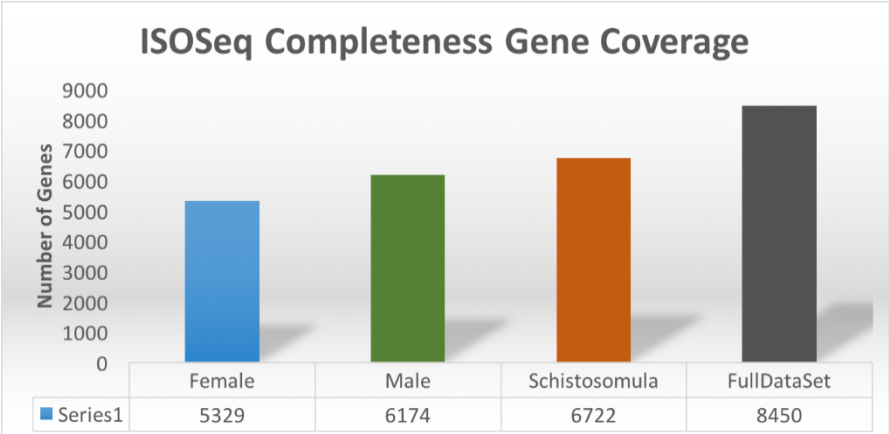

Figura 11: Cobertura de 100% em Extensão dos Genes Codificadores de Proteínas. Em azul estão representados os genes completamente cobertos para os vermes adultos fêmeas, em verde vermes adultos machos, em laranja os esquistossômulos e em cinza escuro a união dos três.

Os resultados do mapeamento contra a v7.1 foram muito próximos aos descritos para a v5.2, os resultados podem ser acessados no Banco de Dados do GeneDB: https://www.genedb.org/#/species/Smansoni.

5.3.1 Resultados da Avaliação da Anotação com os dados de ISO-Seq na v5.2 do Genoma

Caso 1: Correspondência perfeita (*Perfect match*).

No total 75% dos exons apresentaram correspondência completa do início ao fim da sequência. A Figura 12 ilustra dois exemplos de genes que apresentaram correspondência total, de todos os exons, entre o transcrito predito e as *leituras* de ISO-Seq.

Figura 12: Transcritos Preditos com Total Concordância com os dados de ISO-Seq. A Figura mostra dois genes sendo visualizados no programa Artemis, em que os dados do transcritoma mapeados foram adicionados. Letra A: Gene predito está em azul claro, com 7 quadrados azuis representando os exons. A cobertura por RNA-Seq está representada no primeiro quadro, onde as partes em cinza escuro indicam os exons. Letra B: representa gene predito em amarelo, os dados de RNA-Seq foram colocados em outro modo de visualização no programa Artemis, permitindo visualizar a sequência para cada exon.

Caso 2: Exon maior que o ISO-Seq: onde as sequências do ISO-Seq são menores que o exon predito (em quantidade e/ou comprimento). A Figura 13 apresenta um exemplo da complexidade do genoma de *S. mansoni*, onde os dados de ISO-Seq conseguiram ajudar a resolver os erros de predição do gene.

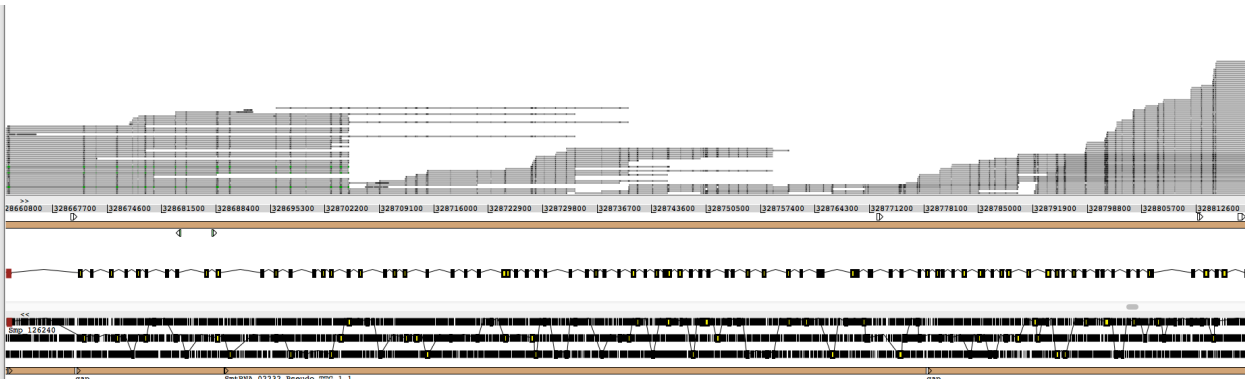

Figura 13: Número de Exon maior que o Suporte por ISO-Seq. A Figura mostra um gene com 94 exons preditos e três isoformas (apenas 88 exons e duas isofromas na imagem). Os dados de RNA-Seq estão representados na parte superior da figura, onde o cinza escuro representa os exons. Alguns destes exons não correspondem aos preditos, representado embaixo em amarelo.

Existem diversos outros casos, onde o ISO-Seq está presente e não existe nenhuma predição gênica no local. Bem como casos onde não existe a presença do ISO-Seq no modelo gênico predito (Ver material suplementar 2 com ilustração de todos os casos encontrados). O *script* ajudou a reportar todos os casos de divergência, mas todo o processo de correção de 2.638 transcritos foi realizado manualmente.

5.3.2 Resultados da Anotação da v7.1 do Genoma de *Schistosoma mansoni*

Foi realizada uma anotação de alta qualidade para a atualização da montagem do genoma do parasito *Schistosoma mansoni*. É necessário ressaltar que a qualidade da montagem influenciou em um bom resultado da anotação. A v5.2 era uma versão fragmentada, onde vários genes estavam incompletos, já a v7.1, apresenta uma montagem contínua.

Os genes codificadores de proteínas preditos e transferidos totalizaram 10.845, com um total de 15.984 proteínas. É possível acessar a versão Sm_V7_prep.gff no material suplementar 3.

5.3.2.1 Avaliação da Qualidade da Anotação e Curadoria Manual

Com o resultado da anotação, o *script* de cobertura: ISOSeqPacBio.pl foi rodado. No total, 8.437 transcritos (incluindo isoformas do mesmo gene), apresentaram concordância completa com as leituras do ISO-Seq (Tabela com os identificadores disponível no material suplementar 4). A Figura 14 evidencia alguns casos onde foi necessário curadoria manual, representando genes modificados entre as versões do genoma.

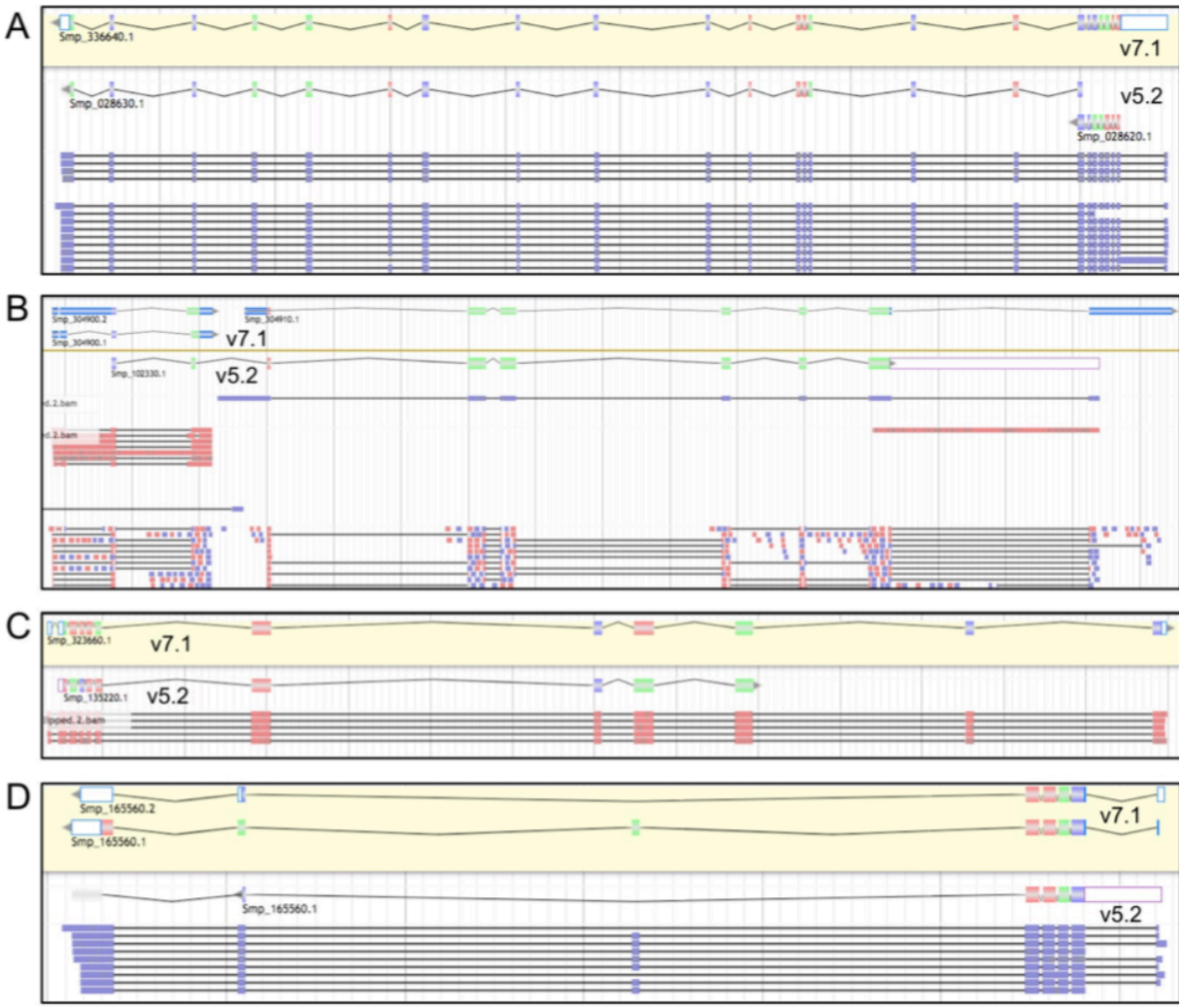

Figura 14. **Mudanças no Modelo de Genes Exemplares Comparadas com a v5.2.** A) Dois genes v5.2 foram fundidos em um gene; B) Um gene v5.2 foi dividido em dois genes; C) O modelo estrutural de 5,25 foi modificado; D) Isoformas adicionais para um gene v5.2

A Tabela 10 apresenta a comparação da v7.1 com a v5.2. após correções manuais.

Tabela 10: Comparação das montagens do genoma v7.1 x v5.2

	v7.1	**v5.2**
Protein-coding genes	10129	10116
Transcripts	14520	11075
New / deleted genes	872**	845
Alternative splicing	27.70%	6.90%
One to one	7139	
Merging	343	746
Splitting	182	95
Struc. change (iden. < 80%)	1012	1012
Pfam	3425+16	3425+40
GO terms	1466+33	1466+17
KEGG Orthologues	3693+253	3693+115

* Based on the gene set from GeneDB as of 10/07/2017 ** Excluding haplotype contigs

A anotação completa pode ser acessada em https://www.genedb.org/#/species/Smansoni. Arquivo Sm_V7_1_prep.gff.

5.4 Considerações

Uma montagem de alta qualidade foi gerada, consequentemente e por meio da integração dos dados de ISO-Seq, uma anotação de alta qualidade também foi alcançada. Com os dados de sequenciamento por RNA-Seq na plataforma PacBio (ISO-Seq), foi possível resolver diversos erros de anotação, particularmente, as características de micro-exons presentes no genoma do parasito puderam ser solucionadas.

Recursos Computacionais: Wellcome Sanger Institute.

6 CONSIDERAÇÕES SOBRE O PRIMEIRO CAPÍTULO

A anotação dos genes codificadores de proteínas dos quatro genomas aqui descritos, foram capazes de fornecer dados valiosos para o levantamento de informações para o desenvolvimento da ferramenta que será descrita no próximo capítulo, bem como forneceram novos dados com qualidade aceitável para a comunidade científica.

O genoma da onça, montagem fragmentada com aproximadamente 5 mil *scaffolds*, permitiu que o desafio da anotação em um genoma fragmentado fosse realizado. A anotação publicada ainda apresenta erros, muitos destes causados pela falta de cobertura e/ou erros de montagem. Outros, porém, poderiam ter sido evitados com uma curadoria manual. O projeto continua em andamento e a montagem e anotação estão sendo melhorados.

O genoma do *Plasmodium coatneyi*, sequenciado unicamente com dados de PacBio, apresentou uma montagem de alta qualidade, porém com algumas regiões dúbias, como por exemplo as regiões finais dos cromossomos 2 e 12 (dados discutidos no artigo). A anotação por sua vez, foi facilitada pela continuidade dos cromossomos, mas novamente, como no genoma da onça, não houve curadoria manual. Vários genes, principalmente os multicópias, apesentam características de difícil previsão por anotadores automáticos. Outra característica observada foi a taxa de erro de sequenciamento, a qual, muito provavelmente, levou à predição errôneas de *pseudogenes*.

Outro genoma estudado foi do outro apicomplexa: *Plasmodium knowlesi*. Com os aprendizados da montagem e anotação de *P. coatneyi*, foi possível realizar um desenho experimental em que não somente dados de PacBio foram utilizados, dessa vez, foram adicionados ao estudo dados de Hi-C, os quais ajudaram a co-localizar regiões de grande importância do gênero *Plasmodium sp.*. Os 14 cromossomos foram montados, mas ainda assim, pequenas regiões não apresentaram cobertura suficiente para serem integradas aos devidos cromossomos. A anotação do genoma foi realizada com muita cautela e por meio de modelos gênicos muito confiáveis. Dessa vez, a curadoria manual foi efetuada para mais de 1.000 genes, os quais apresentaram divergências entre o genoma previamente publicado do parasito e o genoma aqui montado, e para as famílias multigênicas, estas últimas devido à peculiaridade da estrutura dos exons nestes genomas. Sendo assim, foi possível estabelecer uma nova referência de genoma e anotação da espécie.

Por fim, foi descrito na presente tese o estudo do genoma do parasito multicelular *Schistosoma mansoni*. Diferentemente dos genomas anteriores, a sequência do genoma de *S. mansoni* foi apresentada pela primeira vez há mais de dez anos, desde então, a comunidade

científica vem contribuindo com o melhoramento da anotação do genoma, principalmente dos genes codificadores de proteína. Devido aos avanços tecnológicos, o genoma foi ressequenciado, agora pela terceira vez. Deste modo, uma montagem de alta qualidade foi gerada. Mesmo diante de uma montagem de alta qualidade, a anotação dos genes codificadores de proteínas continuava a ser um grande desafio, principalmente devido às características de micro-exons presentes no genoma. Com os dados de sequenciamento por RNA-Seq na plataforma PacBio (ISO-Seq), foi possível resolver diversos erros em modelos gênicos automaticamente previstos e manualmente editados.

Nem mesmo com bons modelos gênicos treinados, os programas de anotação automática foram imunes a erros de predição. A curadoria manual dos genes é essencial para uma anotação confiável, todavia realizar a curadoria manual de um genoma completo demanda tempo. Desenvolver uma ferramenta que aponte para os genes que devem ser conferidos manualmente reduzirá o tempo de curadoria manual, podendo dar mais ênfase aos genes apontados a fim de corrigir possíveis erros.

Diante do exposto, o capítulo 2 abordará o desenvolvimento da ferramenta de avaliação da anotação automática de genes codificadores de proteínas em genomas completos de organismos eucariotos, utilizando os dados gerados no capítulo 1.

CAPÍTULO 2
DESENVOLVIMENTO DA FERRAMENTA

O objetivo central do presente estudo consistiu em propor uma nova ferramenta de avaliação da qualidade da anotação estrutural dos genes codificadores de proteínas em genomas completos de eucariotos por meio da integração de dados.

O trabalho foi embasado em algumas questões centrais: seria possível otimizar a curadoria manual dos genes codificadores de proteínas? Poderia a integração de dados multi-ômicos e de genômica comparativa contribuir para detecção de erros de anotação?

A ferramenta foi desenvolvida em linguagem de programação PERL. Esta, pode ser acessada em: https://github.com/Juassis/ASSIS. Onde encontra-se disponível um tutorial completo e arquivos de teste, bem como o manuscrito a ser submetido.

7 MOTIVAÇÃO

A principal motivação do presente estudo consistiu em realizar a avaliação da qualidade de todos os genes codificadores de proteínas anotados do genoma eucarioto de interesse, sem muito esforço humano. Por três razões: a avaliação humana pode introduzir viés, o tempo é muito importante na "era de *big data*" a qual estamos vivendo, e a principal: reduzir a quantidade de erros nos genomas anotados.

7.1 Avaliação da Qualidade da Anotação do Genoma

ASSIS: Assessing Syntenic Score Integrated System: A Platform to Evaluate Eukaryotic Whole Genome Annotation. ASSIS é um pipeline de avaliação da qualidade da anotação dos genes codificadores de proteínas em genomas completos de eucariotos. A finalidade é permitir que alguns projetos de anotação de genomas de eucariotos avaliem a qualidade dos genes codificadores de proteínas de maneira mais automatizada. ASSIS identifica genes com possíveis erros na anotação e sintetiza automaticamente esses erros em uma escala de pontuação de 1-6 em um novo arquivo de anotação.

Os genes com possíveis erros detectados recebem uma baixa pontuação (mais próximo de 1), enquanto aos genes confiáveis é atribuída uma pontuação mais alta (mais próximo de 6). Assim, o novo arquivo de saída gerado pode ser carregado diretamente em programas como

WeApollo, Artemis, IGV para executar uma curadoria manual naqueles genes com baixa pontuação, reduzindo o tempo de curadoria manual da anotação.

O sistema permite realizar a detecção de erros, utilizando informações da sintenia de genes ortólogos de espécies intimamente relacionadas e por integração de dados de RNA-Seq e informações da estrutura da anotação do gene. No total, o programa contém três módulos: 1-) Sintenia de Ortólogos, 2-) Estrutural e 3-) Transcricional, sendo permitido trabalhar separadamente.

ASSIS é especialmente útil para avaliar anotações de novos genomas, porém este pode ser aplicado a qualquer outra anotação de genoma disponível. Nenhuma grande experiência em bioinformática é necessária, no entanto é fortemente recomendado estudar a característica do genoma antes de iniciar a análise para aumentar a confiabilidade do resultado.

Nenhum gene é removido da anotação, apenas um relatório é gerado contendo todas as pontuações atribuídas a cada gene codificador de proteína. Os arquivos de saída gerados estão no formato Tabular do tipo texto e no formato GFF3.

7.1.1 O que o programa faz?

-> Identifica genes codificadores de proteínas com erros de anotação estruturais;
-> Produz valores de qualidade baseados em evidências para cada gene na anotação do genoma;
-> Gera novo arquivo de anotação com genes e/ou regiões problemáticos para serem corrigidos;

7.1.2 Os três módulos do programa

A ideia geral é classificar os genes codificadores de proteínas pela pontuação de qualidade. Para atingir este objetivo, três diferentes metodologias foram desenvolvidas:

7.1.2.1 Sintenia de Ortólogos

Espera-se um alto nível de sintenia de ortólogos em espécies intimamente relacionadas. Teoricamente, dois genes ortólogos compartilham genes vizinhos ortólogos, no entanto há uma pequena chance de ocorrerem coincidências de homologia por acaso (Jun, Mandoiu, and Nelson 2009). Além disso, rearranjos, inserções e deleções podem levar à perda de sintenia local entre genes ortólogos. Para contabilizar a pontuação desses eventos esperados foi desenvolvido um tipo de abordagem em fila, procurando determinar um tamanho ideal de janela e probabilidade

de correspondência que pudesse identificar de forma confiável os genes ortólogos e os genes ausentes/duplicados/parciais com base na sintenia local.

Outros dois módulos foram desenvolvidos no pipeline: Estrutural e Transcricional para aumentar a precisão das pontuações atribuídas, principalmente para cobrir os casos reais de falta de sintenia acima citados, como: rearranjos, inserções/deleções.

7.1.2.2 Estrutural

A avaliação estrutural consiste na análise da estrutura dos genes. Lembrando que anotação estrutural do genoma é o processo de identificação dos genes, como as estruturas intron-exônicas.

O módulo é capaz de detectar:

-> Códons de Início e Fim ausentes;

-> Erros na fase de leitura (*frameshift*);

-> Genes aninhados (genes com sobreposição);

Todos os prováveis genes problemáticos são relatados para serem inspecionados manualmente.

7.1.2.3 Transcricional

A evidência de transcritos é utilizada para melhorar a qualidade da pontuação nos genes. A cobertura em extensão é contabilizada por meio da ferramenta do BedTools (Quinlan and Hall 2010) e adicionada ao resultado do programa.

7.2 Metodologia

A estrutura do programa é apresentada na Figura 15. O genoma de estudo deve ser fornecido, bem como um ou mais genomas de espécie (s) referência(s) intimamente relacionada(s) ao genoma de estudo. Espécies intimamente relacionadas são adicionadas para avaliação usando a correspondência de ortólogos.

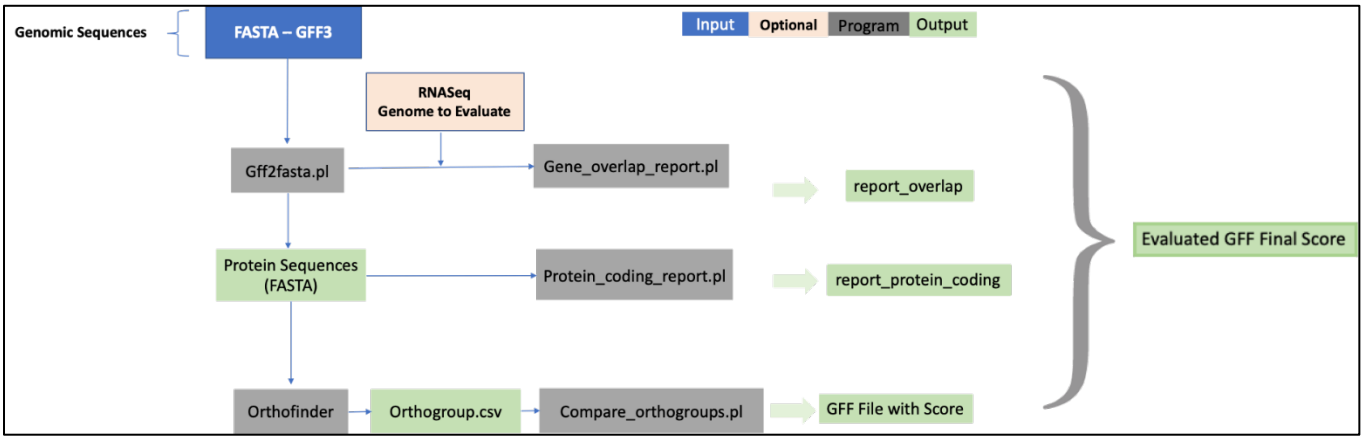

Figura 15: Visão Geral do Programa: Em azul estão representados os arquivos de entrada, FASTA e GFF3. Em rosa claro o arquivo de mapeamento, opcional para o programa. Em cinza estão representados todos os códigos desenvolvidos. Os quadrados verdes representam os arquivos de saída gerados pelo programa.

As sequências dos genes codificadores de proteínas são extraídos dos arquivos dos genomas (FASTA + GFF3) e traduzidos em aminoácido. O proteoma predito extraído é direcionado aos módulos do programa: Sintenia de Ortólogos e Estrutural.

7.2.1 Sintenia de Ortólogos:

Com os arquivos de proteínas extraídos a predição de ortólogos é realizada por meio do programa OrthoFinder. O arquivo de saída do OrthoFinder: Orthogroups.csv, é o arquivo de entrada para o módulo de sintenia de ortólogos do pipeline ASSIS. Orthogroups.csv é um arquivo de texto separado por tabulações. Cada linha contém os genes pertencentes a um único grupo de ortólogo (ortogrupo). Os genes de cada ortogrupo são organizados em colunas, uma por espécie. Como os ortólogos podem ser um para um, um para muitos ou muitos para muitos, no pipeline ASSIS, essas categorias são divididas em: Único (um-para-um), Múltiplo (um-para-muitos e muitos-para-muitos) e Não Atribuídos (genes que não foram atribuídos a nenhum ortogrupo).

Partimos da premissa de que dois genes ortólogos devem compartilhar genes vizinhos ortólogos. A comparação é realizada gene a gene (codificador de proteína) utilizando o arquivo de anotação (GFF3) para cada espécie. Os arquivos GFF3 são ordenados por Cromosomo/Scaffold, com base nas coordenadas genômicas da *feature* de mRNA (ou aquela indicada pelo usuário). O primeiro gene ordenado no genoma de Estudo é cruzado para o conjunto de ortólogos no arquivo tabular do OrthoFinder. O bloco sintênico começa quando a primeira correspondência entre o gene do organismo de estudo e o conjunto de ortólogos únicos

ocorre. O próximo passo é continuar a comparação da vizinhança ortóloga, incluindo todas as informações sobre os ortólogos: Grupos únicos, múltiplos e não atribuídos. O bloco sintênico é quebrado quando ocorrem duas diferenças contínuas.

O algoritmo do módulo sintenia de ortólogos está representando no Quadro 2:

Quadro 2: Algoritmo do Módulo de Sintenia de Ortólogos

Genoma de Estudo = Genoma a ser avaliado
Genoma de Referência = Espécie intimamente relacionada
Arquivo Ortólogo = Orthogoups.csv gerado pelo Orthofinder

1. Procura-se pelo primeiro gene único no genoma de Estudo por chr/scaffold. Pesquisa-se no arquivo dos ortólogos, o gene que pertence ao mesmo grupo ortólogo.
2. Avança-se para o próximo gene no chr/scaffold correspondente.
    2.1. Se ambos os genes: Estudo e Referência possuirem o mesmo grupo de ortólogo, pontua-se 5 e avança-se para o próximo gene.
    2.2. Se o gene do organismo de Estudo não possuir um grupo ortólogo, atribui-se uma pontuação 2 para este gene e o cursor avança para o próximo gene do genoma referência.
    2.3. Se o gene da Referência não possuir um grupo ortólogo o cursor avança para o próximo gene do genoma de Referência.
    2.4 Se ambos os genes apresentarem grupos ortólogos e este grupo ortólogo não for igual:
        2.4.1. Verifica-se se o grupo de ortólogo do genoma de Estudo existe na janela dos próximos 4 genes da Referência, se o mesmo não existir, pontua-se o gene com o valor 2. A mesma lógica para o gene da Referência.
        2.4.2. Se o grupo do gene do genoma de Estudo existe na janela da Referência e o gene da Referência existe na janela do geoma de Estudo, é construido um bloco de 5 genes (o atual gene de comparação, mais os próximos quatro) em ambos os ortólogos e pontua-se esse grupo da seguinte maneira:
            2.4.2.1 Se todos os grupos de ortólogos, em ambas as janelas são idênticos, pontua-se o valor 4 para todos os genes nesta janela,
            2.4.2.2 Se há apenas um grupo de ortólogos no genoma de Estudo diferente, pontua-se o valor 3; se estes apresentarem mais de um gene diferente, é atribuído o valor 2 para toda a janela.

A pontuação atribuída pode ser resumida da seguinte maneira representada no Quadro 3.

Quadro 3: Resumo da Pontuação Atribuída no Módulo de Sintenia de Ortólogos

Pontuação 0: Scaffolds com menos de 10 genes (parâmetro pode ser editado pelo usuário);
Pontuação 1: Genes do início do scaffold/chr, provavelmente pertencentes às famílias multigênicas;
Pontuação 2: Genes do genoma de Estudo que não pertencem à próxima janela da Referência; ou Genes na janela de comparação do genoma de Estudo com mais de 2 genes fora da janela da Referência;
Pontuação 3: Genes na janela de comparação do genoma de Estudo com apenas um gene fora da janela da Referência;
Pontuação 4: Genes na janela de comparação do genoma de Estudo, onde todos os genes estão na janela de comparação da Referência, com um ou mais genes fora de ordem;
Pontuação 5: Gene idêntico ao gene comparado ao ortólogo.

Em caso de genes codificadores de proteínas com mais de uma isoforma, a pontuação é atribuída por isoforma gênica.

7.2.2 Estrutural

As análises deste módulo são realizadas apenas para o genoma de Estudo. Basicamente o arquivo multifasta contendo as sequências proteicas é analisado em relação ao início e fim da proteína. O início da proteína é representado pela Letra M, correspondendo ao aminoácido Metionina. Já o final da proteína é representado pelo caractere *, que corresponde aos três possíveis códons de parada: TAA, TAG e TGA (substituindo o U por T). Caso seja de conhecimento que o organismo apresente códons de iniciação ou parada alternativos, essa informação deve ser fornecida pelo usuário no arquivo de configuração do programa.

A pontuação é atribuída da seguinte forma:

Pontuação 0,5: Códon de Início
Pontuação 0,5: Códon de Parada
Pontuação -1: (n) Códon de Parada Interno

Outra análise realizada neste módulo é a informação de sobreposição gênica. Os genes codificadores de proteínas podem ser organizados dentro do DNA genômico em arranjos espaciais complexos, podendo sobrepor um determinado segmento de DNA genômico, em que mais de um produto gênico é gerado. Os genes sobrepostos podem ser chamados de aninhados e hospedeiro. O aninhado é aquele gene que se encontra dentro de outro gene, no caso o hospedeiro. Ambos diferem dos transcritos processados alternativamente, genes sobrepostos apresentam sequências muito distintas um dos outros, enquanto que formas alternativas diferem em determinadas regiões (Kumar 2009).

Esse tipo de organização genética é excepcionalmente interessante, pois possui implicações biológicas únicas em relação à evolução, função e regulação do gene. No entanto, gene aninhado completamente em região exônica é bastante raro, com poucos exemplos observados em eucariotos. Devido ao fato de ser um acontecimento não tão comum biologicamente, mas muito comum nos preditores gênicos (Pavesi et al. 2018), a ferramenta aqui proposta apenas reporta os casos de sobreposição dos genes, sem pontuar, deixando para o usuário a opção de rever ou não os genes em sobreposição.

Para detectar sobreposição dos genes, o arquivo GFF3, com as coordenadas dos genes codificadores proteína é utilizado. Este programa não considera como genes aninhados, aqueles oriundos de *splicing* alternativo (isoformas alternativas) e as sobreposições de regiões de UTRs. Também não distingue se a sobreposição ocorre em regiões de introns ou exons, como ilustrado na Figura 16.

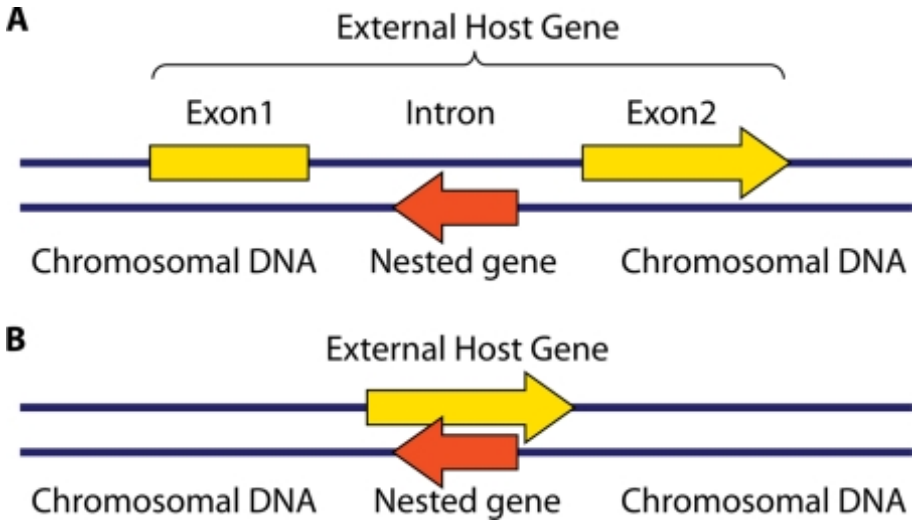

Figura 16: Contexto Cromossômico de um Gene Aninhado. (A) Diagrama de um gene aninhado intrônico. O gene aninhado é representado por uma seta vermelha, enquanto o gene hospedeiro externo é representado por uma seta amarela interrompida por um intron. (B) Diagrama de um gene aninhado não intrônico. O gene aninhado e o gene do hospedeiro são indicados como acima (setas). Fonte:(Kumar 2009).

7.2.3 Transcricional

O único módulo opcional da ferramenta é o transcricional. Por algumas razões: a-) nem sempre os dados de expressão gênica são gerados para estudos de anotação de genomas ou estão disponíveis em bases de dados públicas; b-) mesmo quando gerados os dados de expressão, por meio de RNA-Seq, o resultado é completamente dependente do delineamento experimental. Geralmente estudos de RNA-Seq avaliam condições e tempos diferentes; c-) O estágio de vida e sexo sequenciado são fatores determinantes na expressão gênica.

É válido ressaltar que um gene codificador de proteína que apresente uma cobertura em extensão de 100% da estrutura predita esteja sendo transcrito, e que a estrutura predita corresponda aos dados de expressão. Todavia, tendo em vista o que foi descrito no primeiro parágrafo, a informação de transcritos não é somada na pontuação final dos genes, apenas um relatório é gerado por gene e a informação adicionada ao arquivo final no formato GFF3, em que é indicador por cor se o gene apresenta uma cobertura de 100%. A visualização gráfica para o curador de genomas é muito importante, por isso destaca-se por cor os genes com evidências de transcritos.

Caso o usuário opte por fornecer a informação de transcritos, esta deve ser inserida por meio da ferramenta BedTools. O mapeamento não é realizado aqui por inúmeras razões; entre estas a peculiaridade de cada experimento, tipo de dados, complexidade do genoma dentre outros. O arquivo do mapeamento dos dados gerados contra o organismo de estudo, geralmente em formato BAM, deve ser convertido no arquivo de entrada para a ferramenta. No tutorial da ferramenta está disponível o passo a passo para a conversão. Essa informação será adicionada ao arquivo GFF3 final e evidenciada por cor na *feature* mRNA.

7.2.4 Avaliação da Funcionalidade, Especificidade e Sensibilidade do Programa

A seguir serão apresentadas diferentes métricas de avaliação da funcionalidade da ferramenta desenvolvida. Haja vista a completude dos módulos, a avaliação foi realizada com o resultado oriundo da integração dos módulos: Sintenia de Ortólogos e Estrutural.

Lembrando que a pontuação atribuída ao módulo de Sintenia de Ortólogos vai de 0 a 5 (onde 0 só é atribuído em regiões/scaffolds/contigs os quais são compostos por menos de 10 genes codificadores de proteínas); e a pontuação atribuída ao módulo Estrutural é de 1 ponto positivo ou negativo. Já a pontuação final é decorrente da soma da pontuação dos dois módulos.

O módulo Transcricional por ser opcional não é pontuado, mas a informação de um gene com cobertura de 100% é sinalizada no arquivo final (formato GFF3).

7.2.4.1 Genomas a serem avaliados

Dois genomas de estudo, cujos genomas referências são considerados com alta qualidade de montagem e anotação, foram selecionados para a avaliação. Ambos apresentam características únicas que podem contribuir para uma melhor avaliação:

- **Estudo 1:**

Genoma de Estudo: *Drosophila simulans;*
Genoma de Referência: *Drosophila melanogaster*

- **Estudo 2:**

Genoma de Estudo: *Schistosoma mansoni v5.2*
Genoma de Referência: *Schistosoma mansoni v5.7*

7.2.4.1.1 Da escolha dos genomas do Estudo 1

O cariótipo de *Drosophila melanogaster* é composto por quatro cromossomos: os cromossomos sexuais X e Y (fêmeas XX e macho XY), dois elementos autossômicos maiores: os cromossomos 2 e 3 e o pequeno cromossomo 4 (METZ, 1916; METZ; MOSES, 1923). Em suma, há um total de 10 braços cromossômicos: XL, XR, YL, YS, 2L, 2R, 3L, 3R, 4L e 4R (Deng et al. 2007). A razão para os braços dos cromossomos: X, 2, 3 e 4 serem designados à esquerda (L) e à direita (R), enquanto o Y é designado longo (L) e curto (S), perdeu-se no tempo (Kaufman 2017).

Desde o primeiro sequenciamento, montagem e anotação do genoma, promovida pelo consórcio do Projeto Genoma de *Drosophila* Berkeley (Myers et al. 2000) e Celera Genomics (Adams et al. 2000), o genoma nunca parou de ser melhorado (em termos de sequenciamento, montagem e anotação gênica no geral). Dados de RNA-Seq, melhoria das plataformas de sequenciamento e um grande investimento em curadores do genoma foram e são os principais responsáveis pelo genoma de alta qualidade. Hoje, esse genoma é considerado um dos melhores de metazoários alcançados na atualidade (Kaufman 2017; Shah, Cao, and Ellison 2019).

D. simulans apresenta a mesma estrutura cromossômica e genômica de *D. melanogaster*, e também vem sendo constantemente melhorado em termos de sequenciamento, montagem e anotação (Clark et al. 2007; Hu et al. 2013; McManus et al. 2014). Duas novas

versões do genoma estão em processo de submissão e publicação (http://genomics.princeton.edu/AndolfattoLab/w501_genome.html,https://www.biorxiv.org/content/biorxiv/early/2018/09/24/425710.full.pdf) no entanto, trabalharemos aqui com a versão atualmente disponível.

Ambas, *D. melanogaster* e *D. simunlans* divergiram há aproximadamente 2.5 milhões de anos, portanto são consideradas espécies intimamente relacionadas (Rogers et al. 2014). A Figura 17 ilustra uma árvore filogenética na qual são apresentadas algumas espécies do gênero *Droshopilha spp*. com o tempo de divergência evolutiva estimado.

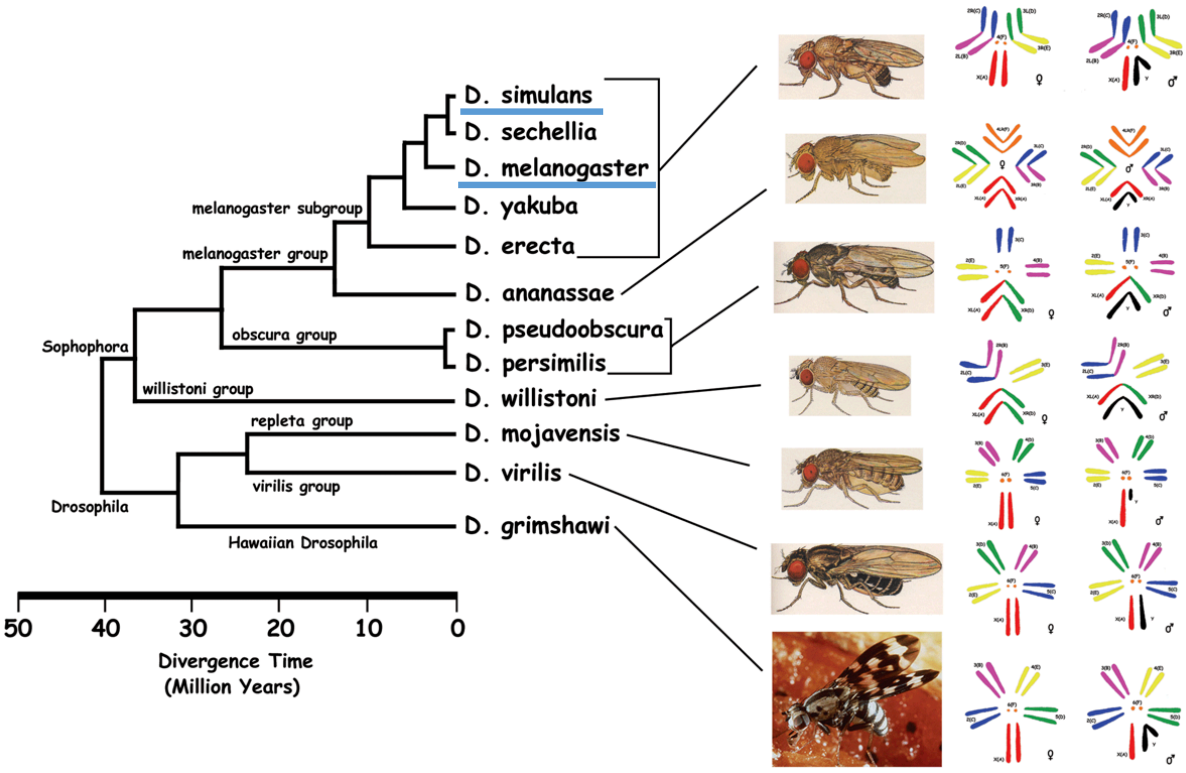

Figura 17: Árvore Filogenética e Tempo de Divergência entre as Moscas da Fruta. A parte à direita da figura mostra uma representação do cariótipo de todas as espécies representadas na árvore. Estão destacadas em azul, *D. melanogaster*, *D. simulans*, alvos de estudo do trabalho. Fonte: http://species.flybase.net/.

As Tabelas 11 e 12 apresentam as informações dos genes codificadores de proteínas distribuídos ao longo dos cromossomos de *D. melanogaster* e *D. simulans*, respectivamente.

Tabela 11: *Drosophila melanogaster* BDGP6.22 (GCA_000001215.4)

Type	Name	RefSeq	Size (Mb)	Protein	Gene	Pseudogene
Chr	X	NC_004354.4	23.54	5,390	2,675	68
Chr	2L	NT_033779.5	23.51	5,694	3,501	47
Chr	2R	NT_033778.4	25.29	6,051	3,628	46
Chr	3L	NT_037436.4	28.11	5,880	3,466	34
Chr	4	NC_004353.4	1.35	295	111	6
Chr	3R	NT_033777.3	32.08	7,205	4,202	36
Chr	Y*	NC_024512.1	3.67	25	113	62
	MT	NC_024511.2	0.02	13	37	-
Un	-	.	6.16	6	5	-

Fonte: Adaptado de : https://www.ncbi.nlm.nih.gov/genome/?term=txid7227[orgn]

Tabela 12: *Drosophila simulans* W501 (ASM75419v2)

Type	Name	RefSeq	Size (Mb)	Protein	Gene	Pseudogene
Chr	2L	NT_479533.1	23.54	4,305	2,837	29
Chr	2R	NT_479534.1	21.54	5,044	3,063	29
Chr	3L	NT_479535.1	24.15	4,821	3,034	29
Chr	3R	NT_479536.1	27.16	5,765	3,670	33
Chr	4	NC_029796.1	1.03	291	96	4
Chr	X	NC_029795.1	20.83	3,635	2,324	37
	MT	NC_005781.1	0.01	13	13	-
Un	-	.	6.69	258	205	4

Fonte: Adaptado de: https://www.ncbi.nlm.nih.gov/genome?term=txid7240%5Borgn%5D

Deste modo, diante dos dados atualmente disponíveis em bases de dados públicas, o genoma/anotação de *D. melanogaster* está em melhor qualidade do que o genoma/anotação de *D. simulans*, à vista disso, *D. melanogaster* será a espécie referência deste estudo.

7.2.4.1.2 Da Escolha dos Genomas do Estudo 2:

O parasito *Schistosoma mansoni* foi alvo de estudo do primeiro capítulo. A nova montagem do parasito (v7.1) é considerada de alta qualidade, bem como a anotação. Mesmo com a correção manual e a integração de ISO-seq na versão antiga (v5.2), a anotação por muitas vezes continuou errada, devido principalmente aos erros da montagem do genoma.

Assim sendo, o genoma da v5.2 foi escolhido como o genoma de Estudo, e o genoma v7.1 o genoma de Referência. Deste modo a ferramenta será testada em uma montagem considerada fragmentada.

7.2.4.2 Métricas de Avaliação do Desempenho da Ferramenta

Os genes com possíveis erros detectados recebem pontuação entre 1 e 2, enquanto aos genes confiáveis é atribuída uma pontuação mais alta 4 a 6. Para avaliação, os genes que recebem pontuação acima de 3 são considerados positivos.

O desempenho da ferramenta foi avaliado, entre alternativas, por meio de algumas métricas. Uma sumarização dos conceitos e métricas de avaliação utilizados estão descritos abaixo:

- Teste de Sanidade (teste de fumaça):

São testes simples, rápidos e pouco específicos, responsáveis por encontrar erros grandes de algoritmos. Os testes verificam se os resultados esperados não estão fora de um intervalo esperado, ou então verificam se os resultados obtidos atendem a certas propriedades relacionadas ao algoritmo. Neste estudo seria o retorno do proteoma analisado contra ele mesmo.

- Verdadeiro-positivos (VP):

É um item corretamente predito como sendo da classe positiva. No estudo, foram chamados genes verdadeiros positivos, aqueles genes que apresentaram evidências com suporte em base de dados curadas e ou evidência completa de transcrição.

- Falso-positivos (FP):

É um item incorretamente predito como sendo da classe positiva. Quando a ferramenta não for capaz de identificar um gene da classe verdadeiro positivo, este será designado como falso negativo.

- Verdadeiro-negativos (VN):

É um item corretamente predito como sendo da classe negativa. No estudo, foram adicionados genes sabidamente errados, quando estes genes forem apontados pela ferramenta como falsos genes serão contabilizados como sendo da classe de verdadeiro negativos.

- Falso-negativos (FN):

É um item incorretamente predito como sendo da classe negativa. Quando os genes falsos adicionados no estudo não forem detectados quando falsos, estes serão atribuídos à classe dos falsos positivos.

- Sensibilidade (SN):

É a capacidade do teste de identificar corretamente os verdadeiros positivos. Logo, a sensibilidade neste estudo indicará a capacidade de atribuir uma pontuação alta aos genes chamados de verdadeiros.

- Especificidade:

É a capacidade do teste de identificar os verdadeiros negativos. A medida visa apontar a capacidade da ferramenta em lidar com a atribuição de baixos valores aos genes errados da anotação

7.2.4.2.1 Teste de Sanidade

O genoma de *D. melanogaster* foi fornecido como espécie de Estudo e Referência no programa. O mesmo foi realizado com o genoma do parasito *S. mansoni* v7.1. É esperado que ambos apresentem pontuação total, haja vista que a sintenia do genoma contra ele mesmo deva ser total.

7.2.4.2.2 Sensibilidade e Especificidade

- Obtenção dos genes verdadeiros positivos:

Para a anotação de *D. simulans,* a base de dados curada do UniProt: Swis-Prot foi consultada. Todos os genes revisados manualmente pertencentes à montagem ASM75419v2 do genoma foram revisados pelos módulos Estrutural e Transcricional, a fim de confirmar a estrutura e expressão dos genes.

No estudo transcricional, dados de RNA-Seq referentes a *D. simulans* foram obtidos por meio de consulta na base de dados do SRA - *Sequencing Read Archive* (Leinonen, Sugawara, and Shumway 2011). Filtros foram realizados nos dados disponíveis, seguindo os seguintes critérios: a-) Apenas dados referentes a mesma cepa da montagem do genoma w501; b-) Plataforma Illumina (não havia dados de ISO-Seq); c-) RNA-Seq total, ou de diferentes partes do organismo onde a maior representação possível de expressão gênica estivesse presente; d-) foram eliminados estudos temporais.

A Tabela 13 apresenta os dados brutos selecionados para o mapeamento das *leituras* contra o genoma. Todos os dados são pertencentes à cepa w501.

Tabela 13: Sequenciamento de RNA *Drosophila simulans* (w501)

BioSample	Run	Sample_Name	sex	tissue	Instrument
SAMN01940564	SRR1511608	w501 ovary	female	NA	Illumina HiSeq 2500
SAMN01940564	SRR1511609	w501 ovary	female	NA	Illumina HiSeq 2500
SAMN02009242	SRR1511610	w501 headless carcass	female	NA	Illumina HiSeq 2500
SAMN02009242	SRR1511611	w501 headless carcass	female	NA	Illumina HiSeq 2500
SAMN02927696	SRR1520537	Drosophila simulans w501 testes	male	virgin testes	Illumina HiSeq 2500
SAMN02927697	SRR1520538	Drosophila simulans w501 male carcass	male	virgin male carcass	Illumina HiSeq 2500
SAMN02927696	SRR1548740	Drosophila simulans w501 testes	male	virgin testes	Illumina HiSeq 2500
SAMN02927696	SRR1548741	Drosophila simulans w501 testes	male	virgin testes	Illumina HiSeq 2500
SAMN02927697	SRR1548743	Drosophila simulans w501 male carcass	male	virgin male carcass	Illumina HiSeq 2500
SAMN02927697	SRR1548805	Drosophila simulans w501 male carcass	male	virgin male carcass	Illumina HiSeq 2500

Filtro:https://www.ncbi.nlm.nih.gov/Traces/study/?WebEnv=NCID_1_63502456_130.14.18.48_5555_1558281126_1818319377_0MetA0_S_HStore&query_key=2

O mapeamento foi realizado por meio da ferramenta bowtie2 (Langmead et al. 2019), com os seguintes parâmetros:

```
hisat2-align-s --wrapper basic-0 -x ref_drosophila_bowtie/dro_indx -S Dro_simu.bam -p
10 --qc-filter -1 D Simulans W501/SRR15 Fow.fastq -2  Simulans W501/SRR15 Rev.fastq"
```

Apesar da chamada do programa utilizar o hisat2, o mesmo não foi utilizado por não ser objetivo de o estudo quantificar o RNA-Seq, mas sim obter a informação de cobertura em extensão em relação ao proteoma predito.

Com o arquivo de mapeamento gerado, no formato BAM, a ferramenta BedTools foi utilizada para inserir a informação de mapeamento no arquivo GFF3, com os comandos descritos abaixo, onde `-s` significa preservar a orientação da fita (strand):

```
bedtools coverage -s -a file.gff3 -b file.bam > newfile.gff3
```

O teste estrutural e transcricional do parasito *Schistosoma mansoni* são proveniente dos estudos do Capítulo 1, Tópico 5.2.1. No entanto alguns filtros foram empregados. Só foram chamados de genes verdadeiros positivos aqueles que atendessem aos seguintes critérios: cobertura completa de ISO-Seq; presença do códon de iniciação e do códon de parada; ausência de códon de parada no meio da sequência.

- Obtenção dos genes representando os verdadeiros negativos,

No estudo de *D. simulans*, estes genes foram provenientes de um genoma externo: *D. ananassae*, o qual pertence ao grupo de *D. melanogaster*, como representado na Figura 17. O objetivo da adição dos genes de um organismo externo foi medir a especificidade do programa.

Como a avaliação de sintenia de ortólogos é dependente das posições contínuas em um genoma, foi necessário a inserção de um *scaffold* completo selecionado aleatoriamente no genoma.

Em *S. mansoni* v5.2, os genes chamados de negativos, apresentavam problemas na estrutura, como por exemplo: genes preditos como sendo diferentes genes (dois ou mais), mas após a curadoria manual utilizando os dados de ISO-Seq como suporte foram preditos como sendo apenas um gene, esquema explicativo ilustrado na Figura 18. Outros exemplos: genes com exons incompletos, genes com códon de parada interno. Todos os genes de teste foram adicionados antes da correção manual.

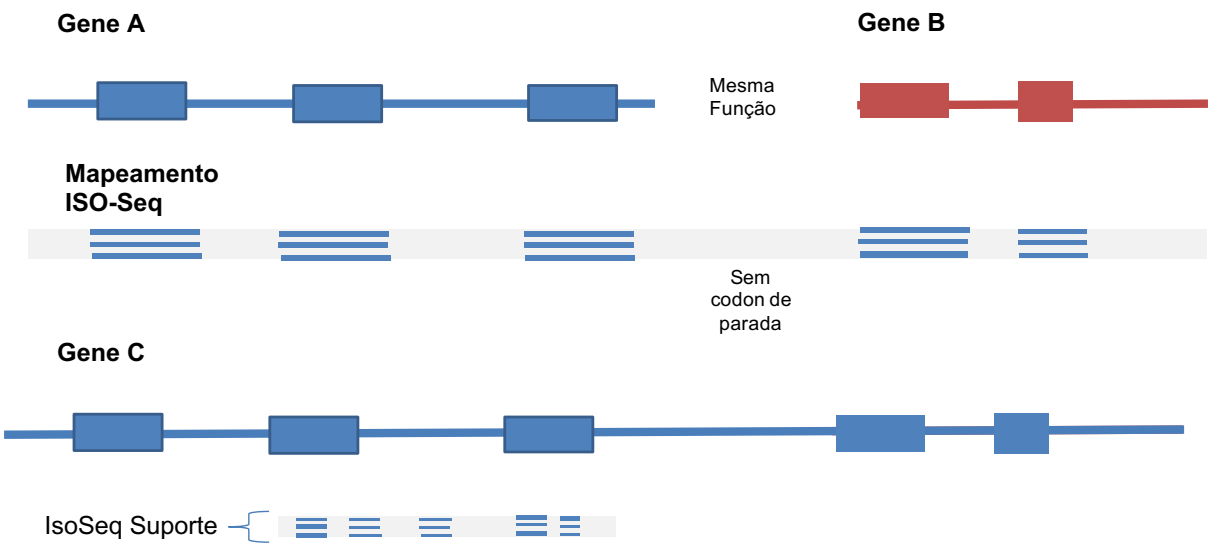

Figura 18: Genes Erroneamente Preditos Na anotação automática do genoma: os Genes A em azul e B em vermelho, foram preditos como genes diferentes. Com o novo dado de suporte do ISO-Seq foi possível constatar que ambos se tratam de apenas um gene: GeneC. Logo, os Genes A e B foram considerados falsos genes.

- Cálculo da Sensibilidade por Pontuação atribuída pela ferramenta. Lembrando que a pontuação é realizada de 1-6, descrita no tópico 7.2.1.

$$Sensibilidade = \frac{Verdadeiro\ Positivo\ (VP)}{Verdadeiro\ Positivo\ (VP)\ /\ Falso\ Negativo\ (FN)}$$

- Cálculo da Especificidade por Pontuação atribuída pela ferramenta. Lembrando que a pontuação é realizada de 1-6, descrita no tópico 7.2.1.

$$Especificidade = \frac{Verdadeiro\ Negativo\ (VN)}{Verdadeiro\ Negativo\ (VN) / Falso\ Positivo\ (FP)}$$

7.3 Resultados

A seguir serão descritos os resultados referentes aos testes de sanidade e sensibilidade e especificidade.

7.3.1 Resultado do Teste de Sanidade

Os Testes de Sanidade foram realizados utilizando o genoma de *Drosophila melanogaster* e com o genoma do parasito multicelular *S. mansoni* v7.1. Ambos os testes foram executados utilizando a sequência do genoma contra a própria sequência. Como resultado, todas as pontuações atribuídas ao módulo Sintenia de Ortólogos foram iguais, indicando assim a funcionalidade da aplicação da sintenia. Resultado no material Suplementar 5 (Arquivo do tipo GFF3).

7.3.2 Resultados Sensibilidade e Especificidade

Os resultados dos testes serão exibidos após os resultados da obtenção dos verdadeiros positivos e verdadeiros negativos para os dois casos de estudo.

- Verdadeiro Positivos e Verdadeiros Negativos, *D. simulans*:

Da avaliação estrutural: não foram encontrados genes com sobreposição no genoma de *D. simulans*. Todos os genes codificadores de proteínas reportaram erro nas 3 últimas bases, não sendo possível identificar os códons de parada, inclusive nos genes da base de dados do Swiss-Prot. A Figura 19 mostra o exemplo do gene: Fbtr0225921 em que se evidencia o erro. É possível visualizar o códon de parada TAG logo em seguida ao final da sequência. Este é um erro comum em anotação de genomas e pode ser corrigido automaticamente, como por exemplo utilizando a ferramenta Artemis ilustrada na Figura 19.

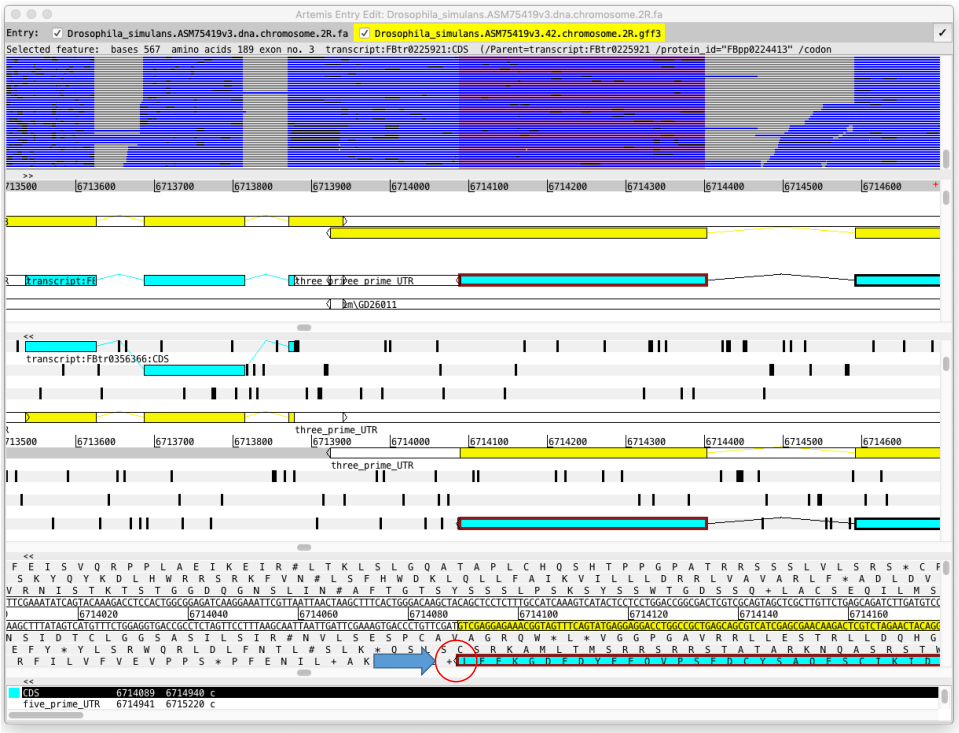

Figura 19: Erros no Códon de Parada. O gene Fbtr0225921 é ilustrado como exemplo na visualização do programa Artemis. Em azul escuro é possível visualizar a cobertura por RNA-Seq, em azul claro os genes codificadores de proteínas preditos (exons). A seta em azul e o círculo em vermelho evidenciam o erro cometido em toda a anotação do genoma de *D. simulans*, o qual não inclui o códon de parada na sequência do gene codificador de proteína.

Transcricional: O mapeamento das leituras de RNA-Seq contra o genoma de *D. simulans* possibilitou contabilizar a cobertura em extensão dos genes preditos. Dos 92 genes da tabela da base de dados do SwissProt, 80 genes apresentaram cobertura de 100%. A correção das últimas 3 bases dos 80 genes foi selecionada para controle positivo. Quadro 4 ilustra um resumo dos genes selecionados para estudo.

Os genes chamados de verdadeiros negativos foram oriundos do genoma externo de *D. ananassae*. O *scaffold*, 12929 foi completamente inserido nas análises, totalizando 326 genes codificadores de proteínas.

Quadro 4: Resumo dos Genes Verdadeiros Positivos e Verdadeiros Negativos em *D. simulans*

VP = 80 genes codificadores de proteínas manualmente anotadas e revisadas adquiridas na base de dados curada do UniProt: Swiss-Prot.
VN = 326 genes codificadores de proteínas do Scaffold_12929 da *Drosophila ananassae*, faz parte do grupo da D. *melanogaster*

- Verdadeiro Positivos e Verdadeiros Negativos, *S. mansoni* v5.2:

Para o parasito *S. mansoni v5.2,* a obtenção dos genes verdadeiros positivos e negativos foi proveniente do estudo descrito no Capítulo 1. Após os filtros realizados, apenas 265 genes passaram no teste de verdadeiros positivos e um total de 311 genes codificadores de proteínas no teste dos verdadeiros negativos. O quadro 5 apresenta o resumo dos verdadeiros positivos e verdadeiros negativos selecionados para os cálculos de sensibilidade e especificidade da ferramenta.

Quadro 5: Resumo dos Genes Verdadeiros Positivos e Verdadeiros Negativos em *S. mansoni*

VP = 265 genes codificadores de proteínas manualmente anotadas e revisadas e validas com o suporte de dados do ISO-Seq
VN = 311 genes codificadores de proteínas revisados manualmente sem suporte de RNA-Seq (Illumina e/ou ISO-Seq)

- Resultado do Cálculo da Sensibilidade e Especificidade:

A Tabela 14 apresenta os resultados referentes à especificidade e à sensibilidade da ferramenta para os dois genomas de estudo de *D. simulans* e *S. manosni* v5.2.

Tabela 14: Sensibilidade e Especificidade

	D. simulans		*S. mansoni v5.2*	
	Sensibilidade	Especificidade	Sensibilidade	Especificidade
Pontuação 1	1	0	1	0
Pontuação 2	1	0	1	0
Pontuação 3	1	1	0.5875	1
Pontuação 4	0.625	1	0.5875	1
Pontuação 5	0.625	1	0.5875	1
Pontuação 6	0.5875	1	0.5875	1

7.4 Da aplicação da Ferramenta Nos Genomas Estudados no Capítulo 1

Após os testes de funcionalidade do programa, o mesmo foi aplicado aos genomas de estudo do Capítulo 1.

O caso de estudo aqui reportado, é a avaliação da qualidade da anotação de *Plasmodium knowlesi*, em que o genoma de Estudo é referente à versão antiga do genoma, chamada de Pknowlesi_H e o genoma referência é o genoma montado e anotado neste estudo, Pknowlesi_Pk1a.

O módulo transcricional não foi aplicado neste estudo de caso, por não terem sido gerados dados de RNA-seq total do parasito em questão.

7.4.1 Resultados Módulo de Sintenia por Ortólogos

Pknowlesi_H apresenta regiões (blocos ou genes) homólogos em diferentes posições quando comparado com Pknowlesi_Pk1a. A Figura 20 representa a Sintenia de Ortólogos entre os genomas, e destacam-se 3 pontos em vermelho:

a-) O primeiro ponto está localizado no Cromossomo 12 de Pknowlesi_H e Cromossomo 11 de Pknowlesi_Pk1A. A nomenclatura dos cromossomos em *Plasmodium knowlesi* foram alteradas devido aos erros de montagem na versão anterior (Pknowlesi_H), em que determinados *scaffolds* foram movidos para os devidos cromossomos (LAPP, S. A. ASSIS *et al.*, 2018).

O cromossomo 12 (PKNH_12) de Pknowlesi_H corresponde ao cromossomo 11 e cromossomo 12 da referência. Com a mudança de lugar dos *scaffolds* que compõem o cromossomo 12 de Pknowlesi_H, os genes foram deslocados para região do outro cromossomo da referência. Esse deslocamento de regiões ocasionou a quebra de Sintenia de Ortólogos, levando todos os genes presentes no cromossomo 12 de Pknowlesi_H, após o deslocamento do *scaffold*, a terem uma pontuação indicando quebra de Sintenia entre os genomas comparados.

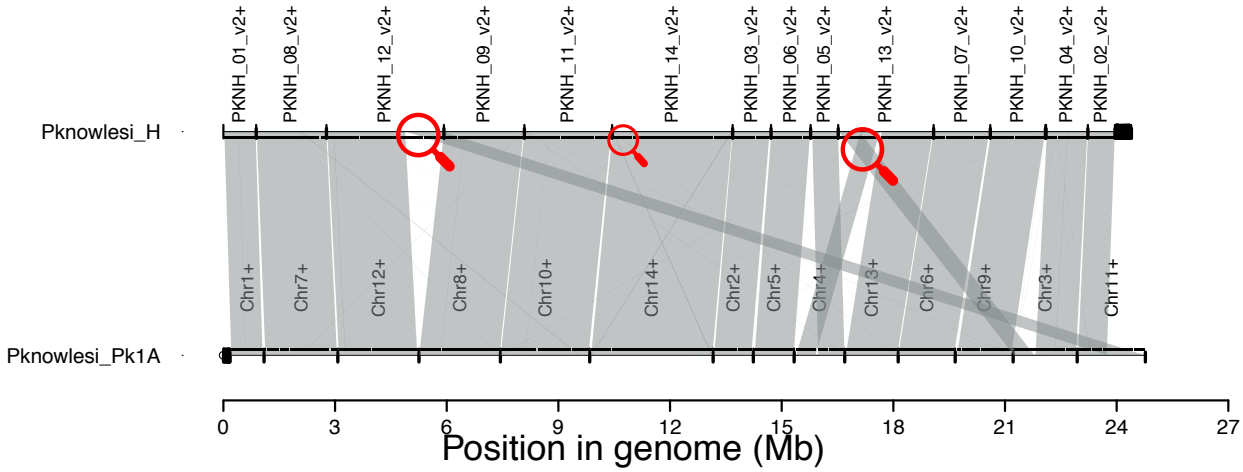

Figura 20: Sintenia de Ortólogos Entre as Diferentes Montagens de *P. knowlesi*. As três lupas em vermelho indicam pontos de quebra de sintenia entre os genomas comparados. O primeiro ponto, da esquerda para a direita, bem como o terceiro ponto, representam erros na montagem do genoma de Pknowlesi_H. Já o segundo ponto (do meio) representa erro na anotação de genes pertencentes a famílias multigênicas.

b-) O segundo ponto destacado na Figura 20, com quebra de Sintenia, corresponde a erros de predição na estrutura dos genes de início e fim do cromossomo em Pknowlesi_H. A Figura 24, apresenta a observação do início do cromossomo 14 de ambos os parasitos, neste caso, o cromossomo 14 é correspondente em ambos os genomas.

No ponto A da Figura 21 está representado o início do chr14 de Pknowlesi_H, onde está destacado em vermelho. O primeiro gene do cromossomo está representado em branco e os 3 exons que o compõem estão representados em amarelo e as CDS em azul (CDS e exon no arquivo GFF3 só exibem diferenças quando a região de UTR está predita e é acrescentada ao exon). O boxe menor indica o nome do gene: SICAvar do Tipo I, fragmentado. O ponto B da Figura 21 apresenta também o início do chr14 em Pknowlesi_Pk1A, onde o gene SICAvar do Tipo I está completo, contendo os 3 exons preditos no genoma de Pknowlesi_H. O gene apenas foi predito na fita (*strand*) errada, o que impossibilitou a continuidade da predição. O programa ASSIS aponta para a quebra de sintenia neste local.

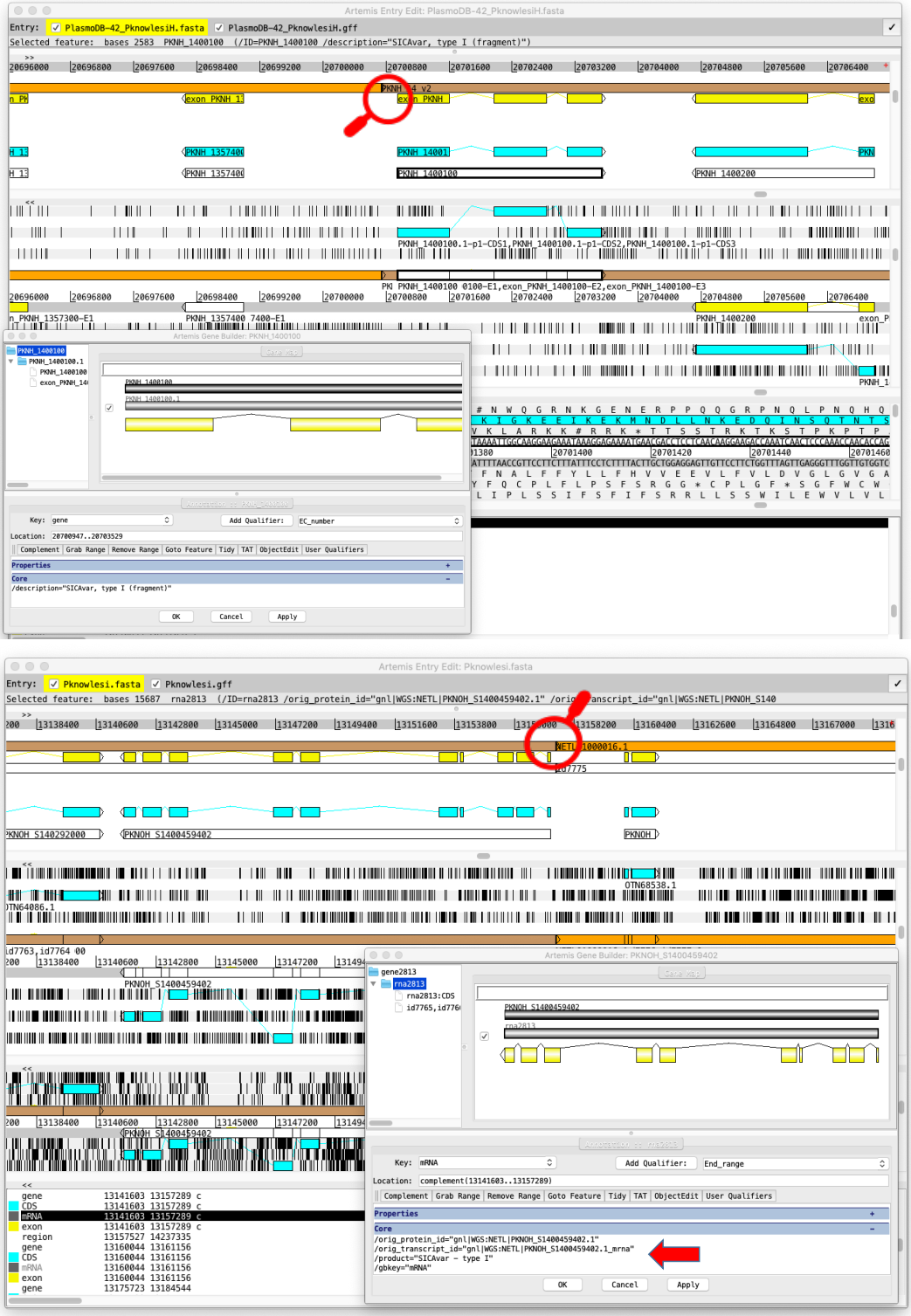

Figura 21: Quebra de Sintenia de Ortólogos entre os genomas de *Plasmodium knowlesi*. O ponto A ilustra o início do chr14 onde o gene SICAvar do tipo I está fragmentado e o ponto B, ilustra a mesma região, predita corretamente em Pknowlesi_Pk1A.

c-) O Terceiro ponto em destaque da Figura 20, também indica para um erro de montagem em Pknowlesi_H. O chr5 de Pknowlesi_H, corresponde agora a dois diferentes cromossomos: chr4 e chr9 em Pknowlesi_Pk1A. Assim como no primeiro caso, todos os genes

deslocados foram penalizados na pontuação de Sintenia de Ortólogos, indicando uma quebra na região completa, mesmo quando o gene é real e predito corretamente, devido a este fator os outros módulos devem ser levados em consideração para avaliar individualmente os genes.

7.4.2 Resultado da Integração dos Módulos de Sintenia de Ortólogos e Estrutural

Dos 5.326 genes codificadores de proteínas presentes em Pknowlesi_H, 5.287 genes foram melhorados com a integração dos módulos. Destes, 4.116 genes apresentaram pontuação máxima. Apenas 41 genes permaneceram com o mesmo valor da pontuação anterior à integração, 46 genes tiveram a pontuação diminuída devido à presença de códons de parada interno. Apenas 4 genes receberam pontuação 1.

A Figura 22 ilustra um exemplo do gene PKNH_0407500, onde a pontuação foi aumentada de 2 para 3, após a integração dos módulos. O gene PKNH_0407500 também apresentou sobreposição com o gene PKNH_0407600, representado no lado A da Figura 22. O lado B, evidencia o gene correspondente em Pknowlesi_Pk1A: PKNOH_S055400800, no entanto não há a predição de gene correspondente a PKNH_0407600. Mesmo com tentativa de inserção manual de um novo gene, não foi possível encontrar uma janela aberta de leitura na vizinhança. A pontuação atribuída a PKNH_0407600 foi o valor 2.

Figura 22. Gene Codificador de Proteína em Sobreposição ao Exon de Outro Gene. O lado A evidencia o gene PKNH_0407500 em sobreposição com o exon do gene PKNH_0407600 no genoma anotado de Pknowlesi_H. O lado B evidencia apenas um gene correspondente em Pknowlesi_Pk1A: PKNOH_S055400800.

Os genes PKNH_0500500.1, PKNH_0500600.1, PKNH_0500700.1 foram pontuados com o valor 1, o mais baixo nível de evidência, ambos foram investigados manualmente e estão representados na Figura 23. Todos estavam incompletos, os mesmos foram identificados por

meio da Ferramenta ASSIS, mas também estavam anotados como fragmentos de gene. Na versão nova do genoma de Pknowlesi_Pk1A, os três genes foram fundidos em apenas um novo gene: SICAvar do tipo I.

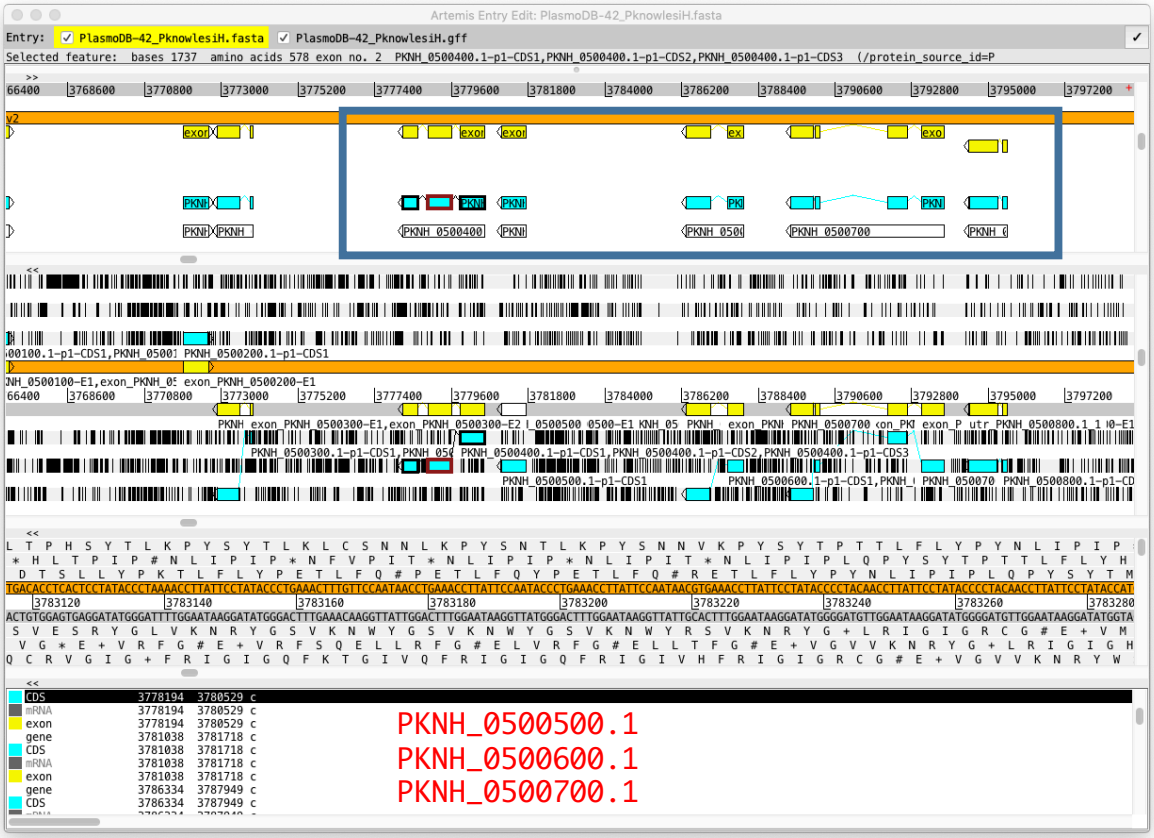

Figura 23: Genes com Pontuação 1. Os três genes: PKNH_0500500.1, PKNH_0500600.1, PKNH_0500700.1 receberam pontuação 1, todos estavam incompletos. Na versão nova do genoma: Pknowlesi_Pk1A os três genes foram fundidos em apenas um novo gene: SICAvar do tipo I.

8 DISCUSSÃO

A comunidade científica tem presenciado nos últimos anos uma mudança que rege o desenvolvimento e o avanço científico. Os desafios e demandas criadas pelo grande ritmo de geração e disponibilização de dados, em particular, dos dados de sequenciamento genômico, tem crescido a cada ano. O grande volume e dinamicidade de dados biológicos vem gerando desafios computacionais e biológicos, principalmente no que diz respeito à extração de conhecimento a partir dessas fontes de informações.

Consideráveis trabalhos de sequenciamento completo do genoma, bem como montagem e anotação, foram e vem sendo desenvolvidos. Cada um utilizando diferentes tipos de tecnologias e abordagens, mas, todos compartilhando de um objetivo em comum: obter a informação genética do organismo de estudo.

A qualidade dos dados e dos resultados gerados nem sempre é alcançada em um nível o qual podemos chamar de bom. Contudo, independentemente da qualidade alcançada, os trabalhos são publicados e toda a informação biológica passa a estar disponível em base de dados públicas. Um sequenciamento incompleto do DNA de determinado organismo, leva a uma montagem incompleta do genoma, que por sua vez, gera uma anotação comprometida.

Cada vez mais as anotações decorrem de estudos comparativos. A genômica comparativa vem permitindo reconhecer as semelhanças e diferenças nos organismos, todavia, esta também exige cautela devido à possibilidade da propagação de erros. Isso ocorre quando um dado errôneo ou incompleto é depositado em base de dados pública e por meio da genômica comparativa o erro é propagado. Um modelo gênico predito errado, depositado em bancos de dados públicos, acaba servindo como base para treinar novos modelos gênicos que serão aplicados ao estudo de novos organismos sequenciados.

Erros de anotação, em grande parte, são provenientes da anotação automática de genomas (por erros no treinamento de modelo, dentre outros), logo, a ideia da curadoria manual parece surgir como uma solução do problema. No entanto, o processo de curadoria manual é custoso para quem o faz, demanda tempo e equipe de trabalho e está susceptível a erros. Por vezes, o processo manual consiste em comparação com genes já depositados em bancos públicos, podendo também estar vulnerável a propagação de erros. Por isto, avaliar a qualidade de todos genes de uma anotação de maneira automática é de extrema importância na atualidade.

Nesse sentido, no presente trabalho foi apresentado uma nova ferramenta para análise da qualidade da anotação dos genes codificadores de proteínas em organismos eucariotos, baseado na integração dos dados.

O trabalho enfrentou inúmeras adversidades e desafios a cada genoma estudado. Começando pela escolha de um organismo modelo o qual esperava-se alcançar uma boa qualidade da anotação dos genes codificadores de proteínas, onde seria possível obter uma referência de genes verdadeiros positivos. No entanto, genomas considerados referências, apresentaram erros basais, como erros na estrutura do gene, por exemplo, o genoma de *D. simulans*, onde faltavam o códon de terminação de todos os genes codificadores de proteínas, provavelmente gerado por descuido dos curadores. Numerosos erros foram encontrados, até mesmo erros de formatação de arquivo nos genomas de referência de *Mus musculus*, *Caenorhabditis elegans*, dados não apresentados.

Diante da dificuldade em encontrar um genoma referência o qual fosse possível contabilizar os genes verdadeiros positivos, dois genomas foram selecionados para o estudo de avaliação da ferramenta: *D. simulans* e *S. mansoni*.

Além dos erros detectados no módulo estrutural onde os códons de parada estavam ausentes, outro problema foi diagnosticado em *D. simulans*. O gene FBtr0354604, considerado no banco de dados do Swiss-Prot como um gene curado manualmente e revisado, apresenta 8 isoformas alternativas. A ferramenta proposta neste estudo, trabalha individualmente com cada uma das isoformas de um gene, logo, para 7 isoformas foi possível obter a pontuação máxima esperada, mas, para a isoforma FBtr0219262, a pontuação atribuída foi 2, sinalizando um problema na isoforma do gene. Por meio da curadoria manual, foi possível levantar a hipótese de que essa isoforma não pertence ao gene a qual é sinalizada.

A Figura 24, apresenta o caso do FBtr035402 destacando a chamada isoforma alternativa: FBtr0219262. Utilizando os dados de expressão de RNA-Seq fita específica, é possível visualizar que FBtr0219262 está sendo transcrito na direção 3'-5', enquanto que todas as outras isoformas do gene FBtr035402 estão sendo transcritos na fita positiva. Buscando por informações na literatura foi possível encontrar o trabalho de Henikoff, 1986, onde foi reportado pela primeira vez a validação de genes em sobreposição em *D. melanogaster* (Henikoff et al. 1986). Por comparação das sequências e tamanho, acredita-se tratar de um gene em sobreposição com outro, onde o gene aninhado seria FBtr0219262 e o gene hospedeiro FBtr035402.

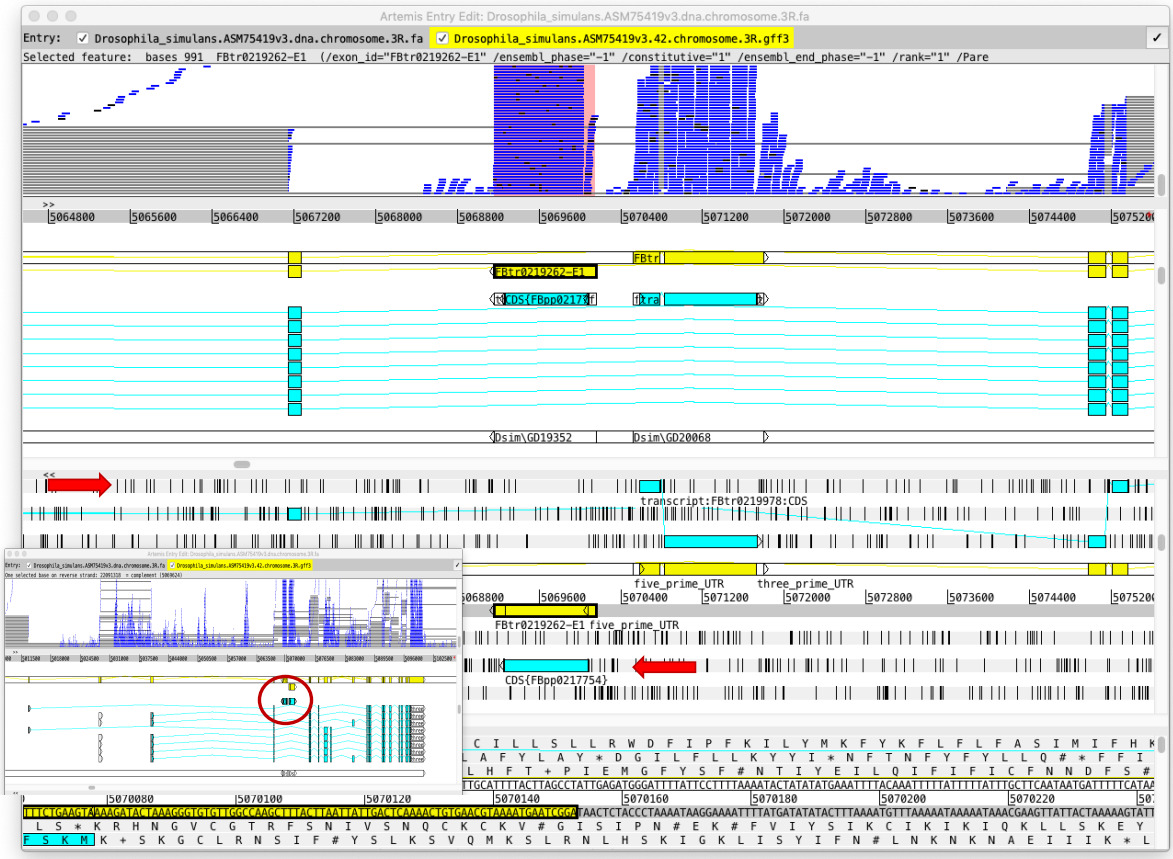

Figura 24: Gene em possível Sobreposição. A isoforma FBtr021962 atribuída ao gene FBtr035402, parece ser na verdade outro gene em sobreposição com este. A parte superior da imagem evidencia a Isoforma FBtr21196, em azul está a informação dos transcritos, onde o perfil de expressão difere das demais isoformas. O boxe inferior mostra a visão geral do FBtr035402 com as demais isoformas. As setas em vermelho indicam o sentido da transcrição.

Erros também foram encontrados em *S. mansoni*, mesmo em genes com cobertura completa do ISO-Seq. O tipo de erro foi provavelmente referente a plataforma de sequenciamento para a montagem (PacBio), com inserção/deleção de bases que levaram a mutações do tipo *frameshift*. Com a correção dos genes de interesse, foi possível estabelecer um conjunto teste para avaliação da ferramenta desenvolvida.

O primeiro módulo da ferramenta proposto neste trabalho, denominado Sintenia de Ortólogos, desenvolvido a partir da premissa que genes ortólogos compartilham genes ortólogos vizinhos, foi pensado para identificar falhas na anotação dos genes. Por meio da quebra de sintenia, seria possível identificar genes faltantes, ou genes parciais no genoma. O módulo foi capaz de identificar a quebra de sintenia, no entanto, a pontuação atribuída foi completamente dependente da continuidade da sintenia. Onde mais de dois genes consecutivos apresentaram divergências, a penalidade da pontuação atribuída foi muito alta, indicando assim

a região onde um possível problema foi identificado. Dessa forma, o módulo foi capaz de fornecer uma visão geral do genoma em estudo, apontando para regiões onde a quebra de sintenia de ortólogos aconteceu. A Figura 20 apresentada nos resultados, onde é exibida a comparação entre as montagens *de P. knowlesi,* configura um exemplo da penalidade da região, indicando a quebra da sintenia e possíveis erros.

Todavia o módulo de Sintenia de Ortólogos quando trabalhado sozinho, gera uma alta taxa de falsos negativos, isto porque, como descrito no parágrafo acima, a penalidade é realizada para a região (*contig*, *scaffold*, *chr*), a partir do momento da quebra da sintenia.

O segundo módulo proposto: Estrutural; é capaz de avaliar a composição da sequência dos genes codificadores de proteínas. Anotadores gênicos, como por exemplo, a plataforma de anotação do Maker2, tem como comando padrão, permitir a predição de genes incompletos, ou seja, genes sem códon de início ou de parada, bem como genes parciais. Muito embora para submeter sequências a base de dados do GeneBank seja necessário passar por um processo de avaliação por meio das ferramentas disponibilizadas pelo próprio banco de dados, os genes considerados incompletos são adicionados com o consentimento do autor que está submetendo. Por muitas vezes, quem submete o genoma anotado só se depara com erro pela primeira vez no momento de submissão a base de dados.

Ferramentas como GAG (Genome Annotation Generator) https://genomeannotation.github.io/GAG/, possibilitam que o usuário visualize os erros da estrutura do gene antes da submissão, no entanto, ferramentas desse tipo são extremamente preocupantes. Isso porque, estas permitem a manipulação dos dados, onde exons/introns inteiros podem ser removidos para atenderem aos requisitos no banco ao qual o usuário está submetendo o genoma. A ferramenta proposta neste estudo, por sua vez, apenas reporta ao usuário o tipo de problema encontrado, mas a reparação do possível erro é de responsabilidade do curador do genoma.

Ao módulo Estrutural a pontuação máxima obtida é de apenas um, seja positivo ou negativo. O resultado é integrado ao módulo anterior, Sintenia de Ortólogos. Como ilustrado no estudo de caso de *P. knowlesi*, a integração da informação estrutural ajuda a melhorar a pontuação individual por gene. Na Figura 21, do estudo de caso de Pkowlesi_H, onde a região recebeu uma pontuação baixa na sintenia indicando problema na região (gene vizinho estava errado), a pontuação estrutural integrada, permitiu indicar um gene predito corretamente. Assim sendo, a completude dos módulos se mostrou de grande importância para a obtenção de um resultado mais próximo ao real.

Por meio da utilização das métricas de sensibilidade e especificidade, foi possível medir a capacidade da ferramenta em contabilizar os verdadeiros positivos e verdadeiros negativos, por pontuação atribuída aos genomas testes.

A especificidade refere-se à capacidade do programa detectar os verdadeiros negativos. Neste estudo, todos os genes sabidamente negativos foram referentes a *D. ananassae* ou aqueles intrínsecos ao genoma de *S. mansoni*. Para as pontuações entre 3 a 6, os valores da Especificidade conseguiram alcançar o valor 1, em ambos os genomas avaliados. O valor 1, indica que todos os genes falsos, foram detectados como genes falsos - verdadeiros negativos. Para as pontuações 1 e 2, a especificidade ficou com valor 0, devido a todos os genes estarem positivos, ou seja, acima de 3.

A Sensibilidade é capacidade de identificar corretamente os verdadeiros positivos. Neste estudo, os valores de Sensibilidade sofreram variação quando a pontuação atribuída aos genes sabidamente positivos foi alterada. No sentido oposto à Especificidade, os valores atribuídos ao valor de corte das pontuações 1 e 2, apresentaram valor referente a 1; no entanto, assim como para especificidade, todos os genes fornecidos foram positivos (acima de 3). Já para os pontos de corte entre as pontuações de 3 a 6, o valor da Sensibilidade alcançou valores próximos a 0.58. A sensibilidade alcançada indica que, dos genomas estudados, seria possível reduzir a necessidade de curadoria manual em aproximadamente 58%.

Uma das perguntas realizadas para teste da ferramenta, foi se a mesma seria aplicável em genomas considerados fragmentados. O genoma de *S. mansoni* em sua versão 5.2, é chamado de fragmentado por conter 380 *scaffolds*, neste genoma, a ferramenta mostrou-se funcional.

Os estudos de anotação de genoma com geração de dados experimentais de transcritos por meio de RNA-Seq ou outra evidência, fornecem uma rica fonte de informação. Dados de sequenciamento de RNA-Seq, principalmente de ISO-Seq mostraram-se de grande importância na curadoria dos genes, principalmente ao nível de isoformas, conseguindo fazer com que a cobertura da isoforma apresentasse concordância total entre a estrutura predita e o transcrito.

As informações de transcritos, conseguem ainda melhorar o resultado após a pontuação atribuída, principalmente para o valor 3, onde as evidências podem não ser suficientes. A Figura 25 ilustra um exemplo de melhoramento do gene onde o suporte por RNA foi essencial. Por falha na predição do gene, um exon foi erroneamente dividido em dois, com os dados de expressão foi possível corrigir o exon.

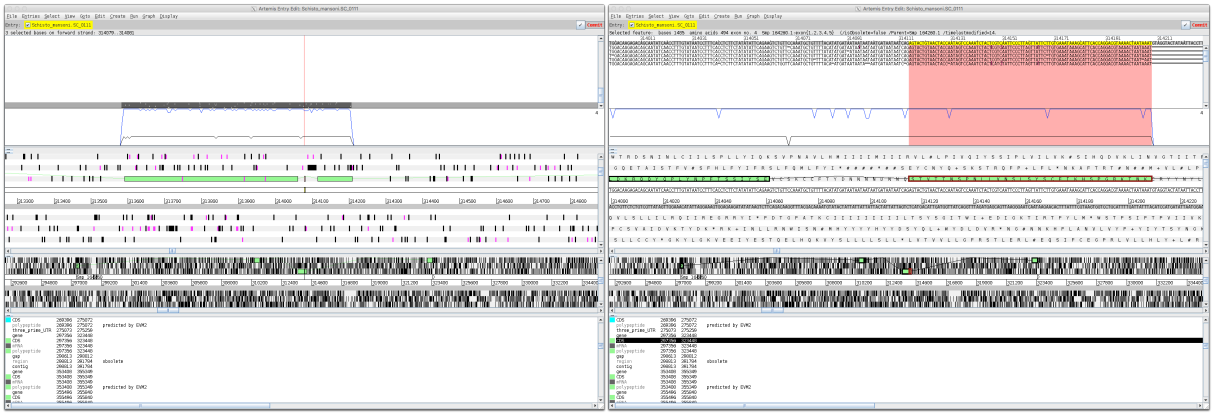

Figura 25: Suporte de ISO-Seq na Correção do Exon. As duas imagens correspondem a mesma região em diferentes perspectivas. Os dois exons preditos estão representados em verde. O suporte do ISO-seq mostra que os dois exons correspondem a apenas um. A imagem a esquerda aproxima a visualização para conferência da sequência.

No entanto, confiar apenas nos dados de RNA-Seq não apresentou ser uma estratégia com uma boa confiança. Por vezes, foi possível obter um gene com total cobertura em extensão, mas quando conferido, este apresentava erros na sequência, onde o módulo Estrutural foi capaz de apontar para o erro. Isso é devido ao modo que a contabilização da cobertura é realizada: as coordenadas dos genes são comparadas com as coordenadas do mapeamento e a cobertura em extensão é adquirida. Por não ser realizado alinhamento de sequência, a mesma pode apresentar divergência em termos de sequência, mas ser igual no tamanho e posição.

Outro exemplo de cobertura total, mas, erro na predição do gene pode ser visualizado na Figura 26, onde está ilustrado dois genes pertencentes ao parasito *S. mansoni* v5.2, ambos os genes com cobertura de 100%. Todavia, quando observado os dados do ISO-Seq, é possível perceber que se trata apenas de um gene. Esse tipo de erro pode ser corrigido por filtros aplicados após o mapeamento, em que uma leitura do RNA-Seq quando mapeada em determinado gene, esta não possa mapear em outro gene (*uniquely mapped reads*). Como o mapeamento deve ser previamente realizado pelo usuário, essa dica é apresentada no tutorial da ferramenta.

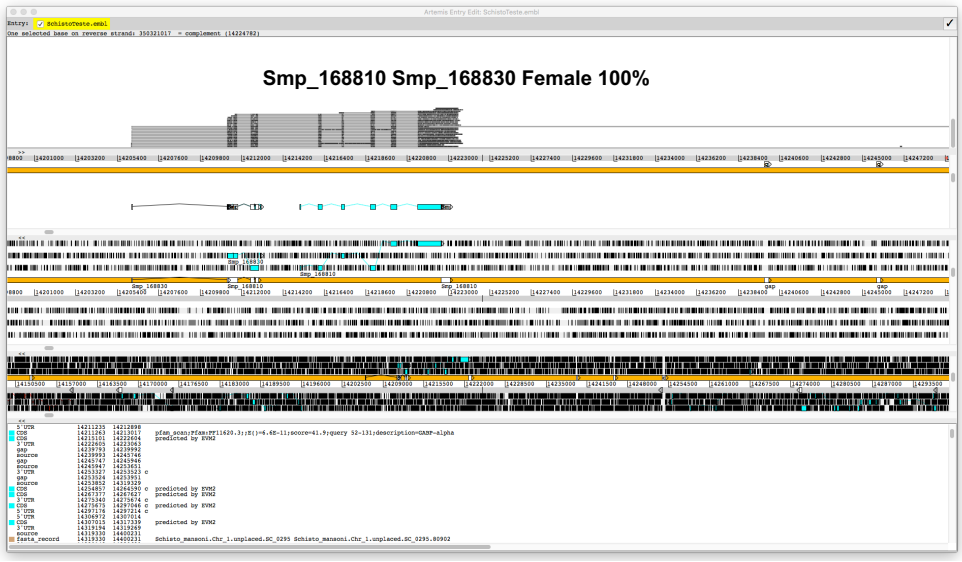

Figura 26: Dois genes Erroneamente Preditos. Os genes do parasito S. mansoni v5.2: Smp_168810 e Smp_168830 estão representados em azul claro, o mapeamento das leituras ISO-Seq corresponde a parte em cinza (cinza escuro corresponde aos exons) na linha superior. Os dados de ISO-Seq, apesar de mapear nos dois genes, evidencia que na verdade trata-se de apenas um gene.

O trabalho aqui desenvolvido não é o único a propor uma pontuação de qualidade para a anotação dos genes codificadores de proteína, é possível encontrar na literatura trabalhos do tipo: (M. A. Dragan et al. 2016; Iliopoulos et al. 2003; Yang, Gilbert, and Kim 2009). Um dos trabalhos, GeneValidator é uma ferramenta online capaz de realizar diversas análises do gene e atribuir uma pontuação a este, contudo, o mesmo não emprega a abordagem de sintenia tampouco de ortólogos; e não consegue avaliar todos os genes da anotação juntos. Outra métrica interessante proposta por esse programa, foi a comparação do comprimento do gene com o gene correspondente do UniProt, todavia, não foi encontrado um valor estatístico que comprava a eficiência nos organismos estudados nesse presente trabalho (dados não mostrados).

Embora não fosse objetivo do presente estudo, o módulo de Sintenia por Ortólogos conseguiu detectar problemas nas montagens de genomas de estudo desse trabalho. Como é de conhecimento, os erros de montagem dos genomas podem afetar a anotação, consequentemente, alguns casos foram apresentados na seção resultados para o genoma de *P. knowlesi*.

Também foi possível detectar erros de montagem ou falta de cobertura do sequenciamento no parasito *S. mansoni* v5.2, onde o gene Smp_041190 (duas isoformas) recebeu pontuação no valor 2 pela ferramenta ASSIS. Esse gene pertencia ao controle positivo por ter passado nos filtros: Estrutural, 100% de cobertura de ISO-Seq, no entanto, o gene está localizado em uma região de gap. Genes anotados em regiões entre ou dentro de intervalos

(gap) estão propensos a erros. Apesar do ISO-seq cobrir toda a região visualizada do gene, por conter um gap no entorno do gene, não é possível afirmar que não exista alguma outra região codificadora que anteceda ao gap. A Figura 27 ilustra esse exemplo.

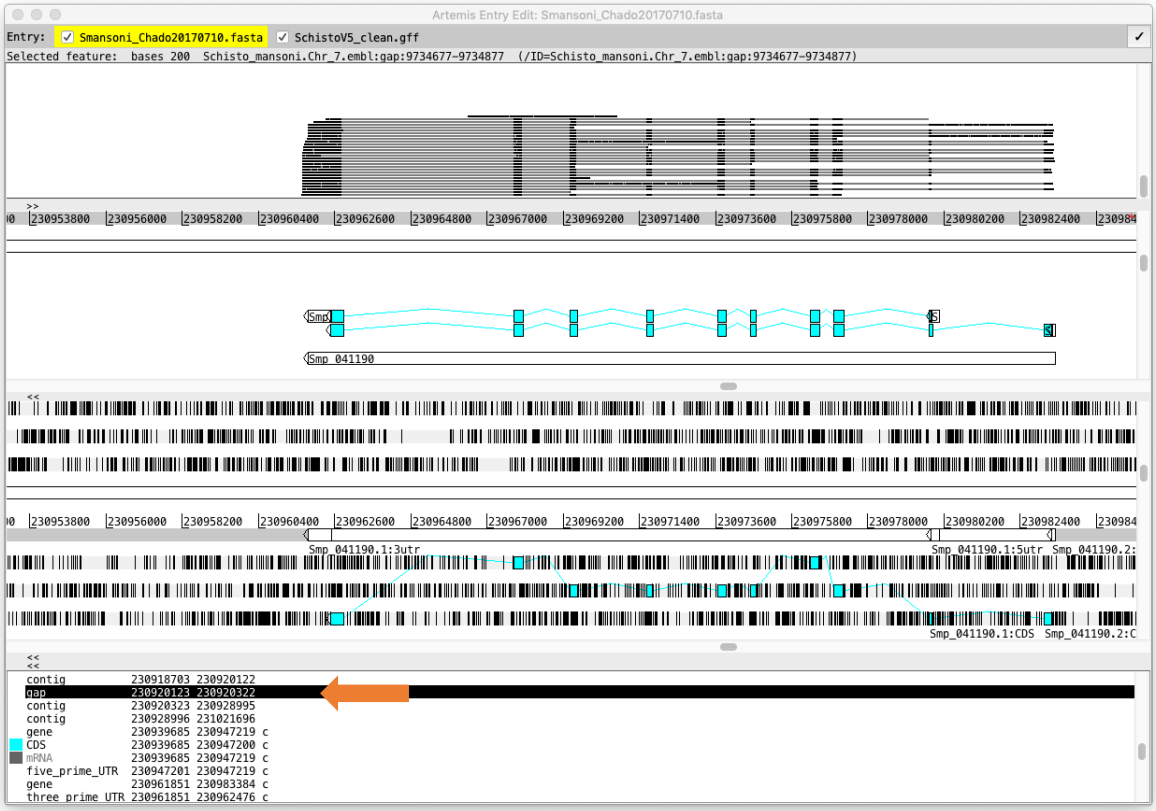

Figura 27: Gap entre contigs. A Seta vermelha indica a região de gap onde estão os dois contigs que constituem o gene.

Outros genomas trabalhados apresentaram problemas referentes a montagem, detectados pela ferramenta ASSIS. A Figura 28 representa dois genomas descritos no Capítulo 1: *P. coatneyi* e *P. knowlesi*. O genoma de *P. knowlesi* é considerado referência devido a qualidade da montagem e anotação. Era esperado uma pequena diferença entre os genomas, haja vista o tempo evolutivo de diferença entre as espécies, no entanto, uma região de destaque é evidenciada na Figura 28, onde o cromossomo 2 de *P. coatneyi* parece se dividir em 2.

Apesar do trabalho de *P. coatneyi* (Chien et al. 2016), ser chamado de alta qualidade, o artigo apresenta uma incongruência na região do cromossomo 2 e 12. Devido a este fato, o genoma de *P. knowlesi* adotou a estratégia de integração dos dados de PacBio com a técnica de Hi-C, para que co-localização pudesse guiar os genes em seus respectivos cromossomos.

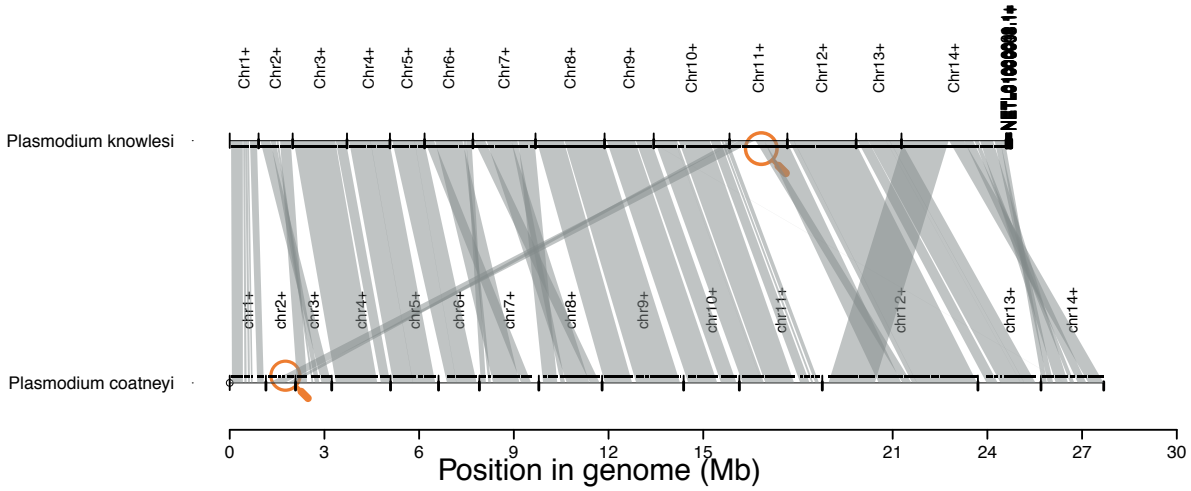

Figura 28: Sintenia de Ortólogos entre *P. knowlesi* e *P. coatneyi*. A figura destaca em vermelho um possível erro de montagem do genoma de *P. coatneyi* detectado pela ferramenta desenvolvido neste estudo.

9 LIMITAÇÕES DAS ANÁLISES

O presente estudo ainda apresenta limitações, entre estas podemos citar:

- Falsos negativos:

O módulo de Sintenia de Ortólogos quando trabalhado sozinho gera falsos negativos, decorrentes da penalidade atribuída por região, penalizando todos os genes após a quebra de sintenia, até que a ordem seja recuperada;

- Pontuação discreta e não contínua:

Os valores atribuídos a cada gene são do tipo discreto, não fracionados. Para uma pontuação mais exata é necessário atribuir peso a cada variável fazendo uma média de aprendizado entre todas;

- Integração do Módulo Transcricional na Pontuação:

Outra limitação apresentada foi a não atribuição de uma pontuação aos dados de evidência de transcrição;

- Genomas Fragmentados:

Certamente, a aplicação do Módulo Sintenia de Ortólogos não será efetiva em genomas muito fragmentados, isto porque a continuidade das regiões é exigida;

- Spliced Leader Trans Splicing:

A maneira atual de representação da anotação, o arquivo do tipo GFF3, apresenta de maneira peculiar a presença de Spliced Leader Trans Splicing. Genes que sofrem esse tipo

de mecanismo são representados em hierarquias de parent;ID como os demais genes, e cada um dos genes contém seu próprio padrão. Até o momento da conclusão deste trabalho não foi possível a adição de organismos que apresentam esse tipo de mecanismo, como por exemplo: *C. elegans*.

10 PERSPECTIVAS

Diante das dificuldades apresentadas neste estudo, limitações ainda estão presentes. Por conseguinte, estas vieram dos testes realizados no presente estudo. Foi possível fazer o levantamento das variáveis e compreende-las, o que está faltando agora é realizar um aprendizado das pontuações, podendo atribuir diferentes pesos a cada evidência.

O trabalho tem como perspectivas:

- Acrescentar os dados de estudos de transcritomas para complementar a pontuação individual dos genes codificadores de proteínas;
- Trabalhar com pesos diferentes para cada evidência, fazendo com que as variáveis discretas se tornem contínuas;
- Aplicar aprendizado de máquina (maching learning) para o reconhecimento de abordagem será possível entender a flutuação de cada variável;
- Adicionar a informação de códon usage para reconhecimento de peseudogenes;
- Integrar em alguma plataforma de anotação.

11 CONCLUSÃO

Diante de todo o trabalho apresentado, é possível concluir que a anotação dos genes codificadores de proteínas continua sendo uma tarefa desafiadora. Como apontado recentemente por Salzberg, 2019; diferentemente da revolução do sequenciamento e das técnicas de montagem dos genomas, a anotação por sua vez, ainda utiliza a mesma tecnologia desenvolvida nas últimas duas décadas. O grande número de genomas requer o uso de procedimentos totalmente automatizados para anotação, mas os erros na anotação são tão prevalentes quanto eram no passado, se não mais (Steven Lb Salzberg 2019).

O trabalho aqui proposto, configura uma saída para apontar os erros na anotação dos genes de uma maneira mais automática. Por meio dos genomas estudados foi possível transcorrer pelos possíveis erros que uma anotação de genes codificadores de proteína pode

apresentar. As evidências fornecidas foram tratadas como variáveis para o desenvolvimento de uma ferramenta automatizada. A ferramenta apresentada, conseguiu reduzir em aproximadamente 60% do esforço da curadoria manual dos genes, nos genomas estudados.

Algumas limitações ainda estão presentes, no entanto, com o conhecimento das variáveis, atender as perspectivas é de interesse da continuidade do trabalho. Podendo alcançar uma ferramenta mais completa que atenda as expectativas de uma necessidade de automatização da anotação com uma alta qualidade.

Recursos Computacionais: Plataforma de Bioinformática Fiocruz Minas e Computadores de uso pessoal.

REFERÊNCIAS

Abdel-Ghany, Salah E et al. 2016. “A Survey of the Sorghum Transcriptome Using Single-Molecule Long Reads. TL - 7.” *Nature communications* 7 VN-re: 11706. http://dx.doi.org/10.1038/ncomms11706.

Adams, Mark D et al. 2000. “The Genome Sequence of Drosophila Melanogaster.” *Science* 287(March): 2196–2204. http://www.ncbi.nlm.nih.gov/pubmed/21793699.

Al-khedery, Basima, John W Barnwell, and Mary R Galinski. 1999. “Antigenic Variation in Malaria : A 3 ¯Genomic Alteration Associated with the Expression of a P . Knowlesi Variant Antigen.” 3: 131–41.

Alexeyenko, Andrey, Ivica Tamas, Gang Liu, and Erik L L Sonnhammer. 2006. “Automatic Clustering of Orthologs and Inparalogs Shared by Multiple Proteomes.” *Bioinformatics* 22(14): 9–15.

Alhakami, Hind, Hamid Mirebrahim, and Stefano Lonardi. 2017. “A Comparative Evaluation of Genome Assembly Reconciliation Tools.” *Genome Biology* 18(1): 1–14. http://dx.doi.org/10.1186/s13059-017-1213-3.

Altenhoff, Adrian M., Adrian Schneider, Gaston H. Gonnet, and Christophe Dessimoz. 2011. “OMA 2011: Orthology Inference among 1000 Complete Genomes.” *Nucleic Acids Research* 39(SUPPL. 1): 289–94.

Ar, Muzaffer. 2016. “Plant Omics: Trends and Applications.” *Plant Omics: Trends and Applications*: 109–35.

Bahl, Amit et al. 2003. “PlasmoDB: The Plasmodium Genome Resource. A Database Integrating Experimental and Computational Data.” *Nucleic Acids Research* 31(1): 212–15.

Baptista, Rodrigo P. et al. 2018. “Assembly of Highly Repetitive Genomes Using Short Reads: The Genome of Discrete Typing Unit III Trypanosoma Cruzi Strain 231.” *Microbial Genomics* 4(4). http://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000156.

Benson, Dennis A Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipma, and James Ostell and EricW. Sayers* N. 2017. “GenBank.” *Nucleic Acids Research* 45(2): 846–60. http://fdslive.oup.com/www.oup.com/pdf/production_in_progress.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27899565.

Berriman, Matthew et al. 2009. “The Genome of the Blood Fluke Schistosoma Mansoni.” *Nature* 460(7253): 352–58. http://www.ncbi.nlm.nih.gov/pubmed/19606141.

Blake, Judith A et al. 2016. “Mouse Genome Database (MGD)-2017: Community Knowledge Resource for the Laboratory Mouse.” *Nucleic acids research* 45(November 2016): gkw1040. http://www.ncbi.nlm.nih.gov/pubmed/27899570.

Borodovsky, Mark. 2015. “NIH Public Access.” 34(3): 474–76.

Bradnam, Keith R et al. 2013. “Assemblathon 2: Evaluating de Novo Methods of Genome Assembly in Three Vertebrate Species.” *GigaScience* 2(1): 10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3844414&tool=pmcentrez&rendertype=abstract (March 21, 2014).

Brendel, Volker, Liqun Xing, and Wei Zhu. 2004. “Gene Structure Prediction from Consensus Spliced Alignment of Multiple ESTs Matching the Same Genomic Locus.” *Bioinformatics* 20(7): 1157–69.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2014. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12(1): 59–60.

Burset, M, and R Guigo. 1998. “Evaluation of Gene Structure Predication Programs.” *UNMC Mcgoogan Library* 655(Sept 10): 12–26. https://ac.els-cdn.com/S0888754396902980/1-s2.0-S0888754396902980-main.pdf?_tid=fc44dde7-7f74-4b09-9982-07089c58d0ec&acdnat=1544102817_45005a695ade8c831a9bc6d39f0822ac.

Camacho, C et al. 2009. “BLAST plus: Architecture and Applications.” *BMC Bioinformatics* 10(421): 1.

Chien, JT et al. 2016. “High-Quality Genome Assembly and Annotation for Plasmodium.” *Genome announcements* 4(5): 4–5.

Clark, Andrew G. et al. 2007. “Evolution of Genes and Genomes on the Drosophila Phylogeny.” *Nature* 450(7167): 203–18.

Clarke, James et al. 2009. “Continuous Base Identification for Single-Molecule Nanopore DNA Sequencing.” *Nature Nanotechnology* 4(4): 265–70.

Crellen, Thomas et al. 2016. “Whole Genome Resequencing of the Human Parasite Schistosoma Mansoni Reveals Population History and Effects of Selection.” *Nature Publishing Group* (October 2015): 1–13. http://dx.doi.org/10.1038/srep20954.

Curwen, Val et al. 2004. “74733_13X.Pdf.” (617): 1–9. sftp://cerca@192.168.2.5/home/cerca/Desktop/data/laptop_files/info/new_mendeley/eukaryotic annotation pipeline/0140942.pdf%5Cnpapers2://publication/uuid/2A400F82-17DB-40CD-A7E5-A5F9FDDD90D1.

Deng, Qiuhong et al. 2007. “Research on the Karyotype and Evolution of Drosophila Melanogaster Species Group.” *Journal of Genetics and Genomics* 34(3): 196–213.

Denton, James F et al. 2014. “Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies.” *PLoS Comput Biol* 10(12).

Dragan, Monica et al. 2016. “GeneValidator: Identify Problems with Protein-Coding Gene Predictions.” *Bioinformatics (Oxford, England)* 2: 1–3. https://github.com/wurmlab/genevalidator.

Dragan, Monica Andreea et al. 2016. “GeneValidator: Identify Problems with Protein-Coding Gene Predictions.” *Bioinformatics* 32(10): 1559–61.

Ederveen, Thomas H.A., Lex Overmars, and Sacha A.F.T. van Hijum. 2013. “Reduce Manual Curation by Combining Gene Predictions from Multiple Annotation Engines, a Case Study of Start Codon Prediction.” *PLoS ONE* 8(5).

Emms, David M, and Steven Kelly. 2015. "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." *Genome Biology* 16(1): 157. http://dx.doi.org/10.1186/s13059-015-0721-2.

Farias, Leonardo P et al. 2011. "Screening the Schistosoma Mansoni Transcriptome for Genes Differentially Expressed in the Schistosomulum Stage in Search for Vaccine Candidates." *Parasitology research* 108(1): 123–35. http://www.ncbi.nlm.nih.gov/pubmed/20852890 (August 28, 2011).

Figueiró, Henrique V. et al. 2017. "Genome-Wide Signatures of Complex Introgression and Adaptive Evolution in the Big Cats." *Science Advances* 3(7): e1700299.

Fu, Limin et al. 2012. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data." *Bioinformatics* 28(23): 3150–52.

Gabaldón, T., and M. A. Huynen. 2004. "Prediction of Protein Function and Pathways in the Genome Era." *Cellular and Molecular Life Sciences* 61(7–8): 930–44.

Goble, Carole et al. 2008. "Data Curation + Process Curation = Data Integration + Science." *Briefings in Bioinformatics* 9(6): 506–17.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews Genetics* 17(6): 333–51. http://dx.doi.org/10.1038/nrg.2016.49.

Grabherr, Manfred G et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature biotechnology* 29(7): 644–52. http://www.ncbi.nlm.nih.gov/pubmed/21572440%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3571712.

Heiko Müller, Naumann, Felix. 2003. "DATA QUALITY IN GENOME DATABASES."

Henikoff, Steven, Michael A. Keene, Kim Fechtel, and James W. Fristrom. 1986. "Gene within a Gene: Nested Drosophila Genes Encode Unrelated Proteins on Opposite DNA Strands." *Cell* 44(1): 33–42.

Heydari, Mahdi et al. 2017. "Evaluation of the Impact of Illumina Error Correction Tools on de Novo Genome Assembly." *BMC Bioinformatics* 18(1): 1–13.

Holt, Carson, and Mark Yandell. 2011. "MAKER2 : An Annotation Pipeline and Genome-Database Management Tool for Second- Generation Genome Projects."

Hu, Tina T., Michael B. Eisen, Kevin R. Thornton, and Peter Andolfatto. 2013. "A Second-Generation Assembly of the Drosophila Simulans Genome Provides New Insights into Patterns of Lineage-Specific Divergence." *Genome Research* 23(1): 89–98.

Iliopoulos, Ioannis et al. 2003. "Evaluation of Annotation Strategies Using an Entire Genome Sequence." *Bioinformatics* 19(6): 717–26.

Illumina. 2017. "An Introduction to Next-Generation Sequencing." *Illumina*: 1–16. https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf.

J., Eid et al. 2009. "Real-Time DNA Sequencing from Single Polymerase Molecules." *Science* 323(5910): 133–38.

http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed11&NEWS=N&AN=354039033.

Jones, Philip et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30(9): 1236–40.

José Domingos Coutinho, Tarcisio, Glória Regina Franco, and Francisco Pereira Lobo. 2015. “Homology-Independent Metrics for Comparative Genomics.” *CSBJ* 13: 352–57. http://creativecommons.org/licenses/by/4.0/.

Jun, Jin, Ion I. Mandoiu, and Craig E. Nelson. 2009. “Identification of Mammalian Orthologs Using Local Synteny.” *BMC Genomics* 10: 1–13.

Kanehisa, M, and S Goto. 2000. “Yeast Biochemical Pathways. KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Res* 28(1): 27–30. http://pathway.yeastgenome.org/biocyc/.

Kanehisa, Minoru et al. 2016. “KEGG as a Reference Resource for Gene and Protein Annotation.” *Nucleic Acids Research* 44(D1): D457–62.

Kaufman, Thomas C. 2017. “A Short History and Description of Drosophila Aberrations , Forward Genetic Screens , and the Nature of Mutations.” *Genetics* 206(June): 665–89.

Kent, W James. 2002. “BLAT--the BLAST-like Alignment Tool.” *Genome research* 12(4): 656–64. http://www.ncbi.nlm.nih.gov/pubmed/11932250%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC187518.

Koonin, Eugene V. 2005. “Orthologs, Paralogs, and Evolutionary Genomics.” *Annual review of genetics* 39: 309–38. http://www.ncbi.nlm.nih.gov/pubmed/16285863.

Korf, Ian. 2004. “Gene Finding in Novel Genomes.” *BMC bioinformatics* 5: 59.

Kumar, Anui. 2009. “An Overview of Nested Genes in Eukaryotic Genomes.” *Eukaryotic Cell* 8(9): 1321–29.

Langmead, Ben, Christopher Wilks, Valentin Antonescu, and Rone Charles. 2019. “Scaling Read Aligners to Hundreds of Threads on General-Purpose Processors.” *Bioinformatics* 35(3): 421–32.

Lapp, S. A. et al. 2018. “PacBio Assembly of a Plasmodium Knowlesi Genome Sequence with Hi-C Correction and Manual Annotation of the SICAvar Gene Family.” *Parasitology* 145(1): 71–84.

Lapp, Stacey A. et al. 2013. “Spleen-Dependent Regulation of Antigenic Variation in Malaria Parasites: Plasmodium Knowlesi SICAvar Expression Profiles in Splenic and Asplenic Hosts.” *PLoS ONE* 8(10): 1–17.

Lapp, Stacey a, Cindy C Korir, and Mary R Galinski. 2009. “Redefining the Expressed Prototype SICAvar Gene Involved in Plasmodium Knowlesi Antigenic Variation.” *Malaria journal* 8: 181.

Lee, Eduardo et al. 2013. “Web Apollo: A Web-Based Genomic Annotation Editing Platform.” *Genome Biology* 14(8): R93. http://genomebiology.com/content/14/8/R93.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway. 2011. “The Sequence Read Archive.” *Nucleic Acids Research* 39(SUPPL. 1): 2010–12.

Li, Li, Christian J Jr Stoeckert, and David S Roos. 2003. “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et Al. 13 (9): 2178 -- Genome Research.” *Genome Research* 13(9): 2178–89. http://genome.cshlp.org/cgi/content/full/13/9/2178.

Logan-klumpler, Flora J et al. 2012. “GeneDB — an Annotation Database for Pathogens.” 40(November 2011): 98–108.

Lu, Zhigang et al. 2016. “Schistosome Sex Matters: A Deep View into Gonad-Specific and Pairing-Dependent Transcriptomes Reveals a Complex Gender Interplay.” *Scientific reports* 6(June): 31150. http://dx.doi.org/10.1038/srep31150.

Marchler-Bauer, Aron et al. 2017. “CDD/SPARCLE: Functional Classification of Proteins via Subfamily Domain Architectures.” *Nucleic Acids Research* 45(D1): D200–203.

McManus, C. Joel et al. 2014. “Evolution of Splicing Regulatory Networks in Drosophila.” *Genome Research* 24(5): 786–96.

Metz, Charles W. “Chromosome Studies on the Diptera 1 1 . the Paired Association.”

METZ, CHARLES W., and MILDRED S. MOSES. 2017. “Chromosomes of Drosophila.” *Journal of Heredity* 14(5): 195–204.

Mitchell, Alex L. et al. 2019. “InterPro in 2019: Improving Coverage, Classification and Access to Protein Sequence Annotations.” *Nucleic Acids Research* 47(D1): D351–60.

Moon, Robert W. et al. 2016. “Normocyte-Binding Protein Required for Human Erythrocyte Invasion by the Zoonotic Malaria Parasite Plasmodium Knowlesi.” *Proceedings of the National Academy of Sciences of the United States of America* 113(26): 201522469. http://www.pnas.org/lookup/doi/10.1073/pnas.1522469113%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27303038.

Myers, Eugene W. et al. 2000. “A Whole-Genome Assembly of Drosophila.” *Science* 287(5461): 2196–2204.

Otto, Thomas D., Gary P. Dillon, Wim S. Degrave, and Matthew Berriman. 2011. “RATT: Rapid Annotation Transfer Tool.” *Nucleic Acids Research* 39(9): 1–7.

Pain, A et al. 2008. “The Genome of the Simian and Human Malaria Parasite Plasmodium Knowlesi.” *Nature* 455(7214): 799–803. http://dx.doi.org/10.1038/nature07306.

Pavesi, Angelo et al. 2018. “Overlapping Genes and the Proteins They Encode Differ Significantly in Their Sequence Composition from Non-Overlapping Genes.” *PLoS ONE* 13(10): 1–24.

Pertea, Mihaela et al. 2015. “StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads.” *Nature Biotechnology* 33(3): 290–95.

Phillippy, Adam M. 2017. “New Advances in Sequence Assembly.” *Genome Research* 27(5): xi–xiii.

Potter, Simon C. et al. 2018. “HMMER Web Server: 2018 Update.” *Nucleic Acids Research*

46(W1): W200–204.

Protasio, Anna V., David W. Dunne, and Matthew Berriman. 2013. “Comparative Study of Transcriptome Profiles of Mechanical- and Skin-Transformed Schistosoma Mansoni Schistosomula.” *PLoS Neglected Tropical Diseases* 7(3): 1–10.

Protasio, Anna V et al. 2012. “A Systematically Improved High Quality Genome and Transcriptome of the Human Blood Fluke Schistosoma Mansoni.” *PLoS neglected tropical diseases* 6(1): e1455. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3254664&tool=pmcentrez&rendertype=abstract (March 9, 2012).

Quinlan, Aaron R, and Ira M Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26(6): 841–42.

Robinson, James T. 2012. “Integrated Genomics Viewer.” *Nature biotechnology* 29(1): 24–26.

Rogers, Rebekah L. et al. 2014. “ Revised Annotations, Sex-Biased Expression, and Lineage-Specific Genes in the Drosophila Melanogaster Group .” *G3: Genes|Genomes|Genetics* 4(12): 2345–51.

Rutherford, K et al. 2000a. “Artemis: Sequence Visualization and Annotation.” *Bioinformatics* 16(10): 944–45.

———. 2000b. “Artemis: Sequence Visualization and Annotation.” *Bioinformatics* 16(10): 944–45. http://www.ncbi.nlm.nih.gov/pubmed/11120685%5Cnhttp://bioinformatics.oxfordjournals.org/content/16/10/944.full.pdf.

Salzberg, Steven L et al. 2012. “GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms Steven L. Salzberg, Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen.” *Genome Research* 22(3): 557–67.

Salzberg, Steven Lb. 2019. “Next-Generation Genome Annotation : We Still Struggle to Get It Right.” : 19–21.

Schnoes, Alexandra M., Shoshana D. Brown, Igor Dodevski, and Patricia C. Babbitt. 2009. “Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies.” *PLoS Computational Biology* 5(12).

Shah, Kinnary, Weihuan Cao, and Christopher E. Ellison. 2019. “Adenine Methylation in *Drosophila* Is Associated with the Tissue-Specific Expression of Developmental and Regulatory Genes.” *G3: Genes|Genomes|Genetics*: g3.400023.2019. http://g3journal.org/lookup/doi/10.1534/g3.119.400023.

Shapland, Elaine B. et al. 2015. “Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process.” *ACS Synthetic Biology* 4(7): 860–66.

Skinner, Mitchell E. et al. 2009. “JBrowse: A next-Generation Genome Browser.” *Genome Research* 19(9): 1630–38.

Slater, Guy St C., and Ewan Birney. 2005. “Automated Generation of Heuristics for

Biological Sequence Comparison." *BMC Bioinformatics* 6: 1–11.

Smith, David Roy. 2016. "Goodbye Genome Paper, Hello Genome Report: The Increasing Popularity of 'genome Announcements' and Their Impact on Science." *Briefings in functional genomics*: elw026. http://www.ncbi.nlm.nih.gov/pubmed/27339634.

Sohn, Jang Il, and Jin Wu Nam. 2018. "The Present and Future of de Novo Whole-Genome Assembly." *Briefings in Bioinformatics* 19(1): 23–40.

Standage, Daniel, and Volker Brendel. 2012. "ParsEval: Parallel Comparison and Analysis of Gene Structure Annotations." *BMC Bioinformatics* 13(1): 187. http://www.biomedcentral.com/1471-2105/13/187%5Cnhttp://www.biomedcentral.com/content/pdf/1471-2105-13-187.pdf.

Stanke, Mario, Mark Diekhans, Robert Baertsch, and David Haussler. 2008. "Using Native and Syntenically Mapped CDNA Alignments to Improve de Novo Gene Finding." *Bioinformatics* 24(5): 637–44.

Steinbiss, Sascha et al. 2016. "*Companion* : A Web Server for Annotation and Analysis of Parasite Genomes." *Nucleic Acids Research* 44(11): gkw292. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkw292.

Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology*: 1–3.

Thibaud-Nissen, Françoise et al. 2013. "Eukaryotic Genome Annotation Pipeline." *The NCBI Handbook [Internet, 2nd Edition]* (Md): 1–24.

Toker, Lilah, Min Feng, and Paul Pavlidis. 2016a. "Whose Sample Is It Anyway? Widespread Misannotation of Samples in Transcriptomics Studies." *F1000Research* 5(0): 2103. http://www.ncbi.nlm.nih.gov/pubmed/27746907.

———. 2016b. "Whose Sample Is It Anyway? Widespread Misannotation of Samples in Transcriptomics Studies." *F1000Research* 5(0): 2103.

Trapnell, Cole et al. 2012. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks." *Nature protocols* 7(3): 562–78. http://dx.doi.org/10.1038/nprot.2012.016.

Vaattovaara, Aleksia, Johanna Leppälä, Jarkko Salojärvi, and Michael Wrzaczek. 2019. "High-Throughput Sequencing Data and the Impact of Plant Gene Annotation Quality." *Journal of Experimental Botany* 70(4): 1069–76.

Vallender, Eric J. 2010. "Evolutionary Relationships." 49(1): 50–55.

Wasmuth, Elizabeth V, and Christopher D Lima. 2016. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45(November 2016): 1–12. http://fdslive.oup.com/www.oup.com/pdf/production_in_progress.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27899565.

Weber, James L. 1987. "Analysis of Sequences from the Extremely A + T-Rich Genome of PIasmodium Falciparum." 52: 103–9.

Wit, Janneke, and John Stuart Gilleard. 2017. "Re-Sequencing Helminth Genomes for

Population and Genetic Studies." *Trends in Parasitology* xx: Accepted. http://dx.doi.org/10.1016/j.pt.2017.01.009.

Wu, Thomas D., and Colin K. Watanabe. 2005. "GMAP: A Genomic Mapping and Alignment Program for MRNA and EST Sequences." *Bioinformatics* 21(9): 1859–75.

Yandell, Mark, and Daniel Ence. 2012. "A Beginner ' s Guide to Eukaryotic Genome Annotation." *Nature Publishing Group* 13(5): 329–42. http://dx.doi.org/10.1038/nrg3174.

Yang, Youngik, Donald Gilbert, and Sun Kim. 2009. "Annotation Confidence Score for Genome Annotation: A Genome Comparison Approach." *Bioinformatics* 26(1): 22–29.

Zimin, Aleksey V., Douglas R. Smith, Granger Sutton, and James A. Yorke. 2008. "Assembly Reconciliation." *Bioinformatics* 24(1): 42–45.

MATERIAL SUPLEMENTAR

Todo o material suplementar do trabalho está disponível em:

https://github.com/Juassis/Material_Suplementar_Tese

ANEXO - PRODUÇÃO CIENTÍFICA

A seguir serão listados os artigos científicos publicados e/ou aprovados os quais pude participar durante o período do doutorado.

Fui responsável pela anotação dos seguintes genomas: Jaguar, *Plasmodium*, *Echinococcus* (não descrito na tese) *Anopheles* (não descrito na tese). O trabalho do parasito *Schistosoma mansoni* está em preparação para submissão.

No trabalho da *Biomphalaria glabrata*, não fiz parte da equipe de montagem e anotação, estando envolvida apenas nas análises de variação. No entanto, no congresso de Esquistossomose de 2018 foi acordado minha participação no melhoramento da anotação.

O trabalho envolvendo o Banco de Dados: ZIKDB, fui responsável pela curadoria manual dos trabalhos.

O trabalho envolvendo *Rhodnius prolixus*, fui responsável por todas as análises de bioinformática.

American Association for the Advancement of *Science:* Science Advances **PMID: 28776029**

Genome-wide signatures of complex introgression and adaptive evolution in the big cats

Henrique V. Figueiró, Gang Li, Fernanda J. Trindade, **Juliana Assis Geraldo,** Fabiano Pais, Gabriel Fernandes, Sarah H. D. Santos, Graham M. Hughes, Aleksey Komissarov, Agostinho Antunes, Cristine S. Trinca, Maíra R. Rodrigues, Tyler Linderoth, Ke Bi, Leandro Silveira, Fernando C. C. Azevedo, Daniel Kantek, Emiliano Ramalho, Ricardo A. Brassaloti, Priscilla M. S. Villela, Adauto L. V. Nunes, Rodrigo H. F. Teixeira, Ronaldo G. Morato, Damian Loska, Patricia Saraguëta, Toni Gabaldón, Emma C. Teeling, Stephen J. O'Brien, Rasmus Nielsen, Luiz L. Coutinho, Guilherme Oliveira, William J. Murphy, Eduardo Eizirik

The great cats of genus Panthera comprise a recent radiation whose evolutionary history is poorly understood. Their rapid diversification poses challenges to resolving their phylogeny, while offering opportunities to investigate the historical dynamics of adaptive divergence. Here we report the sequence, de novo assembly and annotation of the jaguar (Panthera onca) genome, a novel genome sequence for the leopard (P. pardus), and comparative analyses encompassing all living Panthera species. Demographic reconstructions indicated that all of these species have experienced variable episodes of population decline during the Pleistocene, ultimately leading to small effective sizes in present-day genomes. We observed pervasive genealogical discordance across Panthera genomes, caused by both incomplete lineage sorting and complex patterns of historical interspecific hybridization. We identified multiple signatures of species-specific positive selection, affecting genes involved in craniofacial and limb development, protein metabolism, hypoxia, reproduction, pigmentation and sensory perception. There was remarkable concordance in pathways enriched in genomic segments implicated in interspecies introgression and in positive selection, suggesting that these processes were connected. We tested this hypothesis by developing exome capture probes targeting ~19,000 Panthera genes and applying them to 30 wild-caught jaguars. We found at least two genes (DOCK3 and COL4A5, both related to optic nerve development) bearing significant signatures of interspecies introgression and within-species positive selection. These findings indicate that post- speciation admixture has contributed genetic material that facilitated the adaptive evolution of big cat lineages.

Cambridge University Press: Parasitology – **PMID: 28720171**

PacBio assembly of a *Plasmodium knowlesi* genome sequence with Hi-C correction and manual annotation of the SICAvar gene family

S.A. Lapp*, **J.A. Geraldo*,** J-T. Chien, F. Ay, S. Pakala, G. Batugedara, J. Humphrey, J.D. DeBarry, K.G. Le Roch, M.R. Galinski and J.C. Kissinger

* represents equal first authorship

SUMMARY

Plasmodium knowlesi has risen in importance as a zoonotic parasite that has been causing regular episodes of malaria throughout South East Asia. The P. knowlesi genome sequence generated in 2008 highlighted and confirmed many similarities and differences in Plasmodium species, including a global view of several multigene families, such as the large SICAvar multigene family encoding the variant antigens known as the Schizont Infected Cell Agglutination (SICA) proteins. However, repetitive DNA sequences are the bane of any genome project and this and other Plasmodium genome projects have not been immune to the gaps, rearrangements and other pitfalls created by these genomic features. Today, long-read PacBio and chromatin conformation technologies are overcoming such obstacles. Here, based on the use of these technologies, we present a highly refined de novo P. knowlesi genome sequence of the Pk1(A+) clone. This sequence and annotation, referred to as the "MaHPIC Pk genome sequence", includes manual annotation of the SICAvar gene family with 129 full-length members categorized as Type I or Type II. This sequence provides a framework that will permit a better understanding of the SICAvar repertoire, selective pressures acting on this gene family, and mechanisms of antigenic variation in this species and other pathogens.

Key Words: Plasmodium, knowlesi, PacBio, Hi-C, SICAvar, MaHPIC, genome, sequence, annotation, antigenic variation

ASM Journal Press Releases: Genome Announcements - **PMID:** 27587810

High-Quality Genome Assembly and Annotation for *Plasmodium coatneyi*, Generated Using Single-Molecule Real-Time PacBio Technology

Jung-Ting Chien,Suman B. Pakala, **Juliana A. Geraldo**,Stacey A. Lapp, Jay C. Humphrey, John W. Barnwell,Jessica C. Kissinger, Mary R. Galinski

Plasmodium coatneyi is a protozoan parasite species that causes simian malaria and is an excellent model for studying disease caused by the human malaria parasite, *P. falciparum*. Here we report the complete (nontelomeric) genome sequence of *P. coat- neyi* Hackeri generated by the application of only Pacific Biosciences RS II (PacBio RS II) single-molecule real-time (SMRT) high-resolution sequence technology and assembly using the Hierarchical Genome Assembly Process (HGAP). This is the first *Plasmodium* genome sequence reported to use only PacBio technology. This approach has proven to be superior to short-read only approaches for this species.

BioMed Central: BMC Genomics – **PMID: 28241794**

The Echinococcus canadensis (G7) genome:a key knowledge of parasitic platyhelminthhuman diseases

Lucas L. Maldonado1, **Juliana Assis Geraldo**, Flávio M. Gomes Araújo, Anna C. M. Salim, Natalia Macchiaroli,Marcela Cucher, Federico Camicia, Adolfo Fox, Mara Rosenzvit, Guilherme Oliveira and Laura Kamenetzky1

Background: The parasite Echinococcus canadensis (G7) (phylum Platyhelminthes, class Cestoda) is one of thecausative agents of echinococcosis. Echinococcosis is a worldwide chronic zoonosis affecting humans as well asdomestic and wild mammals, which has been reported as a prioritized neglected disease by the World HealthOrganisation. No genomic data, comparative genomic analyses or efficient therapeutic and diagnostic tools areavailable for this severe disease. The information presented in this study will help to understand the peculiarbiological characters and to design species-specific control tools.
Results: We sequenced, assembled and annotated the 115-Mb genome of E. canadensis (G7). Comparativegenomic analyses using whole genome data of three Echinococcus species not only confirmed the status of E.canadensis (G7) as a separate species but also demonstrated a high nucleotide sequences divergence in relation to E. granulosus (G1). The E. canadensis (G7) genome contains 11,449 genes with a core set of 881 orthologs sharedamong five cestode species. Comparative genomics revealed that there are more single nucleotide polymorphisms(SNPs) between E. canadensis (G7) and E. granulosus (G1) than between E. canadensis (G7) and E. multilocularis. Thisresult was unexpected since E. canadensis (G7) and E. granulosus (G1) were considered to belong to the speciescomplex E. granulosus sensu lato. We described SNPs in known drug targets and metabolism genes in the E.canadensis (G7) genome. Regarding gene regulation, we analysed three particular features: CpG island distributionalong the three Echinococcus genomes, DNA methylation system and small RNA pathway. The results suggest theoccurrence of yet unknown gene regulation mechanisms in Echinococcus.
Conclusions: This is the first work that addresses Echinococcus comparative genomics. The resources presentedhere will promote the study of mechanisms of parasite development as well as new tools for drug discovery. Theavailability of a high-quality genome assembly is critical for fully exploring the biology of a pathogenic organism.The E. canadensis (G7) genome presented in this study provides a unique opportunity to address the geneticdiversity among the genus Echinococcus and its particular developmental features. At present, there is no unequivocaltaxonomic classification of Echinococcus species; however, the genome-wide SNPs analysis performed here revealedthe phylogenetic distance among these three Echinococcus species. Additional cestode genomes need to be sequencedto be able to resolve their phylogeny.Keywords: Echinococcus genome, SNPs, Drug targets, Helminth parasites, Comparative genomics

Nature: Nature communication – **PMID: 28508897**

Whole genome analysis of a schistosomiasis-transmitting freshwater snail

Coen M Adema, LaDeana W Hillier, Catherine S Jones, Eric S Loker1, Matty Knight, Patrick Minx, Guilherme Oliveira, Nithya Raghavan, Andrew Shedlock, Laurence Rodrigues do Amaral, Halime D Arican-Goktas, **Juliana A Geraldo**, Elio Hideo Baba, Olga L Baron, Christopher J Bayne, Utibe Bickham-Wright, Kyle K Biggar, Michael Blouin, Bryony C Bonning, Chris Botka, Joanna M Bridger, Katherine M Buckley,Sarah K Buddenborg, Roberta Lima Caldeira, Julia Carleton.., Richard K Wilson2, *et al*

Biomphalaria snails are instrumental in transmission of the human blood fluke Schistosoma mansoni.With the World Health Organization's goal to eliminate schistosomiasis as a global health problemby 2025, there is now renewed emphasis on snail control. Here, we characterize the genome of Biomphalaria glabrata, a lophotrochozoan protostome, and provide timely and importantinformation on snail biology. We describe aspects of phero-perception, stress responses, immune function and regulation of gene expression that support the persistence of B. glabrata in the field and may define this species as a suitable snail host for S. mansoni. We identify several potential targets for developing novel control measures aimed at reducing snail mediated transmission of schistosomiasis.

PLOS Neglected Tropical Diseases **PMID: 27332714**

ZIKV – CDB: A Collaborative Database to Guide Research Linking SncRNAs and ZIKA Virus Disease Symptoms

Victor Satler Pylro, Francislon Silva Oliveira, Daniel Kumazawa Morais, Sara Cuadros-Orellana, Fabiano Sviatopolk-Mirsky Pais, Julliane Dutra Medeiros, **Juliana Assis Geraldo**, Jack Gilbert, Angela Cristina Volpini, Gabriel Rocha Fernandes

Background

In early 2015, a ZIKA Virus (ZIKV) infection outbreak was recognized in northeast Brazil, where concerns over its possible links with infant microcephaly have been discussed. Providing a causal link between ZIKV infection and birth defects is still a challenge. MicroRNAs (miRNAs) are small noncoding RNAs (sncRNAs) that regulate post-transcriptional gene expression by translational repression, and play important roles in viral pathogenesis and brain development. The potential for flavivirus-mediated miRNA signalling dysfunction in brain-tissue development provides a compelling hypothesis to test the perceived link between ZIKV and microcephaly.

Methodology/Principal Findings

Here, we applied in silico analyses to provide novel insights to understand how Congenital ZIKA Syndrome symptoms may be related to an imbalance in miRNAs function. Moreover, following World Health Organization (WHO) recommendations, we have assembled a database to help target investigations of the possible relationship between ZIKV symptoms and miRNA-mediated human gene expression.

Conclusions/Significance

We have computationally predicted both miRNAs encoded by ZIKV able to target genes in the human genome and cellular (human) miRNAs capable of interacting with ZIKV genomes. Our results represent a step forward in the ZIKV studies, providing new insights to support research in this field and identify potential targets for therapy.

Springer: Mammalian Genome - PMID: 27853861

Whole genome sequencing of Guzera cattle reveals genetic variants in candidate genes for production, disease resistance, and heat tolerance

Izinara C. Rosse, **Juliana A. Geraldo**, Francislon S. Oliveira, Laura R. Leite, Flávio Araujo, Adhemar Zerlotini, Angela Volpini, Anderson J. Dominitini, Beatriz C. Lopes, Wagner A. Arbex, Marco A. Machado, Maria G.C.D. Peixoto, Rui S. Verneque, Marta F.Martins, Roney S. Coimbra, Marcos V.G.B. Silva, Guilherme Oliveira, Maria Raquel S. Carvalho

In bovines, artificial selection has produced a large number of breeds which differ in production, envi- ronmental adaptation, and health characteristics. To investigate the genetic basis of these phenotypical differ- ences, several bovine breeds have been sequenced. Mil- lions of new SNVs were described at every new breed sequenced, suggesting that every breed should be sequenced. Guzerat or Guzera´ is an indicine breed resistant to drought and parasites that has been the base for some important breeds such as Brahman. Here, we describe the sequence of the Guzera´ genome and the in silico functional analyses of intragenic breed-specific variations. Mate-paired libraries were generated using the ABI SOLiD system. Sequences were mapped to the Bos taurus refer- ence genome (UMD 3.1) and 87% of the reference genome was covered at a 26X. Among the variants identified, 2,676,067 SNVs and 463,158 INDELs were homozygous, not found in any database searched, and may represent true differences between Guzera and B. taurus. Functional analyses investigated with the NGS-SNP package focused on 1069 new, non-synonymous SNVs, splice-site variants (including acceptor and donor sites, and the conserved regions at both intron borders, referred to here as splice regions) and coding INDELs (NS/SS/I). These NS/SS/I map to 935 genes belonging to cell communication, envi- ronmental adaptation, signal transduction, sensory, and immune systems pathways. These pathways have been involved in phenotypes related to health, adaptation to the environment and behavior, and particularly, disease resis- tance and heat tolerance. Indeed, 105 of these genes are known QTLs for milk, meat and carcass, production, reproduction, and health traits. Therefore, in addition to describing new genetic variants, our approach provided groundwork for unraveling key candidate genes and mutations.

Parasites & Vectors: PMID: 31208458

Transcriptome-based molecular systematics: Rhodnius montenegrensis (Triatominae) and its position within the Rhodnius prolixus-Rhodnius robustus cryptic-species complex

Raíssa N Brito**, Juliana A Geraldo**, Fernando A Monteiro, Cristiano Lazoski, Rita CM Souza, Fernando Abad-Franch

Background: Rhodnius montenegrensis (Triatominae) was described after R. robustus-like bugs from southwestern Amazonia. Mitochondrial cytb sequence near-identity with sympatric R. robustus (genotype II) raised doubts about R. montenegrensis' taxonomic status, but comparative studies reported fairly clear morphologic and genetic differences between R. montenegrensis and laboratory stocks identified as R. robustus. Here, we use a transcriptome-based approach to investigate this apparent paradox.
Results: We retrieved publicly available transcriptome sequence-reads from R. montenegrensis and from the R. robustus stocks used as the taxonomic benchmark in comparative studies. We (i) aligned transcriptome sequence-reads to mitochondrial (cytb) and nuclear (ITS-2, D2-28S, and AmpG) query sequences (47 overall) from members of the R. prolixus-R. robustus cryptic-species complex and related taxa; (ii) computed breadth- and depth-coverage for the 259 consensus sequences generated by these alignments; and, for each locus, (iii) appraised query sequences and full-breadth- coverage consensus sequences in terms of nucleotide-sequence polymorphism and phylogenetic relations. We found evidence confirming that R. montenegrensis and R. robustus genotype II are genetically indistinguishable – and hence implying that they are, in all likelihood, the same species. Furthermore, we found compelling genetic evidence that the benchmark 'R. robustus' stocks used in R. montenegrensis description and in later transcriptome-based comparisons are, in reality, R. prolixus – although likely mixed to some degree with R. robustus (probably genotype II, a.k.a. R. montenegrensis).
Conclusions: We illustrate how public-domain genetic/transcriptomic data can help address challenging issues in disease-vector systematics. In our case-study, taxonomic confusion apparently stemmed from misinterpretation of sequence-data analyses and misidentification of taxonomic-benchmark stocks. More generally, and together with previous reports of mixed and/or misidentified Rhodnius spp. laboratory colonies, our results call into question the conclusions of many studies (on, e.g., morphology, genetics, physiology, behavior, bionomics, or interactions with microorganisms including trypanosomes) based on non-genotyped 'R. prolixus' or 'R. robustus' stocks. Because R. prolixus is a primary vector of Chagas disease, whereas R. robustus s.l. comprises a suite of sylvatic species (including R. montenegrensis) of limited medical relevance, these findings are likely to have public-health implication

Plos One: PMID:31479460

Characterization of the complete mitogenome of *Anopheles aquasalis*, and phylogenetic divergences among *Anopheles* from diverse geographic zones

MARTINEZ-VILLEGAS, L., **ASSISS-GERALDO**, J., KOERICH, L. B., COLLIER, T. C., LEE, Y., MAIN, B. J., LANZARO, G. C., AND PIMENTA, P. F. P.

Whole mitogenome sequences (mtDNA) have been exploited for insect ecology studies, using them as molecular markers to reconstruct phylogenies, or to infer phylogeographic relationships and gene flow. Recent *Anopheles* phylogenomic studies have provided information regarding the time of deep lineage divergences within the genus. Here we report the complete 15,393 bp mtDNA sequence of *Anopheles aquasalis*, a Neotropical human malaria vector. When comparing its structure and base composition with other relevant and available anopheline mitogenomes, high similarity and conserved genomic features were observed. Furthermore, 22 mtDNA sequences comprising anopheline and dipteran sibling species were analyzed to reconstruct phylogenies and estimate dates of divergence between taxa. Phylogenetic analysis using complete mtDNA sequences suggests that *A. aquasalis* diverged from the *Anopheles albitarisis* complex ~28 million years ago (MYA), and ~38 MYA from *Anopheles darlingi*. Bayesian analysis suggests that the most recent ancestor of *Nyssorhynchus* and *Anopheles* + *Cellia* was extant ~83 MYA, corroborating current estimates of ~79-100 MYA. Additional sampling and publication of African, Asian, and North American anopheline mitogenomes would improve the resolution of the *Anopheles* phylogeny and clarify early continental dispersal routes.

Keywords: *Anopheles*, mitogenome, phylogenomics, divergence time, malaria vector, Brazil.

PARTICIPAÇÕES EM CONGRESSOS, CURSOS, ESTÁGIOS

Participação em eventos:
13 th ISCB Student Council Symposium. Orlando, USA. 2016.
24nd Intelligent Systems for Molecular Biology (ISMB). Orlando, USA. 2016.
Xmeeting, 12, 13.
Esquistossomose

Organização de Eventos:
13 th ISCB Student Council Symposium. Orlando, USA. 2016.
1st Brazilian Student Council Symposium - 12th X-meeting 2016
2nd Brazilian Student Council Symposium - 13th X-meeting 2017
3rd Brazilian Student Council Symposium - 14th X-meeting 2018

Cursos:
EupathDB Workshop - University of Georgia, Athens, GA, USA

Estágios:
Doutorado Sanduíche no Instituto de Bioinformática da Unversity of Georgia, Athens – Georgia – Estados Unidos, com bolsa provida pelo programa Ciências Sem Fronteiras, duração de 12 meses.

Estágio de curta duração no Wellcome Sanger Institute, em Hinxton, Inglaterra. Suporte financeiro provido pelo projeto aprovado de Pesquisador Visitante Especial do Ciências Sem Fronteiras.

Estágio de curta duração UGA, 3 meses. Sem bolsa.