UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS BIOLÓGICAS
CURSO DE BACHARELADO EM CIÊNCIAS BIOLÓGICAS

Vilmar Benetti Filho

Revistando o genoma de *Eimeria* spp.: resquícios de marcadores de patogenicidade (ROP, SAG e pseudogenes) de coccidiose aviária

Florianópolis
2021

Vilmar Benetti Filho

Revistando o genoma de *Eimeria* spp.: resquícios de marcadores de patogenicidade (ROP, SAG e pseudogenes) de coccidiose aviária

Trabalho de Conclusão do Curso de Graduação em Ciências Biológicas do Centro de Ciências Biológicas da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel em Ciências Biológicas.
Orientador: Prof. Glauber Wagner, Dr.
Coorientador: Guilherme Augusto Maia, M.e.

Florianópolis
2021

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Benetti Filho, Vilmar
Revistando o genoma de Eimeria spp.: resquícios de marcadores de patogenicidade (ROG, SAG e pseudogenes) de coccidiose aviária / Vilmar Benetti Filho ; orientador, Glauber Wagner, coorientador, Guilherme Augusto Maia, 2021.
71 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Graduação em Ciências Biológicas, Florianópolis, 2021.

Inclui referências.

1. Ciências Biológicas. 2. Bioinformática. 3. Genômica. 4. Patogenicidade. 5. Pseudogenes. I. Wagner, Glauber. II. Maia, Guilherme Augusto. III. Universidade Federal de Santa Catarina. Graduação em Ciências Biológicas. IV. Título.

Vilmar Benetti Filho

Revistando o genoma de *Eimeria* spp.: resquícios de marcadores de patogenicidade (ROP, SAG e pseudogenes) de coccidiose aviária

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de "Bacharel em Ciências Biológicas" e aprovado em sua forma final pelo Curso de Ciências Biológicas

Florianópolis, 19 de Abril de 2021

_____

Prof. Dr. Carlos Roberto Zanetti

Coordenador do Curso

Banca Examinadora:

_____

Prof. Dr. Glauber Wagner

Orientador

UFSC

_____

Profa. Dra. Priscila de Oliveira Moraes

Avaliadora

UFSC

_____

Prof. Dr. Guilherme Toledo e Silva

Avaliador

UFSC

_____

Prof. MSc. Renato Simões Moreira

Avaliador Suplente

IFSC

AGRADECIMENTOS

Meus agradecimentos a todos e todas que contribuíram de forma direta ou indireta para a realização deste trabalho. Primeiramente, gostaria de agradecer à entidade portadora de todo conhecimento e poder, mestre do universo, que nas horas mais difíceis, nos momentos de angústia e aflição abençoou-me e mostrou-me o caminho certo. *Hail* Jonas Guedes.

Agradeço aos meus pais, Leoni e Vilmar, pela educação e valores que me ensinaram. Vocês são meus heróis. Agradeço às minhas irmãs: Luana, minha grande inspiração como cientista, e Beatris pelo incentivo, alegria e carinho sempre presente. Me orgulho muito desta família e amo muito vocês.

Agradeço ao meu orientador, Glauber, pelo acolhimento, paciência e compreensão ao longo dos anos no laboratório; ao meu amigo, colega e coorientador Guilherme Maia, pelas discussões, conselhos e opiniões, que sem dúvida contribuíram para meu crescimento pessoal e visão crítica como cientista; aos meus amigos e colegas de laboratório, ou como o Eric chama: "*a nata da bioinfo*": Eric, Renato, Taty, Karin e Jaime, pelos momentos felizes que passamos juntos. Agradeço aos meus amigos de longa data, Luiz, Eduardo e Guilherme, pela paciência ao me ouvir falar sobre meu trabalho, pelo apoio e amizade incrível de vocês; aos amigos que a biologia me deu, em especial a família Losers: Alisson, Gustavo, Laura, Lígia, Rafael, Garbo e Ross, por existirem e estarem presentes na minha vida. Hoje somos winners. Vocês todos estão em meu coração, metaforicamente.

Agradeço a UFSC pela oportunidade de viver este curso, pelo mix de experiências que me proporcionou e que contribuíram para a construção do homem que sou hoje. Agradeço ao CNPq, instituição financiadora da pesquisa brasileira, importante para o desenvolvimento científico do país, pelos projetos financiados.

“Bioinformatics is both an engineering art and a Science.” (GOODMAN, 2002)

RESUMO

A coccidiose aviária é uma doença causada por sete espécies do gênero *Eimeria*, pertencentes ao filo Apicomplexa, caracterizado principalmente pela presença de uma região denominada complexo apical. A coccidiose aviária causa perdas à indústria avícola na ordem de milhões de dólares anualmente, de modo que o estudo do conteúdo genômico e produtos gênicos destes parasitos pode auxiliar no desenvolvimento de métodos de diagnóstico, profilaxia e vacinação. De acordo com a literatura, os genes pertencentes às famílias gênicas ROP e SAG são os principais fatores responsáveis pela patogenicidade do gênero. Desta forma, o objetivo deste trabalho foi remontar os genomas, predizer e anotar os seus produtos gênicos, assim como buscar pseudogenes marcadores das famílias ROP e SAG nas sete espécies de *Eimeria* causadoras da coccidiose aviária. Primeiramente, foram obtidos os dados de sequenciamento disponíveis na base de dados do GenBank e, em seguida, efetuado o controle de qualidade e montagem dos genomas com base em referência. Realizou-se a predição de proteínas, RNAs e pseudogenes, sendo que suas funções foram estabelecidas com base em similaridade com os bancos de dados SwissProt e ToxoDB. Todas as montagens apresentaram melhoria nos indicadores, com destaque para a redução no número de *scaffolds*, cuja menor redução foi de 17% para *E. necatrix* e a maior redução foi de 73%, para *E. brunetti*. A melhoria dos indicadores foi possível sem perda de informação, ou seja, conteúdo GC, tamanho de genoma e número de repetições mantiveram-se semelhantes aos genomas depositados no GenBank. As predições de RNAs acompanham esta tendência, mantendo o número de genes dos genomas de referência. Foram preditos snoRNAs, cuja comparação com as respectivas referências não é possível, de modo que a identificação desses produtos pode auxiliar a formular e responder perguntas futuramente. Houve um salto no número de proteínas preditas em todos os genomas, bem como a diminuição do número de proteínas hipotéticas. As famílias ROP e SAG foram identificadas em níveis semelhantes ao encontrado nas referências. Os pseudogenes preditos mostram que apenas a família SAG deixa marcas no genoma de possíveis inativações ou fragmentações de genes ao longo do processo evolutivo. Para a família ROP poucos pseudogenes foram identificados. A predição de pseudogenes provoca dúvidas sobre a possível função regulatória destas sequências, já descrita na literatura, mas ainda não identificados para os parasitos causadores da coccidiose aviária. Deste modo, este trabalho representa um salto na melhoria dos genomas e identificação de novas estruturas que possam vir a contribuir com novas pesquisas e estratégias de mitigação da doença.

Palavras-chave: Bioinformática, genômica, patogenicidade, coccídios, pseudogenes.

ABSTRACT

Avian coccidiosis is a disease caused by seven parasites of the genus *Eimeria* and phylum Apicomplexa, which are characterized by the presence of a region known as the apical complex. According to the literature, the genes belonging to the ROP and SAG gene families are the main factors responsible for the pathogenicity of this genus. Avian coccidiosis causes millions of dollars worth of damage to the poultry industry annually, so the study of genomic content and gene products of these parasites can assist in the development of diagnostic, prophylactic, and vaccination methods. This study aimed to reassemble the available genomic data, predict and annotate their gene products, as well as search for pseudogenes markers from the ROP and SAG families in the *Eimeria* species that cause avian coccidiosis. First, the genomic sequencing data available in the GenBank database were obtained and then quality control and genome assembly were carried out based on the current reference. Proteins, RNAs, and pseudogenes were predicted, and their functions were established based on similarity with the SwissProt and ToxoDB databases. All assembly indicators were improved, particularly the number of scaffolds, with reductions ranging from 17%, for *E. necatrix*, to 73%, for *E. brunetti*. The improvement of these indicators was possible without loss of information, that is, GC content, genome size, and the number of repetitions remained similar to the genomes deposited in GenBank. RNA predictions also follow this trend, maintaining the number of genes in the reference genomes. SnoRNAs were predicted, although comparisons with reference genomes were not possible, and because of this, the identification of these products can help formulate and answer questions in the future. There was an increase in the number of predicted proteins in all genomes, as well as a decrease in the number of hypothetical proteins. The ROP and SAG families were identified similarly to those found in the references. The predicted pseudogenes show that only one SAG family leaves marks in the genome of possible inactivations or fragmentations of genes throughout the evolutionary process. For the ROP family, few pseudogenes have been identified. The prediction of pseudogenes causes doubt about the regulatory function of these sequences, as already described in the literature, but have not yet been identified for the parasites that cause avian coccidiosis. In this context, this work represents a leap in the improvement of genome assemblies and the identification of new structures that will contribute to new research and strategies to mitigate the disease.

Keywords: Bioinformatics, genomics, pathogenicity, coccidia, pseudogenes

LISTA DE FIGURAS

FIGURA 1 - OOCISTOS ESPORULADOS DE *Eimeria* SPP. 14
FIGURA 2 - COMPARAÇÃO DE INTESTINO INFECTADO POR E. TENELLA COM O DE UMA AVE SAUDÁVEL..... 15
FIGURA 3 - CICLO DE VIDA DE *Eimeria* SPP..... 16
FIGURA 4 - METODOLOGIA DE MONTAGEM DOS GENOMAS DE *Eimeria* SPP. 34
FIGURA 5 - ADAPTAÇÃO DA ANNOTAPIPELINE UTILIZADA PARA ANOTAÇÃO DAS PROTEÍNAS DE *Eimeria* SPP. 39

LISTA DE TABELAS

TABELA 1 - GENOMAS DE REFERÊNCIA DE *Eimeria* SPP. 20
TABELA 2 - DADOS BRUTOS POR GENOMA 30
TABELA 3 - COMPARATIVO ENTRE AS MONTAGENS DE *Eimeria* SPP. REALIZADAS NESTE ESTUDO E SEUS RESPECTIVOS GENOMAS DE REFERÊNCIA 43
TABELA 4 - PROTEÍNAS PREDITAS E ANOTAÇÕES DE FAMÍLIAS GÊNICAS LIGADAS À PATOGENICIDADE 46
TABELA 5 - PREDIÇÃO DE GENES NÃO CODIFICANTES 48
TABELA 6 - PSEUDOGENES 50
TABELA 7 - PROTEÍNAS ANOTADAS POR VALOR DE COBERTURA 67
TABELA 8 - COMPARATIVO ENTRE OS EXTEMOS DE COBERTURA 67
TABELA 9 - COMPARATIVO DO NÚMERO DE REGIÕES REPETITIVAS ENCONTRADAS NOS GENOMAS MONTADOS E SUAS RESPECTIVAS REFERÊNCIAS 68
TABELA 10 - PROTEÍNAS EM CADA FAMÍLIA GÊNICA LIGADA A PATOGENICIDADE DE *Eimeria* SPP. ENCONTRADAS NOS GENOMAS MONTADOS NESTE ESTUDO 69
TABELA 11 - PROTEÍNAS EM CADA FAMÍLIA GÊNICA LIGADA A PATOGENICIDADE DE *Eimeria* SPP. ENCONTRADAS NOS GENOMAS DE REFERÊNCIA 70

LISTA DE ABREVIATURAS E SIGLAS

DNA	Ácido desoxirribonucleico
DUP	Duplicação
EACV	*E. acervulina*
EBRU	*E. brunetti*
EMIT	*E. mitis*
EMWX	*E. maxima*
ENEC	*E. necatrix*
EPRA	*E. praecox*
EST	Expressed Sequence Tag
ETHE	*E. tenella*
FRAG	Fragmento
GFF	Gene-Finding Format
IMC	Complexo de Membrana Interna
MIC	Proteínas do Micronema
NCBI	National Center for Biotechnology Information
ORF	Open Reading Frame
PACBIO	Pacific Biosciences
PSSD	Pseudogenes Retrotranspostos
RNA	Ácido Ribonucleico
RON	Rhoptry Neck Proteins
ROP	Rhoptry Proteins
SAG	Surface Antigen
SRA	Sequence Read Archive
UTR	Região Não Traduzida
WGS	Whole Genome Sequencing
cDNA	Ácido Desoxirribonucleico Complementar
mRNA	Ácido Ribonucleico Mensageiro
rRNA	Ácido Ribonucleico Ribossomal
snoRNA	Pequeno Ácido Ribonucleico Nucleolar
tRNA	Ácido Ribonucleico Transportador

SUMÁRIO

1 INTRODUÇÃO 13

1.1 Coccidiose aviária 13

1.2 Ciclo de vida 15

1.3 Patogenicidade 17

1.4 Montagem de genoma 19

1.5 Predição e anotação gênica 23

1.6 Pseudogenes 25

2 OBJETIVOS 28

2.1 Objetivo Geral 28

2.2 Objetivos Específicos 28

3 MATERIAL E MÉTODOS 29

3.1 Obtenção dos dados brutos 29

3.2 Limpeza dos dados brutos 30

3.3 Montagem dos genomas 31

3.4 Predição de RNAs e Proteínas 35

3.5 Predição de pseudogenes 38

4 RESULTADOS E DISCUSSÃO 40

4.1 Montagens 40

4.2 Predições gênicas 44

4.3 Pseudogenes 48

5 CONCLUSÃO 51

REFERÊNCIAS 52

APÊNDICE A – COMPARATIVO ENTRE OS MODELOS DE PREDIÇÃO GÊNICA 65

APÊNDICE B – TESTES PARA DEFINIÇÃO DO VALOR DE COBERTURA..... 67

APÊNDICE C – REPETIÇÕES NOS GENOMAS DE *Eimeria* SPP. 68

APÊNDICE D – QUALITATIVO DE PROTEÍNAS DAS CLASSES ROP E SAG..... 69

1 INTRODUÇÃO

1.1 Coccidiose aviária

Os agentes causadores da coccidiose aviária pertencem ao filo Apicomplexa, caracterizados por serem organismos unicelulares e parasitas obrigatórios. A principal característica, que define esse filo, é a presença do complexo apical, que é uma região que contém componentes estruturais e organelas secretoras necessárias para os processos de motilidade do parasito e invasão da célula hospedeira durante o ciclo de vida destes organismos (SEEBER; STEINFELDER, 2016, ŠLAPETA; MORIN-ADELINE, 2011). O filo Apicomplexa compreende vários gêneros de parasitos de interesse econômico e de saúde pública, como *Plasmodium* spp. causador da malária; *Toxoplasma* spp. responsável pela toxoplasmose; *Cryptosporidium* spp. agente responsável pelo desenvolvimento de criptosporidíase; *Theileria* spp. que infecta ruminantes e causa mais de 300 milhões por ano em prejuízo na África subsaariana (SEEBER; STEINFELDER, 2016; MORRISON, 2009).

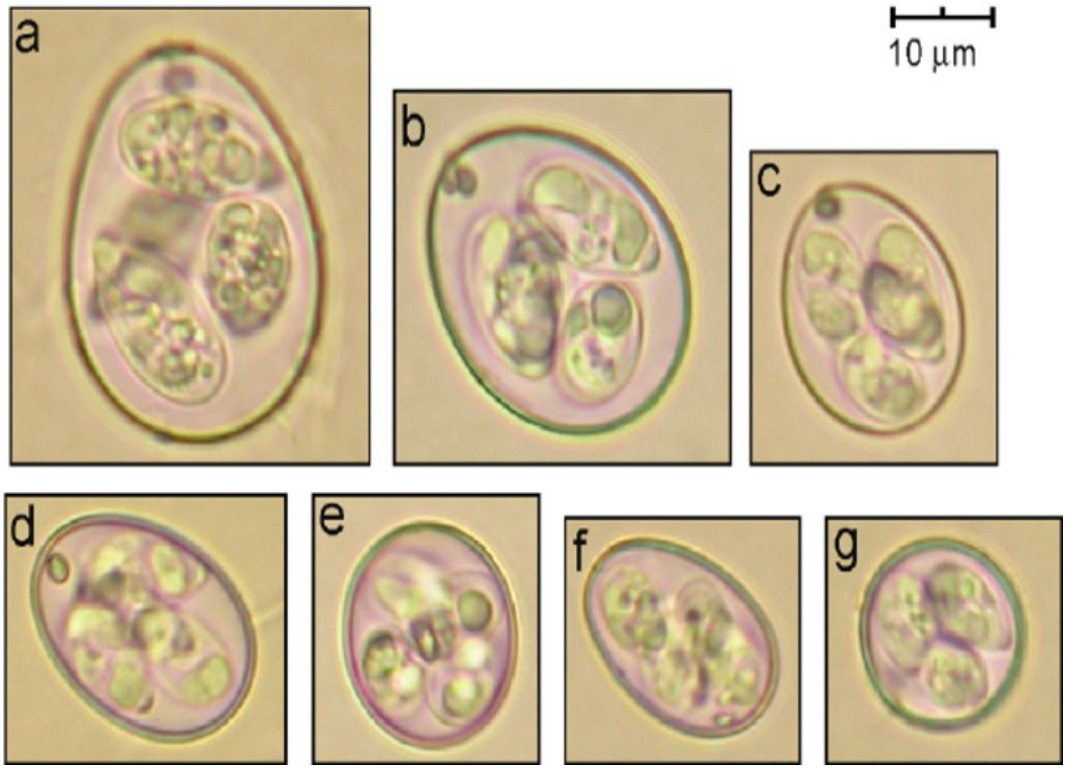
Os apicomplexa englobam ainda: (i) *Neospora caninum*, parasito de ciclo de vida complexo, que infecta o intestino de cães e bovinos, podendo infectar cabras, ovelhas, cavalos e até humanos (DUBEY, 2003). (ii) *Baebesia* spp., transmitida por carrapatos, primariamente descrita como parasita de bovinos e cachorros, mas com relatos de infecção em humanos; (iii) *Sarcocystis neurona*, agente causador de zoonose equina, com hospedeiros intermediários variados de vida selvagem (DUBEY *et al*., 2001; SOLANO-GALLEGO *et al*., 2016).

Por fim, fazem parte deste filo diferentes espécies do gênero *Eimeria* (com aproximadamente 1200 espécies descritas), que infectam o intestino de mamíferos, répteis, anfíbios, peixes e aves, causando zoonoses conhecidas como coccidioses. (CHAPMAN *et al*., 2013). Por este motivo, esses parasitos são comumente chamados de coccídios. A coccidiose aviária é uma doença causada por sete espécies de coccídios do gênero *Eimeria* (*E. acervulina, E. brunetti, E. maxima, E. mitis, E. necatrix, E. praecox* e *E. tenella*) - Figura 1 - que causam severos danos ao epitélio da mucosa intestinal das aves, - demonstrado na Figura 2 - acarretando em: subnutrição, perda de peso e altos índices de mortalidade, que implicam em déficits econômicos por parte da indústria (CHAPMAN, 2014).

Em 1998 Stevens avaliou que um total de 1,5 bilhão de dólares foram gastos combatendo a coccidiose aviária nos Estados Unidos. O valor se aproxima da

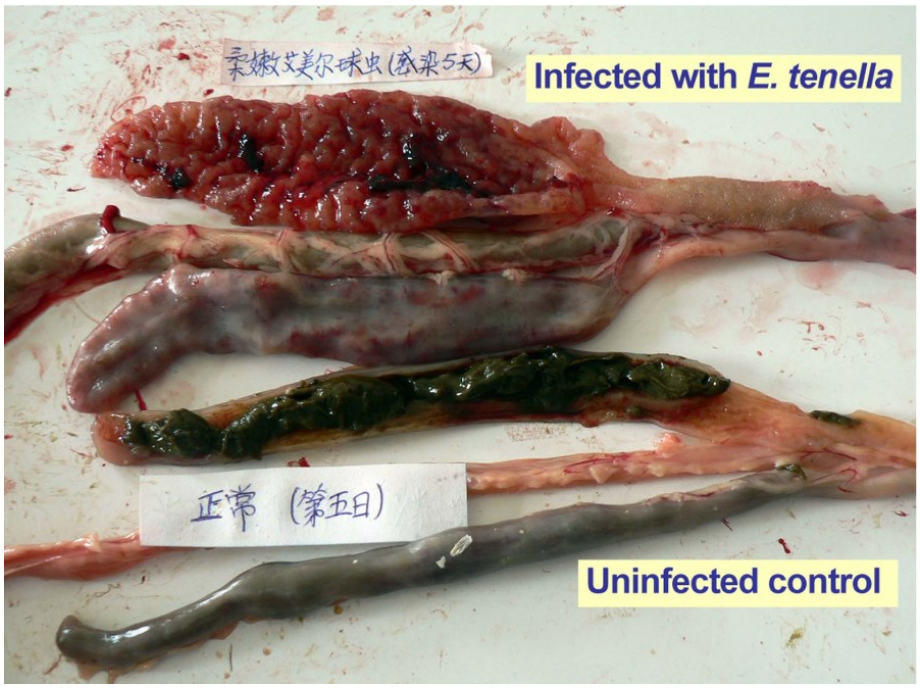
estimativa feita por Bera e colaboradores em 2010, de 1,4 bilhão de dólares para os anos de 2003 e 2004 apenas para a produção indiana. Segundo o mesmo autor, este gasto corresponde à 95% de todo o prejuízo na criação de frangos do mesmo ano. Estes custos englobam profilaxia, tratamento e perda de produção. No Brasil, os dados são de 2002, em que, segundo Pinheiro (2002, apud RAMA, 2016), a doença foi responsável por prejuízos da ordem de 19 milhões de dólares, dos quais 62,4% foram ocasionados por perda de produção e 37,6% com consumo adicional de ração. Recentemente, em 2020, Blake e colaboradores estimaram o gasto com tratamento e perda produtiva em decorrência da coccidiose aviária em 2016. Segundo os autores, o gasto global para o ano foi de aproximadamente 10,36 bilhões de euros, dos quais, o Brasil, segundo maior produtor no ano, teve um gasto estimado de 958 milhões de euros. Devido ao impacto econômico causado pela doença, estudos buscam entender estes patógenos e desenvolver metodologias de diagnóstico e vacinação mais eficazes.

Figura 1 - Oocistos esporulados de *Eimeria* spp.



Legenda: a - *E. maxima*; b - *E. brunetti*; c - *E. tenella*; d - *E. necatrix*; e - *E. praecox*; f - *E. acervulina*; g - *E. mitis*; Fonte: Castañón et al. (2007)

Figura 2 - Comparação de intestino infectado por *E. tenella* com o de uma ave saudável

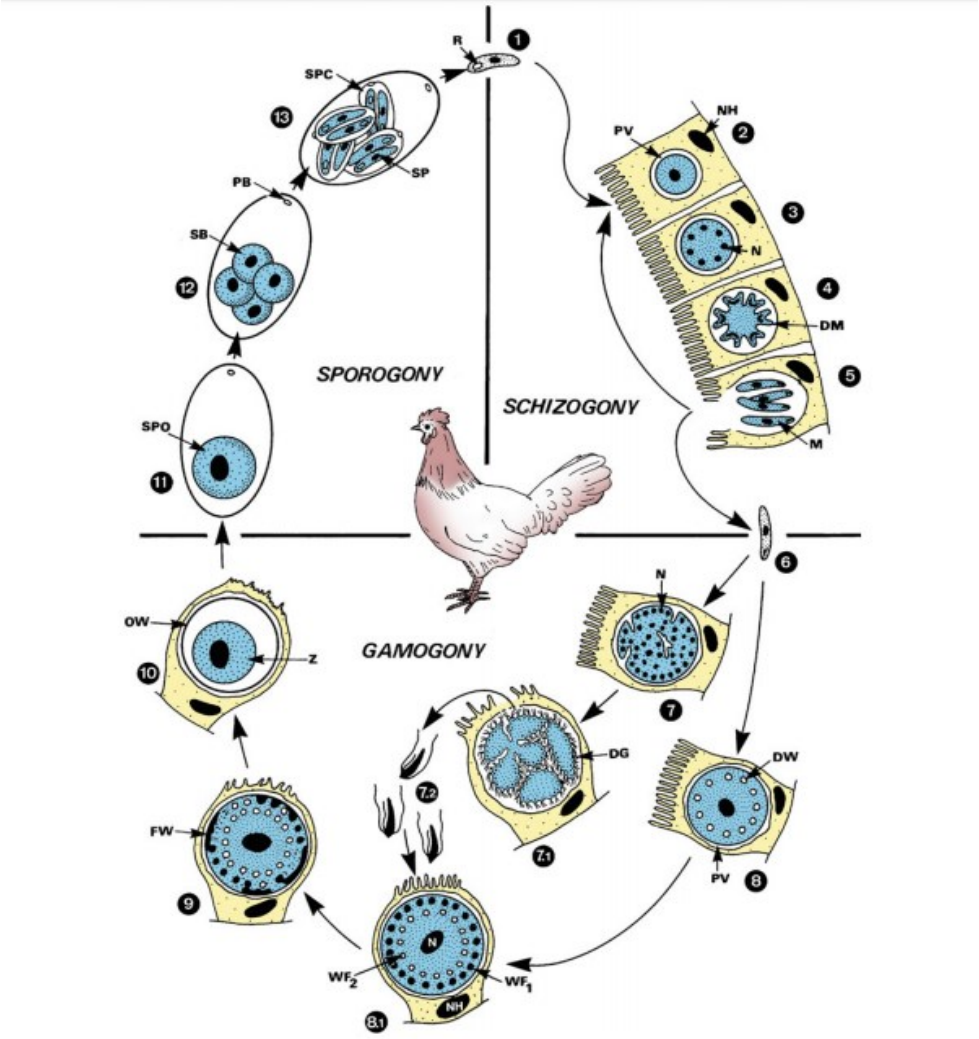


Legenda: A parte superior mostra um intestino, em sua porção cecal, infectado por *E. tenella*, após o 5 dia de infecção, com número estimado de 105 oocistos. A parte inferior da imagem mostra um intestino não infectado, para efeito de comparação. Fonte: Guo et al. (2013).

1.2 Ciclo de vida

O ciclo do parasito ocorre através da ingestão dos oocistos esporulados do parasito provenientes de um local contaminado, por parte das aves (Figura 3). A ação mecânica da ingestão é responsável pela quebra da membrana celular do oocisto e liberação dos esporocistos. A passagem pelo trato intestinal, mais especificamente a interação com o suco pancreático, excita o esporocisto que libera vários esporozoítos, estes movem-se em direção ao tecido epitelial da porção duodenal do intestino, através de movimentos ondulatórios e com o auxílio de um suco proteico secretado pelo esporocisto. Dentro do tecido, os esporozoítos crescem alimentando-se das células hospedeiras, esta fase é denominada trofozóite e causa dano gradual à célula do epitélio intestinal, proporcional ao crescimento do parasito. Em seu tamanho máximo (10 a 12 µm), o parasito causa a atrofia da célula e induz a replicação da mesma de forma a aumentar o número de parasitos (FANTHAM, 1910; CHAPMAN, 2003; CHAPMAN, 2014).

Figura 3 - ciclo de vida de *Eimeria* spp.



Legenda: Fases do ciclo de vida de Eimeria spp. que infectam galináceos. - Após a ingestão do oocisto esporulado (1), os esporozoítos eclodem dos esporocistos e penetram no intestino (2-6). Há a formação dos esquizontes (3), que produzem os merozoítos (DM e M). Os merozoítos eclodem (6) e podem formar macrogametócitos que formam macrogametas (8 e 8.1); ou formar microgametócitos que formam microgametas (7 e 7.1). A fertilização dos gametas forma o zigoto (9) e os oocistos não esporulados são liberados à luz do intestino (10). No ambiente, pode ocorrer a esporulação. A ingestão de um oocisto esporulado recomeça o ciclo. Fonte: MEHLHORN, 2015

A próxima fase, denominada esquizonte, ocorre a partir da fragmentação do núcleo central do parasito e direcionamento desses fragmentos para a extremidade da célula mãe. Há o deslocamento uniforme de citoplasma para o entorno desses fragmentos de núcleo então formados e aglomeração em forma vermiforme; essas formas filhas são denominadas “merozoítos” e são envoltas pela membrana celular da célula mãe. O crescente número de merozoítos induz a secção desta membrana e causa a liberação dos merozoítos para o trato intestinal do animal. Os merozoítos livres passam a se deslocar em direção às células sadias, invadem-nas e repetem o

processo de trofozóite, aumentando o número de parasitos. Assim a infecção se distribui ao longo do intestino do animal. O suporte nutricional que o animal fornece ao parasito é comprometido pela superpopulação destes (efeito notado pela perda de peso por parte do animal). A percepção dessa condição induz o parasito, na fase esquizonte, a acumular os nutrientes disponíveis, crescer e preparar-se para a produção de gametas a fim de perpetuar a espécie (FANTHAM, 1910; CHAPMAN, 2014).

A produção de gametas se dá pela diferenciação dos esquizontes maturados de acordo com o grau de nutrientes acumulado por eles. A diferenciação dos macrogametócitos (que dão origem a um único gameta feminino) ocorre devido ao grande suporte nutricional existente. Os microgametócitos (que dão origem a vários gametas masculinos), possuem menor aporte nutricional. Os macrogametócitos ficam dentro do tecido epitelial do intestino até a completa maturação, quando o tecido é rompido o macrogameta é deslocado de forma a fazer contato com a parede interna do intestino, ainda preso no mesmo. Nessa fase os macrogametas são atacados pelos microgametas que rompem a membrana do microgametócito e se deslocam, através de “movimentos de chicote” e com o auxílio do flagelo, em direção ao gameta feminino para a fecundação (FANTHAM, 1910; CHAPMAN, 2003; CHAPMAN, 2014).

Após a fecundação, há a formação do zigoto que depois da fecundação toma, na maioria dos casos, a forma oval (tamanho e forma dos oocistos possuem grande variação) e adquire uma membrana celular protetora, sendo chamado de oocisto. Eventualmente, os oocistos se desprendem da parede intestinal e ficam livres para serem secretados junto às fezes do animal. Em condições favoráveis, de temperatura, calor e umidade, o oocisto pode ser esporulado; processo que consiste na divisão do núcleo do oocisto e formação de 4 esporocistos individualizados pela massa celular. O oocisto esporulado, quando ingerido por outro animal, sofrerá excitação pelo suco pancreático, liberação dos esporozoítos e início de uma nova infecção (FANTHAM, 1910; CHAPMAN, 2014).

1.3 Patogenicidade

Os apicomplexa possuem mecanismos que permitem invasão e motilidade dentro das células hospedeiras. Estes dois mecanismos são modulados por secreções de duas organelas dos apicomplexa, a roptria e o micronema. O sucesso

destes mecanismos está diretamente ligado à patogenicidade destes agentes infecciosos. O primeiro processo envolvido na patogenicidade dos parasitos é o reconhecimento e adesão à célula do hospedeiro. Antígenos de superfície (SAG) ancorados na membrana do parasito fazem este reconhecimento e permitem que o processo de infecção se inicie. Há diversas classes desses genes, que são descritos como um dos fatores essenciais para a infecciosidade e patogenicidade dos Apicomplexa (CHOW *et al.*, 2011; SOUZA, BELFORD JR., 2014). Durante o processo de invasão os parasitos, em contato com células hospedeiras, induzem a formação de junções móveis, estruturas que realizam um processo semelhante a endocitose (BRADLEY; SIBLEY, 2007). As junções são induzidas pela secreção de proteínas da roptria e micronema conhecidas como RON (do inglês, *rhoptry neck proteins*) e MIC (do inglês, *microneme proteins*) (SHEN; SIBLEY, 2012). Estas proteínas são responsáveis por forçar a invaginação da membrana celular do hospedeiro de forma a projetar o parasito para dentro da célula. Este mecanismo forma um vacúolo que engloba o parasito, denominado vacúolo parasitóforo. As proteínas do micronema e da roptria interagem com a membrana e "filtram" proteínas transmembranas da célula hospedeira que possam interagir com o parasito, formando uma barreira (BAUM *et al.*, 2006; BRADLEY; SIBLEY, 2007).

Uma vez internalizado, o parasito se movimenta por um mecanismo molecular conhecido como motilidade dependente de actina. Os apicomplexa possuem uma membrana interna à membrana celular, denominado complexo de membrana interna - IMC (SOUZA, BELFORD JR., 2014). Neste complexo está ancorado o motor de miosina. Na membrana do parasito estão proteínas adesinas (secretadas pelo micronema) que se associam com receptores na célula hospedeira. Filamentos de F-actina se polimerizam entre o motor de miosina e são ancorados nas proteínas adesinas, em sua porção interna (BAUM *et al.*, 2006). Por gasto ativo de energia, a força motriz do motor de miosina desloca os filamentos de actina e realiza o movimento. Ciclos de clivagem desassociam os filamentos de actina das adesinas, despolimerizando-os e permitindo novos ciclos de polimerização, adesão e movimento (BAUM *et al.*, 2006). Outra classe de proteínas da roptria, ROP (do inglês, *rhoptry proteins*), são secretadas dentro do vacúolo parasitóforo, ou mesmo no citoplasma da célula hospedeira, são proteínas cinases que alteram a dinâmica da célula parasitada e provocam uma série de alterações no metabolismo celular (OAKES *et al.*, 2013).

As secreções destas organelas são essenciais para o sucesso de invasão e infecção do parasito. Proteínas ROP foram descritas como essenciais para que parasitos do gênero *Toxoplasma* associem-se com as mitocôndrias da célula hospedeira, garantindo o aporte energético do parasito (SOUZA, BELFORD JR., 2014).

Desta forma, destaca-se as famílias gênicas RON - essencial para infecciosidade dos parasitos - ROP e SAG fatores que modulam a patogenicidade dos mesmos.

1.4 Montagem de genoma

O processo de sequenciamento de DNA envolve a fragmentação do DNA em vários pedaços, os quais servem como base para o sequenciador, gerando fragmentos de leitura denominados *reads*. No sequenciamento das sete espécies de *Eimeria* estudadas aqui foram utilizados sequenciadores do tipo Illumina, que geram fragmentos em torno de 150 pares de base (YE *et al*., 2012). Os *reads* podem ser gerados únicos (*single end*), pareados (*paired end*), ou pareados com grandes *gaps* (*mate pair*). Em todas as metodologias são indexados adaptadores, que ligam o fragmento de DNA ao sequenciador. A metodologia *single end*, apresenta um adaptador que sequencia uma fita de DNA em um sentido (ILLUMINA, 2010). Os *paired end* apresentam 2 adaptadores ligados no fragmento de DNA e sequenciam um em direção ao outro, com uma pequena porção entre as sequências que é perdida - denominado tamanho de inserção. O *mate pair* apresenta o mesmo princípio do *paired end*, apenas com uma etapa adicional que estende o tamanho do fragmento e por consequência o tamanho de inserção (ILLUMINA, 2010). Os numerosos fragmentos gerados precisam ser remontados por métodos computacionais, de forma a “reconstruir” o genoma, ou seja, *reads* precisam ser alinhados e estruturados na ordem correta. Modelos de grafos (algoritmos baseados em polinômios de probabilidades) são a base para os programas de montagem para dados de sequenciamento e resolvem problemas como: colocar todos os fragmentos em suas devidas direções, tais como no organismo vivo (5’ para 3’ ou 3’ para 5’); identificar *reads* únicos; fazer o alongamento destes com base em *reads* que sobrepõem estas sequências; e garantir a confiabilidade de cada base sequenciada, com base na cobertura daquela determinada sequência (YE *et al*., 2012). Conforme os *reads* vão

sendo agrupados, estes passam a ser chamados de *contigs*, e continuando a escala de grandeza, o agrupamento de *contigs* tornam-se *scaffolds*, e o conjunto destes caracteriza uma sequência que representa o genoma. A máxima redução se dá quando o número de *scaffolds* é igual ao número cromossômico da espécie (WAJID; SERPEDIN, 2012). Atualmente o número de *scaffolds* para cada espécie de *Eimeria* causadora da coccidiose aviária é muito grande, como mostra a Tabela 1, em comparação ao número de cromossomos de *E. tenella*, *E. maxima* e *E. mitis* que possuem 14 cromossomos (CACHO *et al*., 2005; REID *et al*., 2014).

O grande número de *scaffolds* se dá ao fato de o sequenciamento dessas espécies ter sido realizado apenas uma vez, por um único grupo de pesquisa. Para efeito de comparação, o genoma humano foi sequenciado 210 vezes para chegar ao nível cromossômico (Dados: NCBI/GENOME) (REID *et al*., 2014).

Tabela 1 - Genomas de referência de *Eimeria* spp.

Espécie	**Contiguidade do genoma**	**Tamanho (Milhões de bases)**	**Proteínas identificadas**
E. acervulina	Scaffolds (3.415)	45,83 Mb	6.867
E. brunetti	Scaffolds (8.575)	66,89 Mb	8.711
E. mitis	Scaffolds (15.978) *	72,24 Mb	10.077
E. maxima	Scaffolds (3.564) *	45,97 Mb	6.057
E. necatrix	Scaffolds (3.707)	55,00 Mb	8.609
E. praecox	Scaffolds (21.348)	60,08 Mb	7.635
E. tenella	Scaffolds (4.695) *	51,07 Mb	8.609

*Número de cromossomos definidos (1 a 14) Fonte: NCBI/Genome

A montagem não é realizada com os dados brutos fornecidos pelo sequenciador. Antes disso, há um processo fundamental de controle de qualidade. O controle é realizado observando a acurácia de cada base sequenciada em cada *read*. Isto é mensurado pelo que se convencionou como “qualidade Phred”, devido ao programa, de mesmo nome, criado durante o sequenciamento do genoma humano, que calculava a taxa de erro das bases sequenciadas naquele grandioso projeto (LIAO; SATTEN; HU, 2017). A qualidade Phred é dada pela equação 1.

$$Q = -10 \log^{10} P \quad (1)$$

Em que, Q é a qualidade e P é a diferença entre a probabilidade observada e verdadeira de cada base. Assim, uma qualidade Phred de 30, por exemplo, indica 0.1% de chance de erro daquela base em questão. Com base na qualidade Phred, elimina-se bases de baixa qualidade (Q<20), que são substituídas por “N” e podem ser corrigidas durante a montagem, com base na cobertura, ou, no processo de *gap filling* (LIAO; SATTEN; HU, 2017). A qualidade phred dos *reads* pode ser verificada por meio dos programas FastQC (ANDREWS, 2010), ou MultiQC (EWELS *et al.*, 2016). Com base na qualidade inicial, os *reads* são limpos. Neste processo, *reads* que sofrem muitas substituições também são eliminados, tendo em vista que o tamanho de bases restante é reduzido (LIAO; SATTEN; HU, 2017). Um artefato da técnica permite que haja a eliminação de apenas um dos *paired end* ou *mate pair*. A quantidade de *reads* pareados é importante, pois programas que montam genomas com este tipo de dado costumam utilizar os *unpaired reads* apenas para refinar e melhorar a montagem. Desta forma, o uso dos dados não pareados costuma ser um parâmetro opcional. Um exemplo de programa utilizado para limpar os *reads* é o Trimmomatic (BOLGER: LOHSE: USADEL ,2014).

A montagem pode ser realizada de duas formas: *de novo* ou com referência, também denominado método comparativo. Uma montagem *de novo* parte do princípio que serão utilizados para a montagem apenas os dados sequenciados da amostra, sem nenhum modelo fornecido previamente para guiar o programa. Esta metodologia apresenta algumas dificuldades, como identificação de sequências repetitivas dentro do genoma, assim como sua posição (POP, 2004; WAJID; SERPEDIN, 2012; YE *et al.*, 2012). A montagem com referência, como o nome diz, usa uma sequência como referência, e assim, o programa possui um padrão para reconstruir o genoma com os dados sequenciados. A referência pode ser outro genoma, ou sequências nucleotídicas mais longas (denominadas de *long reads*), obtidos por métodos de sequenciamento de terceira geração, em que se obtém fragmentos maiores que 10 mil pares de bases. Um exemplo de sequenciador que gera *long reads* são os do tipo PacBio (HEBERT *et al.*, 2018; WAJID; SERPEDIN, 2012). O método comparativo possui barreiras como a identificação de polimorfismos, que podem ser confundidos com erros de sequenciamento, assim como regiões de DNA divergente, em que o genoma de referência e os *reads* utilizados apresentam mutações que não permitem

que as sequências sejam alinhadas (POP, 2004; WAJID; SERPEDIN, 2012; YE *et al.*, 2012).

Quando se utiliza *long* e *short reads*, da mesma amostra, obtém-se uma montagem híbrida. Assim, emprega-se os *reads* obtidos em outros sequenciadores de maior acurácia – mas de tamanho menor – e *long reads* como referência. Não se utiliza apenas os *long reads* para o processo de montagem pela alta taxa de erro inerente da técnica de sequenciamento (cerca de 13% em tecnologias disponíveis) (WAJID; SERPEDIN, 2012; WENGER *et al*., 2019).

Há programas que realizam montagens *de novo* ou com referência, como o SPAdes (BANKEVICH *et al.*, 2012) e Velvet (ZERBINO; BIRNEY, 2008) e outros que realizam apenas a montagem *de novo* como SOAP (LUO *et al.*, 2012) e ALLPATHS (BUTLER *et al.*, 2008).

Um dos processos finais de montagem de um genoma é o fechamento (*gap filling*) que visa corrigir falhas nos *scaffolds*, hiatos (*gaps*) entre sequências contíguas e bases desconhecidas (marcadas como “N”) geradas durante o processo de sequenciamento ou controle de qualidade. O *gap filling* fecha estes espaços e revisa os N’s baseando-se nos dados utilizados durante a etapa de montagem. Assim, pode-se compensar a baixa qualidade com alta cobertura. Durante esta etapa, pode-se realizar a extensão dos *scaffolds* interpolando *reads* nas extremidades dos *scaffolds* de forma a alongar a sequência, quando possível. Estes métodos não são completamente eficazes, no sentido de que a montagem é melhorada, mas ainda assim, espaços podem ficar no genoma (KAMMONEN *et al.*, 2019).

A qualidade da montagem é verificada comparando métricas já estabelecidas em genomas de referência e, quando não disponíveis, em organismos filogeneticamente aparentados. Essas métricas são: tamanho de genoma, que define quantos pares de base há no genoma; conteúdo GC, que indica a proporção, em porcentagem do genoma, em que estas bases aparecem; N50, que corresponde ao tamanho do menor *contig* ou *scaffold*, após ordená-los de forma que 50% do genoma esteja englobado nos *contigs* ou *scaffolds* acima deste tamanho; L50, que corresponde ao número de *contigs* ou *scaffolds* necessários para atingir o N50; Número de gaps, que representa hiatos preenchidos com bases desconhecidas nas sequências; Cobertura, que indica o quantas vezes o tamanho do genoma foi sequenciado (esta métrica é calculada com a quantidade e tamanho dos *reads* e tamanho do genoma); E profundidade que indica quantos *reads* definiram a mesma

base de cada nucleotídeo nos *scaffolds* formados, em suma, a confiabilidade de cada base nos *scaffolds* (ANGEL *et al*., 2018; CASTRO; NG, 2017). Algumas dessas propriedades da montagem podem ser obtidas pelos programas QUAST v5.02 (GUREVICH *et al*. 2013) ou CAGE (SALZBERG *et al*. 2012).

Nenhuma dessas métricas estabelece qualidade, mas apenas direcionam para um conhecimento pré-estabelecido acerca dos genomas que se deseja montar.

1.5 Predição e anotação gênica

Além do processo de montagem, que foca em dizer “como o DNA está”, há o processo de predição e anotação gênica voltado em dizer “o que o DNA contém”. Assim, com o genoma montado, a predição gênica é o processo pelo qual busca-se padrões, que caracterizem marcadores e genes codificantes (STEIN, 2001; WANG; CHEN; LI, 2004). Este processo visa estabelecer possíveis sequências de DNA que podem ser transcritas (para genes não codificantes, como rRNAs e tRNAs) e traduzidas em proteínas (para genes codificantes). Para tal, busca-se encontrar padrões que descrevem janelas abertas de leitura (do inglês, *Open Reading Frames* – ORFs), que são sequências contíguas com características como *start* e *stop* códon (STEIN, 2001; WANG; CHEN; LI, 2004). Porém, nem toda ORF é um gene, por vezes há ORFs sobrepostas, e falsos positivos. Assim, há outras características que sinalizam para a presença de um gene, como alto conteúdo GC; regiões promotoras que antecedem o gene, altamente conservadas (TATA box para eucariotos e Prinbnow para procariotos); e sinalização para poliadeninação (GHORBANI; KARIMI, 2015). A predição é uma tarefa mais fácil em organismos procariotos, devido à alta densidade de genes, e ausência de sequências que não são traduzidas em proteínas, os íntrons. Em eucariotos, a predição é mais difícil, devido a necessidade de identificação de regiões não traduzidas (UTRs), que são importantes para promover a expressão do gene; identificação de íntrons; e mecanismos de modificação de transcritos como splicing alternativo (STEIN, 2001; WANG; CHEN; LI, 2004).

Os programas de predição utilizam duas metodologias para predizer sequências gênicas: busca por sequências similares e *ab initio* (do início) (ABRIL; CASTELLANO, 2019; STEIN, 2001; WANG; CHEN; LI, 2004). A busca por sequências similares, como o próprio nome diz, visa encontrar regiões com base em estruturas pré-definidas, como etiquetas de expressão (ESTs), proteínas, ou outro tipo de

informação gênica. Deste modo, sequências conservadas evolutivamente são muito bem preditas, se há suficiente informação de respaldo para os programas. Esta metodologia utiliza alinhamento local como base das buscas, de forma que os principais programas utilizam a mesma metodologia do BLAST (ABRIL; CASTELLANO, 2019; STATES; GISH, 1994; STEIN, 2001; WANG; CHEN; LI, 2004). A predição *ab initio* utiliza a estrutura genética de um modelo para predizer sequências que obedecem a este padrão. Com base nesta estrutura é calculada a frequência de utilização de códons, motivos, pontos de ramificação, padrão de sinalização de íntrons, regiões polipirimídicas, que sinalizam para ligação de proteínas e splicing alternativo (ABRIL; CASTELLANO, 2019). Os programas utilizam metodologias de aprendizado de máquina e redes neurais para reconhecer e calcular a probabilidade da existência de genes. Os principais programas utilizados para predição *ab initio* são o Augustus (STANKE; MORGENSTERN, 2005), Genemark (BESEMER; BORODOVSKY, 2005) e Prodigal (HYATT *et al*. 2010).

Isto posto, as duas metodologias podem ser utilizadas em conjunto, de forma que a predição *ab initio* pode ser corroborada por evidências de bancada, como dados de sequenciamento de RNA (RNA-seq), por exemplo, com alta similaridade (ABRIL; CASTELLANO, 2019; STEIN, 2001; WANG; CHEN; LI, 2004).

A partir da identificação dos genes codificantes, busca-se encontrar a funcionalidade dos mesmos, que reflete o RNA ou a proteína gerada, suas variantes de expressão, marcadores de localização celular e a função que desempenham (STEIN, 2001). Para inferir funcionalidade, uma análise de similaridade é realizada com um conjunto de genes com funções previamente conhecidas e descritas, de forma a identificar o papel dos genes preditos (ANGEL *et al*., 2018). Na anotação funcional, buscam-se domínios proteicos e sítios ativos, famílias e função biológica de proteínas, que podem ser obtidas através da organização hierárquica da Gene Ontology (ASHBURNER *et al*. 2000; CONSORTIUM, 2018). Realiza-se anotação da ontologia dos genes, que descreve localização e envolvimento em vias metabólicas, através de busca em bancos de dados como o KEEG (KANEHISA, 2000). Realiza-se anotação da função catalítica ou estrutural, que efetivamente nomeia o gene predito, através de uma análise de similaridade com bancos de dados de proteínas curadas manualmente, como o SwissProt (UNIPROT, 2019) ou automaticamente como o TrEMBL (UNIPROT, 2019). Havendo alta similaridade entre o gene predito e o gene depositado no banco de dados a anotação é transferida e o gene anotado. Isso pode

ser realizado por ferramentas como InterProScan (MULDER; APWEILER, 2007) e MAKER (HOLT; YANDELL, 2011).

Caso não haja similaridade com nenhuma sequência conhecida, o gene é anotado como hipotético, ou seja, possui padrões que caracterizam um gene codificante, mas não há função conhecida. Proteínas hipotéticas podem ser um artefato da técnica, falsos positivos - cerca de 5% em algoritmos de alta performance (STEIN, 2001) - ou genes existentes, por vezes com evidência de expressão, mas que se desconhece a função. Neste ponto, a bioinformática contrasta com o trabalho em bancada, ora guiando estudos que busquem transcritos de proteínas hipotéticas, ora sendo alimentada por tais evidências.

A predição de genes não codificantes é realizada por similaridade, para rRNAs; e *ab initio* para tRNAs, através da predição de estruturas conservadas, e com potencial formação de *hairpins* (STEIN, 2001). O tRNAscan (CHAN; LOWE, 2019) é um exemplo de programa utilizado para predição de tRNAs e o RNAmmer (LAGESEN; *et al*. 2007) para predição de rRNAs. A predição de snRNAs e snoRNAs não é comum e é utilizada principalmente em leveduras (STEIN, 2001). Para estes tipos de RNAs, há um módulo de anotação na Pipeline de Anotação de Eucariotos do GenBank (THIBAUD-NISSEN; SOUVOROV; MURPHY *et al*., 2013), que utiliza busca de similaridade com RNAs descritos no banco de dados Rfam (KALVARI; *et al*. 2018).

Há ainda a identificação de regiões repetitivas, que abrangem grande parte dos genomas (47% do genoma humano), e podem servir como importantes marcadores para metodologias de diagnóstico. Isto porque o tamanho e localização destas repetições no genoma pode variar de espécie para espécie e, inclusive, de indivíduo para indivíduo (STEIN, 2001; YANDELL; ENCE, 2012;). RepeatMasker (SMIT; HUBLEY; GREEN, 2015) e RepeatScout (PRICE; JONES; PEVZNER, 2005) são exemplos de programas utilizados na identificação de estruturas repetitivas.

1.6 Pseudogenes

Durante muito tempo, regiões não codificantes do DNA, que no genoma humano correspondem a 98% das sequências nucleotídicas, eram interpretadas como “DNA lixo”. Hoje, sabemos que estas sequências correspondem a elementos transponíveis, duplicações, repetições, elementos regulatórios e pseudogenes (TUTAR, 2012). Os pseudogenes são genes com modificações em suas sequências

que não os tornam codificantes, mas que podem, em algum momento da linha evolutiva, ter sido. Além disso, eles podem exercer papel regulatório na transcrição de diversos tipos de RNAs e, portanto, regular a ação de genes codificantes (TUTAR, 2012; BALAKIREV; AYALA, 2003). Pseudogenes também podem servir como reservatórios da diversidade genética, permitindo o surgimento de novos genes codificantes, por eventos cumulativos de mutações (BALAKIREV; AYALA, 2003).

Em organismos eucariotos, os pseudogenes são classificados de quatro formas: processados, não processados, unitários e polimórficos. Pseudogenes processados ocorrem por retrotransposição de mRNAs processados, ou seja, o gene é transcrito, sofre processamento, gera um mRNA maduro e este é reincorporado ao DNA por ação da enzima transcriptase reversa. Deste modo, o pseudogene processado é idêntico à sequência nucleotídica codificante dos exons do gene “mãe”, por isso, são identificados como “fragmento” do gene original. Os pseudogenes não processados são aqueles que surgem por duplicação do gene “mãe” e que acumulam mutações que os impedem de serem transcritos (CHEETHAM; FAULKNER; DINGER, 2019). Estas duas classes são as mais comuns. Pseudogenes unitários e polimórficos são duas classes mais raras de classificação. Os unitários correspondem a genes ancestrais que perderam a capacidade de codificar para uma determinada proteína e foram inativados, tornando-se pseudogenes. Os polimórficos representam genes inativados em genomas de referência, mas que são encontrados ativos em outros genomas (CHEETHAM; FAULKNER; DINGER, 2019).

Os pseudogenes podem ser anotados por similaridade, em um programa de predição gênica, como os citados anteriormente, em que se transfere anotação de pseudogenes já identificados. Para predição e anotação de pseudogenes por meio da verificação de troca de quadro de leitura, perda de íntrons ou *stop* códon prematuro são necessários programas específicos para eucariotos, tais como PseudoPipe (ZHANG *et al*., 2006) e PPFINDER (VAN BAREN, 2006).

Organismos do filo Apicomplexa, possuem uma família de genes, denominada de genes SAG, envolvidos no processo de infecção (interação parasito-hospedeiro), sendo que esta classe é destacada nos gêneros *Toxoplasma* e *Eimeria*. Análises do conteúdo genômico das sete espécies de *Eimeria* causadoras da coccidiose aviária, realizada por Reid e colaboradores em 2014, indicam e quantificam a presença de fragmentos de pseudogenes no repertório de genes SAG de *Eimeria* spp.. Assim,

segundo esse grupo de pesquisa, *E. acervulina* apresenta 16 fragmentos; *E. brunetti* 39; *E. maxima* 29; *E. mitis* 128; *E. necatrix* 102; *E. praecox* 20; *E. tenella* 23.

Diante de todo este contexto, a hipótese deste trabalho é que a reavaliação dos dados genômicos das sete espécies de *Eimeria* causadoras de coccidiose aviária possibilitaria uma busca por pseudogenes associados à marcadores de patogenicidade nestes organismos.

2 OBJETIVOS

2.1 Objetivo Geral

Realizar um levantamento da presença de pseudogenes das famílias gênicas ROP e SAG nos genomas das sete espécies de *Eimeria* causadoras da coccidiose aviária.

2.2 Objetivos Específicos

1 Revisitar os dados genômicos públicos de *Eimeria* spp., remontar os genomas e compará-los com os respectivos genomas de referência;
2 Realizar a predição e anotação de genes codificantes e não-codificantes em cada um dos genomas;
3 Efetuar a predição de pseudogenes com base nas proteínas preditas de cada genoma;
4 Identificar os pseudogenes pertencentes às famílias gênicas ROP e SAG nos genomas e relacioná-los com o grau de patogenicidade desses organismos.

3 MATERIAL E MÉTODOS

As análises desenvolvidas neste trabalho foram realizadas no Laboratório de Bioinformática do Departamento de Microbiologia, Imunologia e Parasitologia (MIP), do Centro de Ciências Biológicas (CCB) da Universidade Federal de Santa Catarina (UFSC). A estrutura do laboratório conta também com servidores alocados na Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação (SeTIC) da UFSC para desenvolvimento de plataformas computacionais e demais análises de alta performance.

3.1 Obtenção dos dados brutos

Obteve-se os *reads* das sete espécies de *Eimeria* causadoras da coccidiose aviária pela base de dados do *Sequence Read Archieve* (SRA) do GenBank. Utilizou-se os seguintes dados de sequenciamento com os respectivos códigos de acesso: ERR357127 (*E. acervulina*); ERR357129 (*E. necatrix*); ERR357192 (*E. praecox*); ERR357130 (*E. brunetti*); ERR357128 (*E. maxima*); ERR357191 (*E. mitis*); ERR019305, ERR019306, ERR019307, ERR019308, ERR019309, ERR019310, ERR019311, ERR019312 (*E. tenella* Houghton); ERR296541, ERR296879, (*E. tenella* Nippon NT2). Com exceção de *E. tenella*, que apresenta duas cepas, as outras espécies apresentam apenas um resultado de sequenciamento, aplicado na montagem de seus respectivos genomas de referência. A Tabela 2 exibe um panorama comparativo entre a quantidade de *reads* e os genomas de referência. Obteve-se os dados citados através do programa Fastq-dump v2.8.2 do SRA toolkit (LEINONEN; SUGAWARA; SHUMWAY, 2010). Todos os reads utilizados foram sequenciados pelo método *Illumina Genome Analyzer II* ou *Illumina Genome Analyzer IIx*, ambos geraram *reads* do tipo *paired end*. A média de tamanho dos *reads* é de 75 pares de base para *E. tenella* e 100 pares de base para as demais espécies.

Tabela 2 - Dados brutos por genoma

Espécie/cepa	Quantidade de reads (em milhões)	Tamanho do genoma de referência (em milhões de bases)	Cobertura aproximada
E. acervulina	49,72	45,83	108x
E. brunetti	35,81	66,89	58x
E. praecox	25,54	60,08	42x
E. mitis	130,01	72,24	215x
E. maxima	72,05	45,90	156x
E. necatrix	139,96	55,00	254x
E. tenella Nippon NT2	27,09	51,00*	39x
E. tenella Houghton	68,27	51,00*	100x

** O genoma de referência de* E. tenella *é baseado na cepa de Houghton.*

3.2 Limpeza dos dados brutos

Observou-se a qualidade dos dados brutos obtidos do SRA, individualmente, através do programa FastQC v0.11.8. Valeu-se do programa Trimmomatic v2.0 para remover os *reads* com qualidade Phred abaixo de 25 (`AVGQUAL:25`), assim como as sequências adaptadoras utilizadas no sequenciamento (`ILLUMINACLIP:all_adapters.fa:2:30:10`). Empregou-se parâmetros adicionais de qualidade no início (`LEADING:30`) e fim das sequências (`TRAILING:25`), bem como tamanho mínimo dos *reads* gerados após este processo (`MINLEN:50`).

Depois de limpos, restaram, dos *reads* pareados iniciais: 93,65% de *E. acervulina;* 89,29% de *E. brunetti*; 89,28% de *E. mitis*; 93,08% de *E. maxima*; 88,99% de *E. necatrix*; 90,06% de *E. praecox*. Das cepas de *E. tenella*, restaram, dos reads pareados, para a cepa Houghton: 69,81% do SRA ERR019305; 52,84% do SRA ERR019306; 72,67% do SRA ERR019307, 69,01 do SRA ERR019308; 82,52% do SRA ERR019309; 81,53 do SRA ERR019310; 91,97% do SRA ERR019311; 85,43%

do SRA ERR019312. Para a cepa Nippon NT2, restaram, dos *reads* pareados: 82,62% do SRA ERR296541 e 79,92% do SRA ERR296879.

Utilizou-se dos *reads* pareados limpos de cada espécie para montagem dos respectivos genomas. Agrupou-se os *reads* pareados limpos de *E. tenella* Houghton, valeu-se dos mesmos para realizar a montagem do genoma desta cepa. Repetiu-se o processo para *E. tenella* Nippon NT2.

3.3 Montagem dos genomas

A sequência metodológica de processos realizados para a montagem dos genomas descritos aqui está esquematizada na Figura 4. Realizou-se a montagem da cepa Houghton de *E. tenella* utilizando como referência o genoma de *Toxoplasma gondii* disponível na base de dados de Genomas do GenBank (Acesso: GCA_000006565.2). A montagem sucedeu-se através do programa SPAdes v3.13.0 valeu-se de parâmetros padrão, kmer de tamanho 55 e argumento de correção de erros (`--careful`). Para definição do tamanho de kmer, utilizou-se o programa VelvetOptimizer v2.2.6 (ZERBINO; BIRNEY, 2008). Para *E. tenella* testou-se os kmers ímpares de 33 a 61. Deste modo, estabeleceu-se kmer de 55 para a cepa Houghton e 35 para a cepa Nippon NT2. Para as demais *Eimeria* spp. testou-se os kmers ímpares de 55 a 79. Empregou-se o tamanho de kmer: 77 para *E. praecox*, *E. mitis* e *E. maxima*; 75 para *E. brunetti* e *E. Necatrix*; 67 para *E. acervulina*, para as respectivas montagens. Esta diferença de metodologia para *E. tenella*, se deve ao tamanho dos *reads* obtidos no SRA, que são menores para esta espécie.

Realizou-se a montagem do genoma das outras espécies e da cepa Nippon NT2 de *E. tenella* utilizando o SPAdes v3.13.0, baseando-se no genoma de *E. tenella* Houghton montado primeiramente, após exauridos os processos de montagem e fechamento. Essas montagens utilizaram padrão de cobertura mínima 20 (`--cov-cutoff 20`) e argumento de correção de erros (`--careful`), alterou-se o tamanho de kmer para cada montagem, com base nos resultados do programa VelvetOptimiser. As montagens de *E. mitis*, *E. necatrix* e *E. tenella* Nippon NT2 valeram-se dos *reads* não pareados (`-s`); a montagem de *E. acervulina* foi realizada com padrão de cobertura automático (`--cov-cutoff auto`); *E. preaecox* com padrão de cobertura mínima 15 (`--cov-cutoff 15`) e *E. tenella* com padrão de cobertura mínima 10 (`--cov-cutoff 10`).

Realizou-se a avaliação da qualidade das montagens com auxílio do programa QUAST 5.02. Comparou-se e calibrou-se as métricas geradas com os genomas de referência para definição dos parâmetros de montagem informados anteriormente.

O tamanho de inserção e deleção dos fragmentos foi calculado com o BBmap v38.70 (BUSHNELL, 2014), e padronizado em 300 para as montagens, mesmo tamanho utilizado na montagem dos genomas de referência, informado pelo autor apenas para *E. tenella* (REID, *et al*. 2014). Estabeleceu-se o desvio padrão em 0,2 com base no cálculo do BBmap v38.70 para *E. tenella*. Estes argumentos foram empregados para construção das bibliotecas dos programas SSPACE 3.0 (BOETZER *et al*., 2010) e Gapfiller v1-10 (NADALIN; VEZZI; POLICRITI, 2012).

Utilizou-se o programa SSPACE 3.0 para estender os *scaffolds* geradas e corrigir erros com base na cobertura. Para este processo, executou-se o programa 7 vezes, iterando o resultado de uma rodada como entrada da próxima. Assim, variou-se o tamanho de cobertura necessário para correção dos *scaffolds*. Iniciou-se com 100 (`-k 100`), depois 50, 30, 20, 10, 5 e 5. Definiu-se o parâmetro de extensão dos *contigs* (`-x 1`), remoção de fragmentos pequenos, menores que 500 pares de base (`-z 500`) para as 6 primeiras iterações e `-z 1000` para a última iteração com `-k 5`.

Após este processo, utilizou-se o programa Gapfiller v1-10 para fechar *gaps* entre os *scaffolds* e remover bases não identificadas. Definiu-se número de sobreposição mínimo de *reads* de 20 (`-n 20`); diferença máxima de *gaps* entre os *scaffolds* e as sobreposições de 50 nucleotídeos (`-d 50`); número de nucleotídeos removidos das extremidades dos *gaps* de 10 (`-t 10`). Executou-se o Gapfiller v1-10 com 10 iterações (`-i 10`).

A fim de melhorar a estrutura dos genomas, executou-se o programa SSPACE-LongRead, v1.1 (BOETZER; PIROVANO, 2014), para cada montagem realizada. Utilizou-se como *long reads* os genomas de referência respectivos de cada espécie, disponíveis no GenBank. O programa foi executado em seis iterações, alterando o tamanho de kmer de 100, para 50, 30, 20, 10 e 5.

Após avaliação, os genomas utilizados para o processo final de polimento foram: *E. acervulina*, *E. brunetti* e *E. tenella* Houghton resultantes das seis iterações do SSPACE-LongRead; *E. mitis*, *E. praecox* e *E. tenella* Nippon NT2 resultantes da primeira iteração do SSPACE-LongRead. Para os genomas de *E. necatrix* e *E.*

maxima a metodologia não foi eficaz e, portanto, manteve-se os genomas montados após o processo de fechamento de espaços.

Com esses genomas, executou-se novamente o programa Gapfiller v1-10, com os mesmos parâmetros informados anteriormente. Por fim, uma etapa final de polimento das montagens foi executada através do programa Pilon v1.23 (WALKER *et al*. 2014), valendo-se dos argumentos: `--diploid` e `--defaultqual 25`.

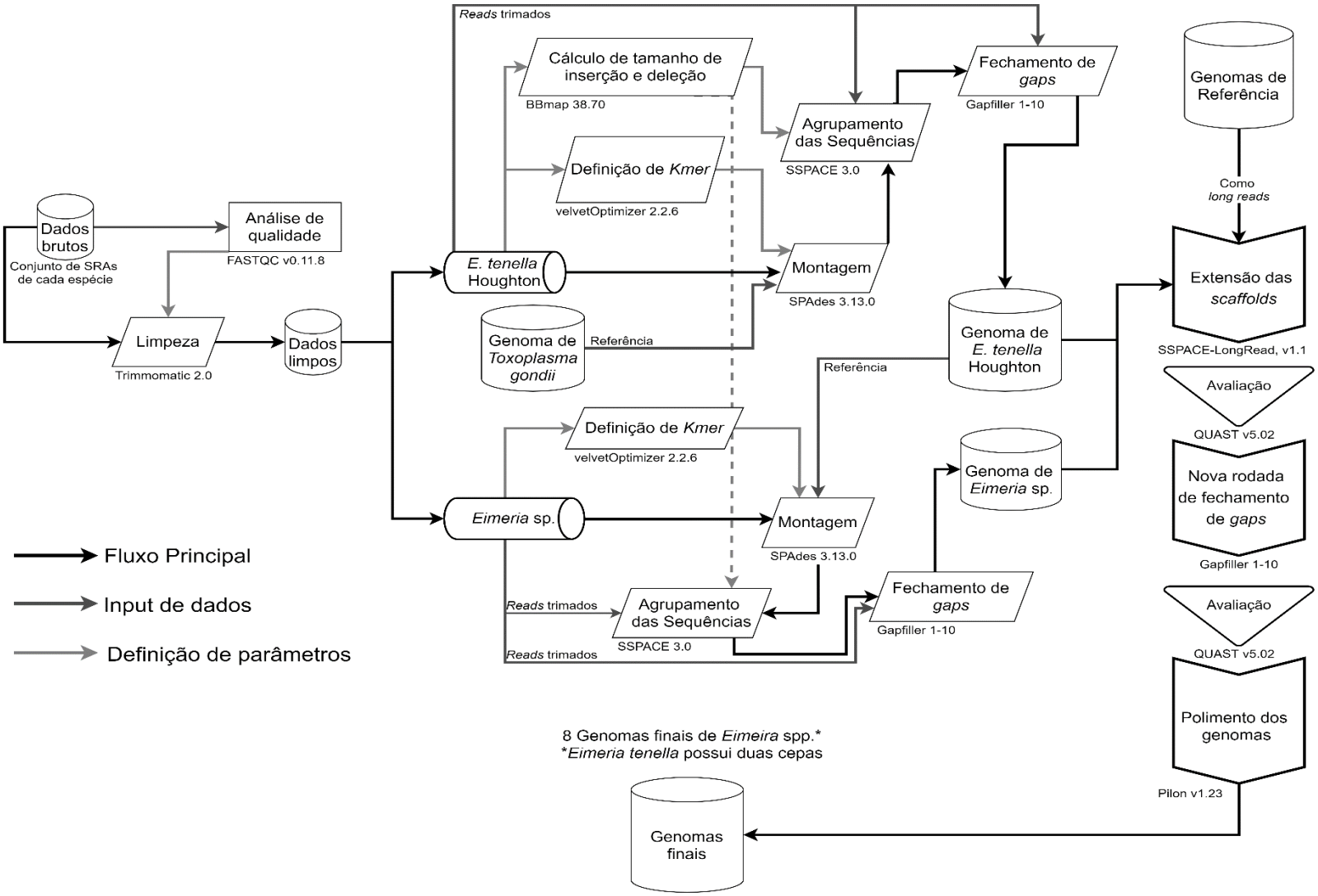
Calculou-se a cobertura dos genomas através da fórmula 2.

`Número de reads * Tamanho dos reads / tamanho do genoma` (2)

Calculou-se a profundidade e realizou-se o mapeamento dos *reads* nos genomas com os programas BWA v0.7.17-r1188 (LI; DURBIN, 2009) e SAMtools v1.7 (LI *et al*. 2009).

Valeu-se do programa RepeatMasker versão 4.1.0 para buscar as regiões repetitivas nos genomas com base em similaridade com as bases de dados Dfam 3.1 e RepBase-20181026, nos quais, restringiu-se a busca a nível de filo (Apicomplexa).

Figura 4 - Metodologia de montagem dos genomas de *Eimeria* spp.



Legenda: Representação gráfica dos processos e programas utilizados para realização das montagens.

3.4 Predição de RNAs e Proteínas

Realizou-se a predição de tRNAs com o programa tRNAscan v2.0, executado com parâmetros padrão. A predição de rRNAs foi realizada pelo programa Infernal v1.2 baseando-se na homologia com dados disponíveis na base de dados do RFAM, obtida no dia 11/05/2020. O banco de dados do RFAM (https://rfam.xfam.org/) foi preparado com o algoritmo `cmpress`. Para a análise com o algoritmo `cmscan` os seguintes argumentos foram utilizados: `-rfam`, que corresponde ao modo rápido do algoritmo; `-Z`, que corresponde ao número de nucleotídeos de cada genoma vezes 2 (fita senso e antisenso), em megabases, calculado separadamente para cada genoma. `--nohmmonly`, para usar modelos de covariância (CM) em conjunto com modelos ocultos de Markov (*Hidden Markov Models* - HMM); `--notrunc`, para evitar resultados truncados; `--clanin`, para usar o arquivo preparado com o cmscan; `--cut_ga`, para utilizar os valores de *bit score* determinados pelo banco de dados curado ao considerar uma sequência homóloga.

A predição de snoRNAs foi realizada pelo programa SnoReport v2.0 (OLIVEIRA *et al*. 2016). O programa foi executado individualmente para cada *scaffold* de cada um dos genomas montados, com configurações padrão e, ao final, agrupados os resultados de cada genoma. Validou-se as predições de tRNAs, rRNAs e snoRNAs com sequências obtidas da base de dados RNACentral (disponível em: https://rnacentral.org/, acessado no dia 16/05/2020). Considerou-se uma predição válida caso houvesse identidade e cobertura maior ou igual a 90 com alguma sequência depositada no banco de dados, sendo o processo realizado por meio do algoritmo do BLASTn. Manteve-se todas as predições de tRNAs e rRNAs. Para snoRNAs manteve-se apenas as predições com alta similaridade com as sequências depositadas no RNACentral. Para as comparações de tRNAs, utilizou-se 121 de *E. acervulina*; 177 de *E. brunetti*; 130 de E. necatrix; 230 de *E. praecox*; 180 de *E. mitis*; 163 de *E. maxima* e 206 de *E. tenella* oriundas do RNACentral. Para as comparações de rRNAs, utilizou-se 93 sequências de *E. acervulina*; 106 de *E. brunetti*; 62 de *E. necatrix*; 95 de *E. praecox*; 129 de *E. mitis*; 200 de *E. maxima* e 210 de *E. tenella* obtidas do RNACentral. Para as comparações de snoRNAs, utilizou-se 9 sequências de *E. acervulina*; 20 de *E. brunetti*; 12 de *E. maxima*; 28 de *E. mitis*; 13 de *E. necatrix*; 22 de *E. praecox* e 8 de *E. tenella* obtidas do RNACentral.

Para a predição de proteínas, testaram-se diversas abordagens. Utilizou-se o programa Augustus v3.3.3, GeneMark-ES Suite v4.57_lic (LOMSADZE, 2005) e BRAKER v2.1.5 (GREMME; 2013; HOFF *et al*., 2019; LOMSADZE; BORODOVSKY; STANKE, 2019). Criaram-se dois modelos de predição com o Augustus, um com o Genemark e um com o BRAKER. O modelo do Genemark foi executado com configurações padrão e sem utilização de evidências externas, visto que é um algoritmo que se auto-treina e, dessa forma, foi criado um modelo preditivo para cada genoma. Para criação do modelo de predição do BRAKER, foi fornecido, para cada espécie, o modelo de predição criado pelo Genemark, assim como o proteoma da espécie, disponíveis no GenBank. Além disso, utilizou-se o *Genome Threader* (GREMME; 2013) para gerar os *hints* com base em similaridade (`--prg=gth`); utilizou-se das estruturas geradas pelo *Genome Threader* (`--gth2traingenes`) em conjunto com a predição do Genemark para treinar o Augustus (`--trainFromGth`). Assim como o modelo do Genemark, cada predição do BRAKER é personalizada e específica para cada genoma.

O primeiro modelo preditivo para o Augustus foi criado valendo-se de 100 genes aleatórios do genoma de referência de *E. tenella* Houghton disponível no GenBank. Empregou-se cDNA da base de dados *Eimeria Transcript* DB (http://coccidia.icb.usp.br/eimeriatdb/, acessado em: 14/05/2020) como suporte para predição dos genomas de E. *acervulina*, *E. maxima, E. tenella* Houghton e *E. tenella* Nippon NT2. Utilizou-se este modelo para realizar a predição gênica de todas os genomas. As sequências dos transcritos foram preparadas utilizando o software BLAT v34 (KENT, 2002) com identidade mínima de 92% (`minIdentity=92`). Converteu-se o alinhamento múltiplo em *hints* com o script do Augustus `blat2hints.pl`. Empregou-se os *hints* como evidências de predição no Augustus valendo-se do argumento `--hintsfile` e `--extrinsicCfgFile=extrinsic.E.cfg`. O arquivo de configuração foi fornecido pelo Augustus. Por fim, o segundo modelo preditivo do Augustus procedeu da mesma forma, mas utilizando o arquivo `.gff` da referência de *E. tenella*. Realizou-se a predição utilizando os mesmos *hints* para as espécies de *E. tenella*, *E. acervulina* e *E. maxima*, e sem este dado para as demais.

Comparou-se todos os modelos preditivos com relação à: número de genes preditos; tamanho das proteínas deduzidas; quantidade de proteínas com menos de

100 resíduos aminoacídicos; e, mais importante, número de genes preditos com algum correspondente no ToxoDB (https://toxodb.org/toxo/) versão 46. Os números de predições obtidos em cada modelo podem ser consultados no APÊNDICE A. O modelo que apresentou a menor quantidade de proteínas que não encontraram correspondência com a base de dados do ToxoDB (Augustus com gff de *E. tenella*) foi utilizado para as etapas seguintes de anotação e análise de ortologia. Este procedimento foi adotado para tentar reduzir o número de falsos positivos. Posteriormente, removeu-se predições menores que 100 aminoácidos, a fim de evitar outros possíveis falsos positivos.

Todas as predições realizadas valeram-se dos genomas contendo as regiões repetitivas, devido ao fato de regiões homopoliméricas serem muito comuns em proteínas do gênero, descrito por Reid e colaboradores em 2012.

Realizou-se a anotação gênica por similaridade utilizando os bancos de dados SwissProt (versão 11/2019) e ToxoDB versão 46. Foi empregado uma versão customizada da AnnotaPipeline (MAIA, 2019). O diagrama sequencial dos processos está exposto na Figura 5. Utilizou-se valores de cobertura maior ou igual à 30, identidade maior ou igual à 40 e positividade maior ou igual à 60 para transferir anotação, valendo-se do maior *bitscore* como critério de seleção.

Adotou-se os valores padrão para identidade e positividade (MAIA, 2019). O padrão mínimo de cobertura foi estabelecido após a execução de testes para verificar alterações significativas na troca de anotação ao alterar o valor mínimo de cobertura. Tais resultados estão disponíveis no APÊNDICE B. Proteínas com anotações contendo as palavras chave: `hypothetical`, `unspecified`, `fragment`, `partial`, `unknown`, foram consideradas como hipotéticas. Anotações sem estas palavras consideradas, portanto não hipotéticas, prevaleceram. Após atribuição de função por similaridade, realizou-se a anotação funcional por meio do HMMscan versão 3.1b2 (FINN; CLEMENTS; EDDY, 2011; POTTER *et al.*, 2018), RPSBlast versão 2.9.0 (STATES; GISH, 1994) e InterProscan versão 2.0 (MULDER; APWEILER, 2007). A informação de funcionalidade foi inserida (quando presente) na anotação das proteínas.

Para avaliação do número de proteínas hipotéticas e anotadas, considerou-se anotações que continham a palavra "*hypothetical*" como hipotéticas e as que não tinham como anotadas. Devido à forma como o proteoma de referência de *E. mitis* foi

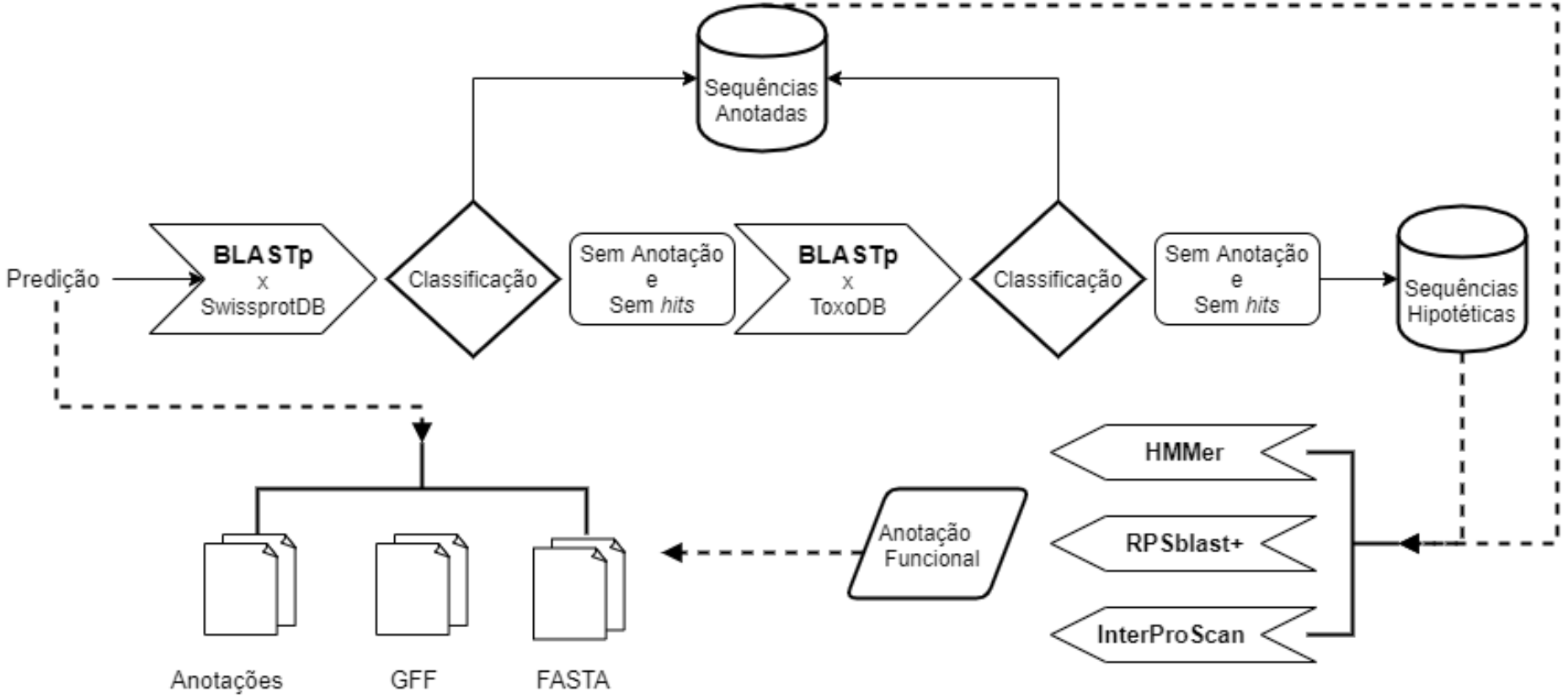
anotado, para este, considerou-se hipotéticas as anotações que continham a palavra "*uncharacterized*" e "*hypothetical*".

3.5 Predição de pseudogenes

Executou-se o programa PseudoPipe para realizar a predição de pseudogenes individualmente para cada genoma. Para esta análise, utilizou-se a localização dos exons de cada proteína predita para o genoma obtido através dos resultados da predição de proteínas, além do genoma e proteoma correspondentes da espécie. Desta forma, o programa foi executado em sua configuração padrão.

Foram utilizados scripts *in-house*, desenvolvidos em linguagem computacional Python (versões 3 e 3.6), para processar os dados em cada uma das etapas do processo de predição.

Figura 5 - Adaptação da AnnotaPipeline utilizada para anotação das proteínas de *Eimeria* spp.



As predições foram comparadas com a base de dados do SwissprotDB. Proteínas não anotadas por similaridade com o SwissprotDB foram comparadas com o ToxoDB, pelo mesmo processo. Ao fim desta etapa, gerou-se dois conjuntos de dados: sequências anotadas e sequências hipotéticas. Ambos os conjuntos foram submetidos à anotação funcional, realizada pelos programas HMMer, RPSblast+ e InterproScan. O compilado destas informações resultou em três arquivos: o de anotações, que exibe a sequência e sua anotação; o GFF que concatena dados de localização do gene; e o arquivo fasta, que contém a anotação, a sequência aminoacídica e informações da localização. Fonte: MAIA, 2019, adaptado.

4 RESULTADOS E DISCUSSÃO

4.1 Montagens

As montagens apresentam métricas melhores que os genomas de referência depositados no GenBank. Todos os genomas montados neste projeto apresentam menor número de *scaffolds* e menos espaçamentos que os respectivos genomas de *Eimeria spp.* conforme a Tabela 3. Isto permite dizer que as montagens estão mais contíguas, agrupadas e sem perda de dados, evidenciado pelo conteúdo GC e tamanho de genoma, próximos dos padrões encontrados nas respectivas referências (NCBI/Genome). As métricas de N50 e L50 estão melhores para a maioria dos genomas, assim como bem próximas da referência para os genomas de *E. tenella*.

As montagens de *E. tenella* são casos particulares, uma vez que apresentam maior número de bases N e tamanho de genoma maiores que a referência. Escolheu-se utilizar os dados estruturais do genoma de referência para obter um genoma maior e mais contíguo, em detrimento de algumas métricas, uma vez que o tamanho esperado do genoma de *E. tenella* corresponde a 60Mb (SHIRLEY, 2000), enquanto a referência possui genoma descrito de 51.8Mb. Em decorrência disso, algumas métricas, como bases N a cada 100 mil pares de base foram infladas, pois a montagem feita neste trabalho conseguiu estabelecer uma estrutura maior do genoma, próximo do tamanho esperado, mas com regiões de baixa precisão na chamada de bases. Assim, essas duas montagens de *Eimeria tenella*, contribuem para próximos estudos que sequenciem e identifiquem tais bases, valendo-se da estrutura montada neste estudo, de forma que os genomas se tornem mais próximos da realidade.

O genoma de *E. mitis* apresenta tamanho muito inferior ao genoma de referência (12Mb a menos). Porém, *E. mitis*, em particular, possui duas versões do genoma publicados. O atual genoma de referência (*GenBank assembly accession*: GCA_000499745.2) permitiu maior identificação de proteínas que sua versão antiga (*GenBank assembly accession*: GCA_000499745.1), porém em detrimento de algumas métricas. A montagem realizada neste estudo reuniu as duas vantagens destas versões dos genomas: melhora de métricas e maior número de proteínas preditas. Neste sentido, o fator que destoa é o tamanho de genoma (72Mb para a atual referência), métrica que conserva a estatística do primeiro genoma (60Mb) na

proposta atual de montagem, e que vai de encontro ao esperado para o gênero, segundo a literatura (SHIRLEY, 2000).

Ao mapear os genomas de referência com suas respectivas montagens, para verificar a integridade das informações contidas nas referências, obtemos alto grau de correspondência: 99,55% para *E. acervulina*, 100%; para *E. brunetti*; 99,92% para *E. necatrix*; 99,99% para *E. mitis* e *E. maxima*; 97,72% para *E. tenella* Nippon NT2; 82,02% para *E. tenella* Houghton. O mapeamento corrobora as estatísticas da Tabela 3, que indica melhora dos genomas dessas espécies sem perda de informação em relação aos genomas de referência.

Outra estatística importante é o número de repetições nos genomas. Os genomas de referência, segundo o grupo de pesquisa responsável pelo estudo, apresentam de 20 a 30% de repetições simples no genoma. Ao aplicar a mesma metodologia para identificar tais repetições nos genomas montados e nos genomas de referência, encontramos uma quantidade de repetições que varia de 13 a 25%, mantendo o padrão descrito por REID e colaboradores em 2014, em que estas repetições, em quase que sua totalidade, são repetições simples. As maiores diferenças entre um genoma de referência e um genoma montado neste estudo se dão para os dois genomas de *E. tenella*, cuja variação chega a 2%. Apesar disso, os números absolutos (pares de bases com repetições) diferem em 214 mil pares de bases para *E. tenella* Houghton e 432 mil pares de bases para *E. tenella* Nippon NT2, 5,8% e 2,8% do total, respectivamente. Isto explica-se pelo fato de as montagens estarem maiores que o genoma de referência usados para comparação, como discutido anteriormente. Este comparativo está exposto no APÊNDICE C.

As montagens apresentam grande redução na cobertura comparado à referência, isto ocorre em decorrência dos valores de qualidade Phred utilizados para limpeza dos dados brutos. Porém, o mapeamento dos *reads* com os genomas finais mostra que a maioria dos *reads* limpos foram utilizados nas respectivas montagens. Obtivemos mapeamento acima de 96% para todos os genomas, exceto *E. praecox* e *E. tenella* Houghton, com 94.14% e 81.05% dos *reads* mapeados, respectivamente.

Para fins de comparação, um estudo realizado em 2013 por Desai e colaboradores, para montagem *de novo* de genomas pequenos (menores que 100 milhões de bases) mostra que uma cobertura em torno de 50x é suficiente para obter boas montagens. Apenas três genomas não condizem com esta recomendação, mas é importante ressaltar que a cobertura se mostra satisfatória, uma vez que os

genomas foram montados com base em uma referência. O mesmo método foi utilizado para a montagem do genoma de referência de *Toxoplasma gondii* ME49*,* que apresenta cobertura de 26.55x (NCBI, Genome). T. *gondii* é um organismo filogeneticamente aparentado de *Eimeria* spp. com tamanho de genoma e conteúdo GC semelhantes (NCBI, Genome), tornando a comparação válida.

Além disso, há profundidade suficiente para garantir a acurácia nas chamadas das bases presentes nos *scaffolds*. Não há um padrão para definir uma boa profundidade de genoma, mas, para métodos de sequenciamento w*hole genome sequencing* (WGS), a Illumina (empresa responsável pela tecnologia de sequenciamento empregada nas amostras de *Eimeria* spp.) recomenda profundidade em torno de 50x para genomas humanos (ILLUMINA, 2020). Novamente, trabalhando com dados genômicos de humanos, Ajay e colaboradores, em 2011, citam que acima de 50x, a profundidade permite sensibilidade para identificação de genótipos diferentes, interessante para análises de polimorfismos de nucleotídeos únicos (SNP) ou uso clínico. Este não é o propósito deste estudo, porém, a profundidade média se mostra muito acima do recomendado, reiterando o alto aproveitamento dos *reads* citado ao apresentar as métricas de mapeamento.

Tabela 3 - Comparativo entre as montagens de *Eimeria* spp. realizadas neste estudo e seus respectivos genomas de referência

	Scaffolds		**N50**		**L50**		**Espaços**		**N / 100kpbs**		**Cobertura**		**Conteúdo GC**		**Dados adicionais**	
Espécie	**Referência**	**Montagem**	**Referência**	**Montagem**	**Referência**	**Montagem**	**Referência**	**Montagem**	**Referência**	**Montagem**	**Referência**	**Montagem**	**Referência**	**Montagem**	**Tamanho do genoma**	**Profundidade média**
EACV	3.415	1.786	33.010	82.446	351	165	1.532	836	334	448	199x	101x	48,3	48,1	47	181x
EBRU	8.575	2.293	33.515	71.624	531	262	19.411	2.693	2 794	3.119	143x	47x	47,9	47,8	64	92x
EPRA	21.348	6.446	8.846	17.087	1.438	670	34.977	5.851	8 843	6.534	102x	38x	45,7	46,0	51	82x
EMIT	15.978	6.615	9.328	18.348	2.055	2.054	56.820	4.029	7.398	4.234	520x	192x	47,6	47,3	60	352x
EMWX	3.564	2.730	26.759	45.864	452	275	1.006	342	218	11	258x	146x	46,6	46,6	45	245x
ENEC	3.707	3.062	40.326	47.128	291	261	960	698	174	22	559x	226x	51,0	51,2	54	411x
ETHE Houghton*	4.665	2.208	200.914	172.693	68	86	8.064	4.120	1.322	17.161	168x	72x	51,3	51,4	57	113x
ETHE Nippon NT2*		1.703		209.502		75		1.671		18.720		32x		51,2	58	63x

O tamanho de genoma é dado em milhões de bases. O conteúdo GC é dado em porcentagem. EACV corresponde à *Eimeria acervulina*; EBRU à *E. brunetti*; EPRA à *E. praecox*; EMIT à *E. mitis*; EMWX à *E. maxima*; ENEC à *E. necatrix*; ETHE à *E. tenella*.

* O genoma de referência utilizado para comparação é da cepa Houghton.

4.2 Predições gênicas

Todas as predições do modelo definido (Tabela 4) apresentam valores superiores aos encontrados nos genomas de referência. Ao analisar os dados de outros coccídios depositados no GenBank (*T. gondii*, *P. vivax*, *P. falciparum*, *Theileria* spp. e *N. caninum*), observa-se que o número de proteínas preditas varia entre cinco e oito mil. Considerando o parentesco filogenético desses parasitos, assim como o alto número de proteínas sem qualquer semelhança com outras predições, fica evidente que, apesar dos testes realizados, o modelo de predição pode ser melhorado a fim de reduzir falsos positivos.

Apesar disso, há um grande número de proteínas preditas que possuem alguma anotação, métrica que chegou a 59% para *E. tenella* Nippon NT2, e 34% para *E. praecox*, com variações entre estes dois extremos. Proteínas hipotéticas podem ser um artefato da técnica e a redução deste número é notável quando comparado com os genomas de referência.

Com relação às três classes de genes ligadas a patogenicidade, há uma uniformidade no número de genes SAG encontrados nos genomas montados, quando comparados com suas respectivas referências. Os genes da família ROP apresentaram maior variação, embora sejam poucos genes por espécie. Com relação aos produtos dessas famílias, de modo geral, houve ganho de informação (encontro de novas proteínas dentro da família) para todas as classes. Na família ROP foram encontradas novas proteínas em todos os genomas, destacando a proteína ROP14, encontrada em *E. brunetti*, *E. mitis* e *E. maxima*. As proteínas ROP21 e ROP38 presentes no proteoma de referência de *E. acervulina* não foram encontradas na predição realizada para a espécie. Apesar disso, o número de proteínas preditas ainda está distante dos 28 genes ROP encontrados para *E. tenella*, segundo análise de Talevich e Kannan em 2013. Com relação à família SAG, apenas *E. tenella*, em seu proteoma de referência, apresenta anotação específica, contendo 22 anotações. Nas predições realizadas neste trabalho, foram encontradas apenas quatro anotações específicas nos proteomas de *E. tenella*, de forma que apesar da uniformidade no número de genes, há diferença em suas anotações. Neste sentido, *E. necatrix*, que não possui anotação específica para SAG em seu proteoma de referência, apresentou 12 anotações, das quais três são compartilhadas com os proteomas de *E. tenella*. Estas informações estão disponíveis no APÊNDICE D.

Dentre muitas proteínas compartilhadas, destaca-se a ROP38, encontrada nos proteomas de *E. tenella* e *E. brunetti*, duas espécies que em conjunto com *E. necatrix* são classificadas como hemorrágicas, causando maior dano aos hospedeiros. A proteína ROP38 é encontrada no proteoma de referência de *E. acervulina,* mas não é encontrada no proteoma realizado neste estudo, para a mesma espécie. Esta proteína é uma das abordagens recentes no desenvolvimento de vacinas contra *T. gondii* (NOSRATI *et al.*, 2020). Outra proteína exclusiva é a ROP27 de *E. acervulina*, que não possui anotação para os proteomas de referência, sendo encontradas em gêneros como *Toxoplasma* e *Plasmodium* (NCBI, Gene). Além disso, há nove das 12 anotações de genes SAG encontradas em *E. necatrix* que são exclusivas da mesma. Porém, como discutido anteriormente, a maioria destes marcadores está presente no proteoma de referência de *E. tenella*, mas não foram identificados no proteoma produzido por este estudo, para a espécie. Esses marcadores compartilhados e exclusivos merecem atenção quando o objetivo é o diagnóstico ou tratamento específico. Apesar destes achados, um estudo de expressão é necessário para verificar a veracidade destas predições no modelo *in vivo*. As predições *in silico* podem apresentar artefatos, de modo que genes com expressão variável, descritos em apicomplexa (YEOH *et al.*, 2019), podem ser parcialmente identificados, de forma a alterar a anotação atribuída à proteína.

Tabela 4 - Proteínas preditas e anotações de famílias gênicas ligadas à patogenicidade

	Número de proteínas		Proteínas hipotéticas									
			Totais		Proteínas com anotação funcional*	Totais		ROP		SAG	
Espécie	**Montagem**	**Referência**	**Montagem**	**Referência**	**Montagem**	**Montagem**	**Referência**	**Montagem**	**Referência**	**Montagem**	**Referência**
EACV	8.905	6.867	4.490	4.453	600	4.415	2.414	6	8	16	16
EBRU	11.470	8.711	6.726	6.462	1.138	4.774	2.249	8	7	120	105
EMIT	11.627	10.073	7.229	8.076	1.269	4.398	1.997	8	6	202	173
EMWX	7.980	6.057	4.312	3.929	541	3.668	2.128	5	5	33	39
ENEC	9.822	8.609	4.473	5.851	775	5.349	2.758	9	7	103	120
EPRA	10.792	7.635	7.053	6.070	1.008	3.676	1.565	5	4	20	19
ETHE Houghton	9.682	8.609	4.298	5.839	480	5.384	2.770	9	7	91	87
ETHE Nippon NT2	8.855		3.616		449	5.239		9		83	

EACV corresponde à *Eimeria acervulina*; EBRU à *E. brunetti*; EPRA à *E. praecox*; EMIT à *E. mitis*; EMWX à *E. maxima*; ENEC à *E. necatrix*; ETHE à *E. tenella.* As proteínas hipotéticas com anotação funcional apresentam indicadores de função biológica provenientes da *Gene Ontology*. O genoma de referência de *E. tenella* utilizado para comparação é da cepa Houghton.
*Não se aplica aos genomas de referência

As predições de genes não codificantes estão expostas na Tabela 5. A maioria das predições foram validadas com dados disponíveis no RNACentral. Não há menção de predição de estruturas não codificantes na publicação dos genomas de referência de *Eimeria* spp. e por isso não é passível de comparação. Os dados de *T. gondii* mostram 183 predições de tRNAs, número inferior ao número de sequências utilizadas para comparação em cada espécie, obtidas do RNACentral, conforme descrito na metodologia (NCBI, Genome). Apesar disto, há um grande número de sequências semelhantes com evidências, que indicam a possibilidade de genes multicópias.

Com relação aos rRNAs, há menos predições do que sequências disponíveis na base de dados do RNACentral. Ao analisar coccídeos próximos obtém-se dois extremos, *T. gondii* com alto número de rRNAs (superior a 400) e *Plasmodium* spp. (22 a 71) com números próximos das predições realizadas (NCBI, Genome).

As predições de snoRNAs não são comuns em organismos parasitários. Por se tratarem de pequenas sequências, a chance de falsos positivos é maior, fato que influenciou no filtro destas sequências. Deste modo, os resultados expostos na Tabela 5 constituem apenas sequências com alta similaridade com os poucos dados disponíveis na literatura para esses transcritos. Assim, a predição de snoRNAs realizadas neste projeto representa um grande salto no número de sequências encontradas e pode sinalizar para a presença de multicópias destes genes e possíveis variantes fenotípicas.

Tabela 5 - Predição de genes não codificantes

Espécie	tRNAs		rRNAs		snoRNAs
	Total	Validado	Total	Validado	Total
EACV	492	458	40	36	43
EBRU	311	305	49	45	46
EPRA	303	297	29	27	116
EMIT	247	231	56	49	95
EMWX	312	305	60	58	115
ENEC	290	273	30	30	33
ETHE Houghton	259	256	24	24	19
ETHE Nippon NT2	252	250	24	22	20

EACV corresponde à *Eimeria acervulina*; EBRU à *E. brunetti*; EPRA à *E. praecox*; EMIT à *E. mitis*; EMWX à *E. maxima*; ENEC à *E. necatrix*; ETHE à *E. tenella.* As validações foram realizadas com base na similaridade com sequências depositadas no RNACentral. As comparações foram realizadas a nível de gênero para os rRNAs e a nível de espécie para os tRNAs e snoRNAs. O número total de snoRNAs corresponde apenas aos validados. As predições de rRNAs agrupam genes de 28s, 18s e 5.8s.

4.3 Pseudogenes

A busca de pseudogenes, expostos na Tabela 6, retornou muitos fragmentos de genes (FRAG), algumas duplicações (DUP) e pseudogenes oriundos de retrotransposição (PSSD). Ao analisar as classes de proteínas relacionadas a patogenicidade, há poucos pseudogenes da família ROP - estão presentes em menor número nos genomas - e muitas predições para pseudogenes da família SAG.

Com relação à SAG, as espécies hemorrágicas apresentam um número próximo de pseudogenes, com pequena variação para a cepa Nippon NT2 de *E. tenella*. *E. acervulina* e *E. praecox*, que são espécies que causam menos dano e compartilham o mesmo local de infeção (QUIROZ-CASTAÑEDA; DANTÁN-

GONZÁLEZ, 2015) possuem poucos pseudogenes desta família. *E. mitis* apresenta muitos pseudogenes e difere deste padrão.

Com exceção das cepas de *E. tenella*, todas as outras predições feitas neste trabalho subestimam os números de pseudogenes encontrados por Reid e colaboradores em 2014. Os pseudogenes da classe SAG estão categorizados como duplicações ou fragmentos, classificando-os como genes não processados, de forma que estes permanecem como resquícios de genes outrora ativos. O fato de muitos pseudogenes (os três pseudogenes de *E. acervulina*, por exemplo) localizarem-se antes ou depois de genes SAG preditos ativos reforça este argumento. É comum encontrar blocos de genes SAG em sequência em um mesmo *scaffold* (REID *et al.*, 2014) e o fato de pseudogenes estarem próximos destes blocos pode sinalizar para inativações ao longo do curso da evolução dos patógenos.

A localização destes pseudogenes também pode indicar regiões regulatórias da expressão de genes SAG que o acompanham. Mecanismos de regulação gênica pré-/pós-transcricional e epigenética, mediados por pseudogenes, são descritos na literatura (CHEETHAM; FAULKNER; DINGER, 2020; MURO; MAH; ANDRADE-NAVARRO, 2011). Apesar disso, nenhuma regulação do tipo foi descrita para os genes de *Eimeria*, o que suscita perguntas para futuros estudos.

Os pseudogenes da família ROP aparecem em apenas quatro espécies, uma delas da classe das mais patogênicas, *E. brunetti*. *E. maxima* foi a única a apresentar um pseudogene oriundo de duplicação gênica (ROP35). Todos os demais pseudogenes foram identificados como fragmentos – ROP14 (*E. brunetti*), ROP27 (*E. acervulina*) e ROP35 (*E. mitis*), além de um pseudogene inespecífico de *E. brunetti*. São poucas predições, ao mesmo tempo que são poucas proteínas preditas. Não havendo resquícios de pseudogenes, ao menos pseudogenes de genes da própria espécie, podemos inferir que estes constituem um cerne muito restrito de proteínas essenciais, conservadas desde seu surgimento evolutivo. De outro modo, fragmentos e duplicações de genes estariam distribuídos no genoma, na forma de pseudogenes, como ocorre com os genes SAG.

Estudos com pseudogenes de organismos do filo apicomplexa são raros. Identificação de pseudogenes entre as diferentes espécies pode demonstrar fenômenos parecidos como o descrito por Reid em 2012, que descreveu o gene codificante da proteína ROP18 de *T. gondii* estava “pseudogenizado” em *N. caninum*. A proteína ROP18 não foi identificada nas predições gênicas e, portanto, não faz parte

das predições de pseudogenes. Esta é uma perspectiva de como a identificação de pseudogenes pode levar a compreensão do compartilhamento destes marcadores de patogenicidade entre os parasitos causadores da coccidiose aviária.

Tabela 6 - Pseudogenes

Espécie	Total	Classes			Anotações	
		PSSD	DUP	FRAG	ROP	SAG
EACV	1.324	166	323	835	1	3
EBRU	3.553	454	321	2.778	2	33
EMIT	3.728	329	620	2.779	1	101
EMWX	775	82	119	574	1	7
ENEC	1.606	203	262	1.141	-	36
EPRA	1.244	125	120	999	-	4
ETHE Houghton	1.375	220	160	995	-	31
ETHE Nippon NT2	753	93	142	518	-	46

EACV corresponde à *Eimeria acervulina*; EBRU à *E. brunetti*; EPRA à *E. praecox*; EMIT à *E. mitis*; EMWX à *E. maxima*; ENEC à *E. necatrix*; ETHE à *E. tenella*. PSSD corresponde à pseudogenes produzidos por retrotransposição; FRAG corresponde à pseudogenes identificados como fragmento de genes; DUP corresponde à pseudogenes identificados como duplicação de genes.

5 CONCLUSÃO

O protocolo aplicado para remontar os genomas das sete espécies de *Eimeria* causadoras da coccidiose aviária mostrou-se eficaz, uma vez que as métricas de montagem foram melhoradas e, na maioria dos casos, sem perda de informação. Este novo conjunto de dados genômicos pode se provar de extrema importância e relevância para este e tantos outros grupos de pesquisa.

A identificação de pseudogenes nos genomas das sete espécies de *Eimeria* causadoras da coccidiose aviária representa uma pesquisa de base essencial para o avanço na compreensão do papel destas sequências no genoma dos parasitos. Neste contexto, o presente trabalho faz contribuições importantes para melhor guiar os futuros estudos que possam surgir acerca de pseudogenes em coccídeos.

Os pseudogenes dos marcadores de patogenicidade mostram que apenas os genes da família SAG deixaram marcas significativas no genoma desses organismos, de forma que perspectivas como análises de blocos de sintenia entre os pseudogenes encontrados, assim como a identificação de pseudogenes de genes de outras espécies são alçadas para pesquisas futuras.

REFERÊNCIAS

ABRIL, J. F.; CASTELLANO, S. Genome Annotation. **Encyclopedia of Bioinformatics and Computational Biology**, p. 195-209, 2019. Elsevier. http://dx.doi.org/10.1016/b978-0-12-809633-8.20226-4.

AJAY, S. S.; PARKER, S. C. J.; OZEL A., H.; FUENTES F., K. V.; MARGULIES, E. H. 2011. Accurate and comprehensive sequencing of personal genomes. **Genome Research**, 21(9). doi:10.1101/gr.123638.111.

ANDREWS, S. 2010. **FastQC: a quality control tool for high throughput sequence data**. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

ANGEL, V. D.; HJERDE, E.; STERCK, L.; CAPELLA-GUTIERREZ, S.; NOTREDAME, C.; PETTERSSON, O. V.; AMSELEM, J.; BOURI, L.; BOCS, S.; KLOPP, C. Ten steps to get started in Genome Assembly and Annotation. **F1000Research**, v. 7, p. 148, 2018. https://doi.org/10.12688/f1000research.13598.1.

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T. Gene Ontology: tool for the unification of biology.: tool for the unification of biology. **Nature Genetics**, v. 25, n. 1, p. 25-29, mai. 2000. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/75556

BALAKIREV, E. S.; AYALA, F. J. Pseudogenes: Are They “Junk” or Functional DNA? Annual Review of Genetics, v. 37, n. 1, p.123-151, dez. 2003. **Annual Reviews**. http://dx.doi.org/10.1146/annurev.genet.37.040103.103949.

BANKEVICH, A.; NURK, S.; ANTIPOV, D.; GUREVICH, A.; DVORKIN, M.; KULIKOV, A. S.; LESIN, V.; NIKOLENKO, S.; PHAM, S.; PRJIBELSKI, A.; PYSHKIN, A.; SIROTKIN, A.; VYAHHI, N.; TESLER, G.; ALEKSEYEV, M. A.; PEVZNER, P. A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of Computational Biology**, 2012. http://dx.doi.org/10.1089/cmb.2012.0021.

BESEMER, J.; BORODOVSKY, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.: web software for gene finding in prokaryotes, eukaryotes and viruses. **Nucleic Acids Research**, v. 33, p. 451-454, 1 jul. 2005. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gki487.

BERA, A.K.; BHATTACHARYAA, D.; PANA, D.; DHARAB, A.; KUMARC, S.; DAS, S.K. Evaluation of Economic Losses due to Coccidiosis in Poultry Industry in India. **Agricultural Economics Research**, Kolkata, v. 23, n. 1, p.91-96, 12 jan. 2010. http://dx.doi.org/10.22004/ag.econ.92156.

BLAKE, D. P.; KNOX, J.; DEHAECK, B.; HUNTINGTON, B.; RATHINAM, T.; RAVIPATI, V.; AYOADE, S.; GILBERT, W.; ADEBAMBO, A. O.; JATAU, I. D. Re-calculating the cost of coccidiosis in chickens. **Veterinary Research**, v. 51, n. 1, p. 115, 2020. https://doi.org/10.1186/s13567-020-00837-2.

BRADLEY, P. J.; SIBLEY, L. D. Rhoptries: an arsenal of secreted virulence factors. **Current Opinion in Microbiology**, v. 10, n. 6, p. 582–587, 2007. https://doi.org/10.1016/j.mib.2007.09.013.

BAUM, J.; PAPENFUSS, A. T.; BAUM, B.; SPEED, T. P.; COWMAN, A. F. Regulation of apicomplexan actin-based motility. **Nature Reviews Microbiology**, v. 4, n. 8, p. 621–628, 2006. https://doi.org/10.1038/nrmicro1465.

BOETZER, M.; HENKEL, C. V.; JANSEN, H. J.; BUTLER, D.; PIROVANO, W. Scaffolding pre-assembled contigs using SSPACE. **Bioinformatics**, v. 27, n. 4, p. 578-579, 12 dez. 2010. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btq683.

BOETZER, M.; PIROVANO, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information.: scaffolding bacterial draft genomes using long read sequence information. **Bmc Bioinformatics**, v. 15, n. 1, p. 1-12, 20 jun. 2014. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2105-15-211.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p.2114-2120, 1 abr. 2014. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btu170.

BUSHNELL, B. “BBMap: A Fast, Accurate, Splice-Aware Aligner.” 2014.

BUTLER, J.; MACCALLUM, I.; KLEBER, M.; SHLYAKHTER, I. A.; BELMONTE, M. K.; LANDER, E. S.; NUSBAUM, C.; JAFFE, D. B. ALLPATHS: *De novo* assembly of whole-genome shotgun microreads. **Genome Research**, v. 18, n. 5, p.810-820, 21 fev. 2008. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/gr.7337908.

CACHO, E. del; PAGES, M.; GALLEGO, M.; MONTEAGUDO, L.; SÁNCHEZ-ACEDO, C. Synaptonemal complex karyotype of *Eimeria tenella*. **International Journal for Parasitology**, v. 35, n. 13, p.1445-1451, nov. 2005. Elsevier BV. http://dx.doi.org/10.1016/j.ijpara.2005.06.009.

CASTAÑÓN, C. A. B.; FRAGA, J. S.; FERNANDEZ, S.; GRUBER, A.; COSTA, L. da F. Biological shape characterization for automatic image recognition and diagnosis of protozoan parasites of the genus *Eimeria*. **Pattern Recognition**, v. 40, n. 7, p.1899-1910, jul. 2007. Elsevier BV. http://dx.doi.org/10.1016/j.patcog.2006.12.006.

CASTRO, C. J.; NG, T. F. F. U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. **Journal of Computational Biology**, v. 24, n. 11, p.1071-1080, nov. 2017. Mary Ann Liebert Inc. http://dx.doi.org/10.1089/cmb.2017.0013.

CHAN, P. P.; LOWE, T. M. TRNAscan-SE: searching for trna genes in genomic sequences.: Searching for tRNA Genes in Genomic Sequences. **Methods in Molecular Biology**, p. 1-14, 2019. Springer New York. http://dx.doi.org/10.1007/978-1-4939-9173-0_1.

CHAPMAN, H. D. Origins of Coccidiosis Research in the Fowl—The First Fifty Years. **Avian Diseases**, v. 47, n. 1, p.1-20, jan. 2003. American Association of Avian Pathologists (AAAP). http://dx.doi.org/10.1637/0005-2086(2003)047[0001:oocrit]2.0.co;2.

CHAPMAN, H. D.; BARTA, J. R.; BLAKE, D.; GRUBER, A.; JENKINS, M.; SMITH, N. C.; SUO, X.; TOMLEY, F. M. A Selective Review of Advances in Coccidiosis Research. **Advances In Parasitology**, p.93-171, 2013. Elsevier. http://dx.doi.org/10.1016/b978-0-12-407705-8.00002-1.

CHAPMAN, H. D. Milestones in avian coccidiosis research: A review. **Poultry Science**, v. 93, n. 3, p.501-511, 26 fev. 2014. Oxford University Press (OUP). http://dx.doi.org/10.3382/ps.2013-03634.

CHEETHAM, S. W.; FAULKNER, G. J.; DINGER, M. E. Overcoming challenges and dogmas to understand the functions of pseudogenes. **Nature Reviews Genetics**, p.1-11, 17 dez. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41576-019-0196-1.

CHOW, Y. P.; WAN, K. L.; BLAKE, D. P.; TOMLEY, F.; NATHAN, S. Immunogenic *Eimeria tenella* glycosylphosphatidylinositol-anchored surface antigens (SAGs) induce inflammatory responses in avian macrophages. **PLoS ONE**, v. 6, n. 9, 2011. https://doi.org/10.1371/journal.pone.0025233.

CONSORTIUM, The Gene Ontology. The Gene Ontology Resource: 20 years and still going strong.: 20 years and still Going strong. **Nucleic Acids Research**, v. 47, n. 1, p. 330-338, 5 nov. 2018. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gky1055.

DESAI, A.; MARWAH, V. S.; YADAV, A.; JHA, V.; DHAYGUDE, K.; BANGAR, U.; KULKARNI, V.; JERE, A. Identification of Optimum Sequencing Depth Especially for *De Novo* Genome Assembly of Small Genomes Using Next Generation Sequencing Data. **Plos One**, v. 8, n. 4, p. 1-10, 12 abr. 2013. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0060204.

DUBEY, J. P.; LINDSAY, D. S.; SAVILLE, W. J. A.; REED, S. M.; GRANSTROM, D. E.; SPEER, C. A. A review of Sarcocystis neurona and equine protozoal myeloencephalitis (EPM). **Veterinary Parasitology**, v. 95, n. 2-4, p.89-131, fev. 2001. Elsevier BV. http://dx.doi.org/10.1016/s0304-4017(00)00384-8.

DUBEY, J. P. Review of *Neospora caninum* and neosporosis in animals. **The Korean Journal of Parasitology**, v. 41, n. 1, p.1-10, 2003. Korean Society for Parasitology. http://dx.doi.org/10.3347/kjp.2003.41.1.1.

EWELS, P.; MAGNUSSON, M.; LUNDIN, S.; KÄLLER, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, v. 32, n. 19, p.3047-3048, 16 jun. 2016. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btw354.

FANTHAM, H. B. 1910. The morphology and life-history of *Eimeria* (coccidium) *avium*: A sporozoön causing a fatal disease among young grouse. **Christ's College Proc. Zool. Soc. Lond.**

FINN, R. D.; CLEMENTS, J.; EDDY, S. R. HMMER web server: Interactive sequence similarity searching. **Nucleic Acids Research**, v. 39, n. SUPPL. 2, 2011. https://doi.org/10.1093/nar/gkr367.

GHORBANI, M.; KARIMI, H. 2015. **Bioinformatics Approaches for Gene Finding**. IJSRST. 4. 12-15. Disponível em: < https://www.researchgate.net/publication/305720975_Bioinformatics_Tools_for_Protein_Analysis>.

GREMME, G. 2013. Computational Gene Structure Prediction. PhD thesis, **Universität Hamburg**.

GUO, A.; CAI, J.; GONG, W.; YAN, H.; LUO, X.; TIAN, G.; ZHANG, S.; ZHANG, H.; ZHU, G.; CAI, X. Transcriptome Analysis in Chicken Cecal Epithelia upon Infection by *Eimeria tenella In Vivo*. **Plos One**, v. 8, n. 5, p.1-10, 30 mai. 2013. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0064236.

GUREVICH, A.; SAVELIEV, V.; VYAHHI, N.; TESLER, G. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072–1075, 2013. https://doi.org/10.1093/bioinformatics/btt086.

HEBERT, P. D. N.; BRAUKMANN, T. W. A.; PROSSER, S. W. J.; RATNASINGHAM, S.; DEWAARD, J. R.; IVANOVA, N. V.; JANZEN, D. H.; HALLWACHS, W.; NAIK, S.; SONES, J. E. A Sequel to Sanger: amplicon sequencing that scales. **Bmc Genomics**, v. 19, n. 1, p.1-14, 27 mar. 2018. Springer Nature. http://dx.doi.org/10.1186/s12864-018-4611-3.

HOFF, K. J.; LOMSADZE, A.; BORODOVSKY, M.; STANKE, M. Whole-Genome Annotation with BRAKER. **Methods in Molecular Biology**, p. 65-95, 2019. Springer New York. http://dx.doi.org/10.1007/978-1-4939-9173-0_5.

HOLT, C.; YANDELL, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.: an annotation pipeline and genome-database management tool for second-generation genome projects. **Bmc Bioinformatics**, v. 12, n. 1, p. 1-24, dez. 2011. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2105-12-491.

HYATT, D.; CHEN, G.; LOCASCIO, P. F.; LAND, M. L.; LARIMER, F. W.; HAUSER, L. J. Prodigal: prokaryotic gene recognition and translation initiation site identification.: prokaryotic gene recognition and translation initiation site identification. **Bmc Bioinformatics**, v. 11, n. 1, p. 1-15, 8 mar. 2010. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2105-11-119.

ILLUMINA (San Diego) (Org.). Genomic Sequencing. 2010. Disponível em: <https://www.illumina.com/documents/products/datasheets/datasheet_genomic_sequence.pdf>. Acesso em: 23 fev. 2020.

ILLUMINA. **Coverage depth recommendations**: Learn how to estimate the depth of sequencing coverage needed for your research. 2020. Disponível em: <https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html>. Acesso em: 14 ago. 2020.

KALVARI, I.; NAWROCKI, E. P.; ARGASINSKA, J.; QUINONES-OLVERA, N.; FINN, R. D.; BATEMAN, A.; PETROV, A. I. Non-Coding RNA Analysis Using the Rfam Database. **Current Protocols in Bioinformatics**, v. 62, n. 1, p. 51-60, jun. 2018. Wiley. http://dx.doi.org/10.1002/cpbi.51.

KAMMONEN, J. I.; SMOLANDER, O. P.; PAULIN, L.; PEREIRA, P. A. B.; LAINE, P.; KOSKINEN, P.; JERNVALL, J.; AUVINEN, P. GapFinisher: A reliable gap filling pipeline for SSPACE-LongRead scaffolder output. **Plos One**, v. 14, n. 9, p.1-12, 9 set. 2019. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0216885.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic acids research**, v. 28, n. 1, p. 27–30, 1 jan. 2000. https://dx.doi.org/10.1093%2Fnar%2F28.1.27.

KENT, W. J. BLAT---The BLAST-Like Alignment Tool. **Genome Research**, v. 12, n. 4, p. 656-664, 20 mar. 2002. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/gr.229202.

LAGESEN K, HALLIN P. F.; RØDLAND E.; STÆRFELDT H. H.; ROGNES T.; USSERY D. W. RNammer: consistent annotation of rRNA genes in genomic sequences. **Nucleic Acids** Res. 2007 Apr 22. https://doi.org/10.1093/nar/gkm160.

LEINONEN, R.; SUGAWARA, H.; SHUMWAY, M. The Sequence Read Archive. Nucleic Acids Research, v. 39, p. 19-21, 9 nov. 2010. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkq1019.

LIAO, P.; SATTEN, G. A.; HU, Y. PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. **Genetic Epidemiology**, v. 41, n. 5, p.375-387, 31 mai. 2017. Wiley. http://dx.doi.org/10.1002/gepi.22048.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078-2079, 8 jun. 2009. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btp352.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754-1760, 18 maio 2009. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btp324.

LOMSADZE, A. Gene identification in novel eukaryotic genomes by self-training algorithm. **Nucleic Acids Research**, v. 33, n. 20, p. 6494-6506, 27 nov. 2005. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gki937.

LUO, R.; LIU, B.; XIE, Y.; LI, Z.; HUANG, W.; YUAN, J.; HE, G.; CHEN, Y.; PAN, Q.; LIU, Y. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. **Gigascience**, v. 1, n. 1, p.1-15, dez. 2012. Oxford University Press (OUP). http://dx.doi.org/10.1186/2047-217x-1-18.

MAIA, G. A. **Ferramenta integrada para anotação de proteínas hipotéticas: estudo de caso utilizando análises proteogenômicas em *Trypanosoma rangeli***. 2019. 79 f. Dissertação (Mestrado) - Curso de Biotecnologia e Biociências, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, Florianópolis, 2019. Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/215541/PBTC0302-D.pdf?sequence=-1&isAllowed=y. Acesso em: 03 jan. 2021.

MURO, E. M.; MAH, N.; ANDRADE-NAVARRO, M. A. Functional evidence of post-transcriptional regulation by pseudogenes. **Biochimie**, 2011. https://doi.org/10.1016/j.biochi.2011.07.024.

MEHLHORN, H. *Eimeria* Species. **Encyclopedia of Parasitology**, p.1-13, 2015. Springer Berlin Heidelberg.

MORRISON, D. A. Evolution of the Apicomplexa: where are we now? **Trends in Parasitology**, v. 25, n. 8, p.375-382, ago. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.pt.2009.05.010.

MULDER, N.; APWEILER, R. InterPro and InterProScan. **Comparative Genomics**, p. 59-70, 2007. Humana Press. http://dx.doi.org/10.1007/978-1-59745-515-2_5.

NADALIN, F.; VEZZI, F.; POLICRITI, A. GapFiller: a *de novo* assembly approach to fill the gap within paired reads.: a de novo assembly approach to fill the gap within paired reads. **Bmc Bioinformatics**, v. 13, n. 14, p. 1-20, set. 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2105-13-s14-s8.

NOSRATI, M. C.; GHASEMI, E.; SHAMS, M.; SHAMSINIA, S.; YOUSEFI, A.; NOURMOHAMMADI, H.; JAVANMARDI, E.; KORDI, B.; MAJIDIANI, H.; GHAFFARI, A. D. *Toxoplasma gondii* ROP38 protein: Bioinformatics analysis for vaccine design improvement against toxoplasmosis. **Microbial Pathogenesis**, v. 149, 2020. https://doi.org/10.1016/j.micpath.2020.104488.

OAKES, R. D.; KURIAN, D.; BROMLEY, E.; WARD, C.; LAL, K.; BLAKE, D. P.; REID, A. J.; PAIN, A.; SINDEN, R. E.; WASTLING, J. M. The rhoptry proteome of *Eimeria tenella* sporozoites. **International Journal for Parasitology**, v. 43, n. 2, p. 181–188, 2013. https://doi.org/10.1016/j.ijpara.2012.10.024.

OLIVEIRA, J. V. de A.; COSTA, F.; BACKOFEN, R.; STADLER, P. F.; WALTER, M. E. M. T.; HERTEL, J. SnoReport 2.0: new features and a refined support vector machine to improve snorna identification. **Bmc Bioinformatics**, v. 17, n. 18, p. 1-15, dez. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s12859-016-1345-6.

PRICE, A. L.; JONES, N. C.; PEVZNER, P. A. *De novo* identification of repeat families in large genomes. **Bioinformatics**, v. 21, n. 1, p. 351-358, 1 jun. 2005. Oxford University Press (OUP). https://doi.org/10.1093/bioinformatics/bti1018.

POP, M. Comparative genome assembly. **Briefings in Bioinformatics**, v. 5, n. 3, p.237-248, 1 jan. 2004. Oxford University Press (OUP). http://dx.doi.org/10.1093/bib/5.3.237.

POTTER, S. C; LUCIANI, A.; EDDY, S. R.; PARK, Y.; LOPEZ, R.; FINN, R. D. HMMER web server: 2018 update. **Nucleic Acids Research**, v. 46, n. W1, p. W200–W204, 2018. https://doi.org/10.1093/nar/gky448.

QUIROZ-CASTAÑEDA, R. E.; DANTÁN-GONZÁLEZ, E. Control of Avian Coccidiosis: Future and Present Natural Alternatives. **Biomed Research International**, v. 2015, p.1-11, 2015. Hindawi Limited. http://dx.doi.org/10.1155/2015/430610.

RAMA, J. D. *Eimeria acervulina* E *Eimeria tenella*: Estudo de casos na avicultura de corte industrial. 2016. 38 f. TCC (Graduação) - Curso de Medicina Veterinária, **Universidade de Brasília faculdade de Agronomia e Medicina Veterinária**, Brasília, 2016. Disponível em: <https://bdm.unb.br/bitstream/10483/16322/1/2016_JessicaDelazzeriRama_tcc.pdf>. Acesso em: 16 fev. 2020.

REID, A. J.; VERMONT, S. J.; COTTON, J. A.; HARRIS, D.; HILL-CAWTHORNE, G. A.; KÖNEN-WAISMAN, S.; LATHAM, S. M.; MOURIER, T.; NORTON, R.; QUAIL, M. A.Comparative Genomics of the Apicomplexan Parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia Differing in Host Range and Transmission Strategy. **Plos Pathogens**, v. 8, n. 3, p.1-13, 22 mar. 2012. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.ppat.1002567.

REID, A. J.; BLAKE, D. P.; ANSARI, H. R.; BILLINGTON, K.; BROWNE, H. P.; BRYANT, J.; DUNN, M.; HUNG, S. S.; KAWAHARA, F.; MIRANDA-SAAVEDRA, D. Genomic analysis of the causative agents of coccidiosis in domestic chickens. **Genome Research**, v. 24, n. 10, p.1676-1685, 11 jul. 2014. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/gr.168955.113.

SALZBERG, S. L.; PHILLIPPY, A. M.; ZIMIN, A.; PUIU, D.; MAGOC, T.; KOREN, S.; TREANGEN, T. J.; SCHATZ, M. C.; DELCHER, A. L.; ROBERTS, M. GAGE: A critical evaluation of genome assemblies and assembly algorithms**. Genome Research**, v. 22, n. 3, p.557-567, 6 jan. 2012. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/gr.131383.111.

SEEBER, F.; STEINFELDER, S. Recent advances in understanding apicomplexan parasites. **F1000research**, v. 5, p.1369-1380, 14 jun. 2016. F1000 Research Ltd. http://dx.doi.org/10.12688/f1000research.7924.1.

SHEN, B.; SIBLEY, L. D. The moving junction, a key portal to host cell invasion by apicomplexan parasites. **NIH Public Access**, 2012. https://doi.org/10.1016/j.mib.2012.02.007.

SHIRLEY, M. W. The genome of *Eimeria* spp., with special reference to *Eimeria tenella* - A coccidium from the chicken. **International Journal for Parasitology**, v. 30, n. 4, p. 485–493, 2000. https://doi.org/10.1016/S0020-7519(99)00183-6.

SMIT, A.; HUBLEY, R.; GREEN, P. *RepeatMasker Open-4.0*. 2013-2015. Disponível em: <http://www.repeatmasker.org>.

STATES, D. J.; GISH, W. QGB: combined use of sequence similarity and codon bias for coding region identification: Combined Use of Sequence Similarity and Codon Bias for Coding Region Identification. **Journal of Computational Biology**, v. 1, n. 1, p. 39-50, jan. 1994. Mary Ann Liebert Inc. http://dx.doi.org/10.1089/cmb.1994.1.39.

STANKE, M.; MORGENSTERN, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. **Nucleic Acids Research**, v. 33, p.465-467, 1 jul. 2005. Oxford University Press. https://dx.doi.org/10.1093%2Fnar%2Fgki458.

STEIN, L. Genome annotation: from sequence to biology. **Nature Reviews Genetics**, v. 2, n. 7, p.493-503, jul. 2001. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/35080529.

STEVENS, D.D. 1998. Coccidiosis: Encyclopedia of Immunology, Vol 1. Eds: P.J. Delves and I.M Roitt. **Academic Press**, London, pp. 591-593.

SOLANO-GALLEGO, L.; SAINZ, A.; ROURA, X.; ESTRADA-PEÑA, A.; MIRÓ, G. A review of canine babesiosis: the European perspective. **Parasites & Vectors**, v. 9, n. 1, p.1-10, 11 jun. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s13071-016-1596-0.

SOUZA, W.; BELFORD JR., R. comp. Toxoplasmose e *Toxoplasma gondii*. **SciELO Books**, p. 1-206, 2014. Editora FIOCRUZ. http://dx.doi.org/10.7476/9788575415719.

ŠLAPETA, J.; MORIN-ADELINE V. 2011. **Apicomplexa Levine 1970**. Sporozoa Leucart 1879. Version 18, mai. 2011. Disponível em: <http://tolweb.org/Apicomplexa/2446/2011.05.18>.

TALEVICH, E.; KANNAN, N. Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors. **BMC Evolutionary Biology**, v. 13, n. 1, 2013. https://doi.org/10.1186/1471-2148-13-117.

THIBAUD-NISSEN F., SOUVOROV A., MURPHY T.; DICUCCIO, M.; KITTS, P. Eukaryotic Genome Annotation Pipeline. 2013 Nov 14. In: **The NCBI Handbook**. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Disponível em: <https://www.ncbi.nlm.nih.gov/sites/books/NBK169439/>. Acesso em: 30 abr. 2020.

TUTAR, Y. Pseudogenes. Comparative and Functional Genomics, v. 2012, p.1-4, 2012. **Hindawi**. http://dx.doi.org/10.1155/2012/424526.

UNIPROT, Consortium, UniProt: a worldwide hub of protein knowledge, **Nucleic Acids Research**, Volume 47, Issue D1, 08 January 2019, Pages 506–515. https://doi.org/10.1093/nar/gky1049.

VAN BAREN, M. J. Iterative gene prediction and pseudogene removal improves genome annotation. **Genome Research**, v. 16, n. 5, p.678-685, 1 mai. 2006. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/gr.4766206.

WAJID, B.; SERPEDIN, E. Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers. **Genomics, Proteomics & Bioinformatics**, v. 10, n. 2, p.58-73, abr. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.gpb.2012.05.006.

WALKER, B. J.; ABEEL, T.; SHEA, T.; PRIEST, M.; ABOUELLIEL, A.; SAKTHIKUMAR, S.; CUOMO, C. A.; ZENG, Q.; WORTMAN, J.; YOUNG, S. K. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. **Plos One**, v. 9, n. 11, p. 1-17, 19 nov. 2014. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0112963.

WANG, Z.; CHEN, Y.; LI, Y. A Brief Review of Computational Gene Prediction Methods. **Genomics, Proteomics & Bioinformatics**, v. 2, n. 4, p. 216-221, nov. 2004. Elsevier BV. http://dx.doi.org/10.1016/s1672-0229(04)02028-5.

WENGER, A. M.; PELUSO, P.; ROWELL, W. J.; CHANG, P.; HALL, R. J.; CONCEPCION, G. T.; EBLER, J.; FUNGTAMMASAN, A.; KOLESNIKOV, A.; OLSON, N. D. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. **Nature Biotechnology**, v. 37, n. 10, p. 1155-1162, 12 ago. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41587-019-0217-9.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, v. 13, n. 5, p. 329-342, 18 abr. 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrg3174.

YE, C.; MA, Z. S.; CANNON, C. H.; POP, M.; YU, D. W. Exploiting sparseness in *de novo* genome assembly. **Bmc Bioinformatics**, v. 13, n. 6, p.1-8, 19 abr. 2012. Springer Nature. http://dx.doi.org/10.1186/1471-2105-13-s6-s1.

YEOH, L. M.; LEE, V. V.; MCFADDEN, G. I.; RALPH, S. A. Alternative splicing in apicomplexan parasites. v. 10, n. 1, p. 1-10, 19 fev. 2019. **American Society for Microbiology**, 2019. https://doi.org/10.1128/mBio.02866-18.

ZERBINO, D. R.; BIRNEY, E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. **Genome Research**, v. 18, n. 5, p.821-829, 21 fev. 2008. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/gr.074492.107.

ZHANG, Z.; CARRIERO, N.; ZHENG, D.; KARRO, J.; HARRISON, P. M.; GERSTEIN, M. PseudoPipe: an automated pseudogene identification pipeline. **Bioinformatics**, v. 22, n. 12, p.1437-1439, 30 mar. 2006. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btl116.

APÊNDICE A – Comparativo entre os modelos de predição gênica

A Planilha 1 exibe o número de predições; a média de tamanho das proteínas preditas; número de proteínas menores que cem aminoácidos; mediana e moda do tamanho das predições. A planilha compara os quatro modelos preditivos com a referência e ajudou na escolha do modelo de predição adotado (Augustus_gff). A referência de *Eimeria tenella* é da cepa Houghton.

Planilha 1 - Comparativo modelos de predição por espécie

Eimeria acervulina

	Referência	Augustus_gb	Augustus_gff	Braker	Genemark
Predições	6.867	10.288	9.518	9.820	8.954
Média	669,49	538,73	568,19	541,15	507,25
Menores que cem aminoácidos	293	1.176	973	972	36
Mediana	460	315	338	326	313
Moda	235	87	81	95	114

Eimeria brunetti

	Referência	Augustus_gb	Augustus_gff	Braker	Genemark
Predições	8.711	13.272	12.944	16.665	10.640
Média	546,95	460,16	466,16	391,35	312,74
Menores que cem aminoácidos	479	1.630	1.474	2.328	75
Mediana	347	261	265	223	221
Moda	133	111	98	88	99

Eimeria necatrix

	Referência	Augustus_gb	Augustus_gff	Braker	Genemark
Predições	8.609	10.802	10.665	12.090	8.935
Média	548,67	551,55	551,36	485,53	504,37
Menores que cem aminoácidos	510	879	843	1.099	29
Mediana	372	342	342	298	332
Moda	174	118	103	130	103

Continua

Eimeria praecox

	Referência	Augustus_gb	Augustus_gff	Braker	Genemark
Predições	7.635	13.053	12.647	15.475	8.498
Média	478,48	382,92	388,23	342,37	555,82
Menores que cem aminoácidos	551	2.045	1.855	2.448	13
Mediana:	307	222	225	202	354
Moda	121	85	109	97	122

Eimeria mitis

	Referência	Augustus_gb	Augustus_gff	Braker	Genemark
Predições	8.748	13.578	13.313	17.573	9.210
Média	446,54	428,40	432,21	378,23	609,84
Menores que cem aminoácidos	535	1.885	1.686	2.430	17
Mediana	303	247	251	223	381
Moda	198	126	126	93	102

Eimeria maxima

	Referência	Augustus_gb	Augustus_gff	Braker	Genemark
Predições	6.057	9.427	9.145	8.686	7.893
Média	664,68	539,45	546,67	567,05	563,38
Menores que cem aminoácidos	249	1.318	1.165	622	23
Mediana	459	305	317	367,5	344
Moda	154	78	126	122	108

Eimeria tenella Houghton

	Referência	Augustus_gb	Augustus_gff	Braker	Genemark
Predições	8.609	10.862	10.847	9.745	8.144
Média	505,06	504,57	502,48	513,05	521,05
Menores que cem aminoácidos	662	1.147	1.165	650	35
Mediana	342	318	316	343	344.5
Moda	121	103	89	112	112

Eimeria tenella Nippon NT2

	Referência	Augustus_gb	Augustus_gff	Braker	Genemark
Predições	8 609	11.586	9.587	10.097	7.509
Média	505,06	471,40	535,49	505,87	655,46
Menores que cem aminoácidos	662	1.407	732	754	10
Mediana	342	292	351	325	445
Moda	121	67	104	116	136

Média, mediana e moda são dados em relação ao tamanho das proteínas. “Menores que cem aminoácidos” são valores absolutos em relação ao total de predições.

APÊNDICE B – Testes para definição do valor de cobertura

O teste foi realizado com a predição realizada para *Trypanosoma rangeli* realizado autor da AnnotaPipeline construção da mesma em 2019 (MAIA, 2019). O banco de dados utilizado foi o TrytripDB em sua versão número 43. Para transferir a anotação para uma proteína considerou o melhor alinhamento que satisfez os valores mínimos de: cobertura (variável), positividade (maior ou igual à 60%) e identidade (maior ou igual à 40%) (MAIA, 2019).

Com base nos testes apresentados na Tabela 7, realizou-se as comparações apresentadas na Tabela 8.

Tabela 7 - Proteínas anotadas por valor de cobertura

Valor mínimo de cobertura	Proteínas Anotadas	Proteínas hipotéticas
0 %	6.079	4.269
30 %	6.019	4.329
40 %	6.009	4.339
50 %	5.988	4.360
60 %	5.947	4.401
70 %	5.873	4.475
80 %	5.698	4.650
90 %	5.289	5.059

Fonte: Elaborado pelo autor (2020)

Tabela 8 - Comparativo entre os extemos de cobertura

Valores mínimos de cobertura comparados	Número de anotações iguais	Troca de anotação	Troca de significado
0 e 90	4.979	168	55
0 e 30	6.003	2	1
30 e 60	5.576	19	5

Fonte: Elaborado pelo autor (2020)

Considerando o baixo número de troca de significado (1,1% no comparativo mais extremo), optou-se por definir o valor de corte de cobertura em 30, amparado pelos valores limítrofes de identidade ou positividade, desde que um deles satisfaça os valores mínimos. Entende-se troca de significado como: alteração significativa da função da proteína (“transialidase” por “ATP sintase”, por exemplo). A troca de anotação engloba a definição anterior e ainda trocas que não alteram a descrição da função molecular da proteína (SAG por SAG Family, por exemplo). O comparativo foi realizado manualmente.

APÊNDICE C – Repetições nos genomas de *Eimeria* spp.

Tabela 9 - Comparativo do número de regiões repetitivas encontradas nos genomas montados e suas respectivas referências

Espécie	Pares de bases		Repetições Simples	
	Montagem	Referência	Montagem	Referência
E. acervulina	10.666.171 (22,51%)	10.315.130 (22,51%)	20,90	20,90
E. brunetti	14.537.005 (22,48%)	14.723.369 (22,01%)	20,64	20,14
E. mitis	15.048.680 (24,69%)	16.023.450 (26,52%)	22,44	23,88
E. maxima	9.586.176 (20,90%)	9.503.645 (20,67%)	19,47	19,29
E. necatrix	9.786.682 (17,88%)	9.417.571 (17,12%)	16,81	16,06
E. praecox	13.119.171 (25,43%)	14.951.328 (24,88%)	23,59	22,93
E. tenella Houghton	7.676.434 (13,23%)	7.890.969 (15,21%)	12,29	14,05
E. tenella Nippon NT2	7.458.932 (12,96%)		12,01	

A porcentagem de pares de base é dada sobre o tamanho dos genomas. As repetições simples são informadas em porcentagem, sobre o tamanho dos genomas.

APÊNDICE D – Qualitativo de proteínas das classes ROP e SAG

A Tabela 10 apresenta as proteínas ROP e SAG encontradas nos genomas montados neste estudo.

Tabela 10 - Proteínas em cada família gênica ligada a patogenicidade de *Eimeria* spp. encontradas nos genomas montados neste estudo

Espécie	ROP	SAG
EACV	17, 23, 25, 27, 30, 35	-
EBRU	14, 17, 23, 25, 30, 35, 38	-
EMIT	14, 17, 21, 23, 25, 30, 35	-
EMWX	14, 17, 21, 23, 35	-
ENEC	17, 21, 23, 25, 30, 35	3, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21
EPRA	17, 23, 25, 30, 35	-
ETHE Houghton	17, 21, 23, 25, 30, 35, 38	10, 16, 17,19
ETHE Nippon NT2	17, 21, 23, 25, 30, 35, 38	10, 16, 17,19

EACV corresponde à *Eimeria acervulina*; EBRU à *E. brunetti*; EPRA à *E. praecox*; EMIT à *E. mitis*; EMWX à *E. maxima*; ENEC à *E. necatrix*; ETHE à *E. tenella*. As proteínas de cada família estão separadas por vírgula. As espécies que não apresentam tipificação de proteínas SAG possuem anotação genérica: *"SAG Family"*

A Tabela 11 apresenta as proteínas ROP e SAG encontradas nos genomas de referência.

Tabela 11 - Proteínas em cada família gênica ligada a patogenicidade de *Eimeria* spp. encontradas nos genomas de referência

Espécie	ROP	SAG
EACV	17, 21, 23, 25, 30, 35, 38	-
EBRU	17, 21, 23, 25, 30, 35	-
EMIT	17, 23, 25, 30, 35	-
EMWX	17, 21, 23, 35	-
ENEC	17, 21, 23, 25, 30, 35	-
EPRA	17, 23, 35	-
ETHE	17, 21, 23, 25, 30,35	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23

EACV corresponde à *Eimeria acervulina*; EBRU à *E. brunetti*; EPRA à *E. praecox*; EMIT à *E. mitis*; EMWX à *E. maxima*; ENEC à *E. necatrix*; ETHE à *E. tenella*. As proteínas de cada classe estão separadas por vírgula. As espécies que não apresentam tipificação de proteínas SAG possuem anotação genérica: *"SAG Family"*