UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

ELOY DA SILVA SEABRA JUNIOR

A predição gênica do genoma de *Triatoma infestans* Klug, 1835

Rio de Janeiro

2019

Eloy da Silva Seabra Junior

A predição gênica do genoma de *Triatoma infestans* Klug, 1835

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Bioquímica, Instituto de Química, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Bioquímica.

Orientador: Rafael Dias Mesquita

Rio de Janeiro

2019

Mxxx

Seabra-Junior, Eloy da Silva

A predição gênica do genoma de *Triatoma infestans* Klug, 1835 / Eloy da Silva Seabra Junior. – Rio de Janeiro: UFRJ/ IQ, 2019.

x f.

Orientador: Rafael Dias Mesquita.

Tese (Mestrado em Ciências) - Universidade Federal do Rio de Janeiro, Instituto de Química, Programa de Pós-Graduação em Bioquímica, 2019.

1. Genômica. 2. Bioinformática. 3. *Triatoma infestans*. 4. Predição Gênica.

I. Mesquita, Rafael Dias. (Orient.). II. Universidade Federal do Rio de Janeiro. Instituto de Química. Programa de Pós-Graduação em Bioquímica. III. A predição gênica do genoma de *Triatoma infestans* Klug, 1835.

Eloy da Silva Seabra Junior

A predição gênica do genoma de *Triatoma infestans* Klug, 1835

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Bioquímica, Instituto de Química, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Bioquímica.

Aprovado em:

_____

Dr. Rafael Dias Mesquita (orientador)

_____

Dra. Mônica Ferreira Moreira Carvalho Cardoso (Membro Interno)

_____

Dr. Pedro Lagerblad de Oliveira (Membro Externo)

_____

Dr. Francisco Prosdocimi de Castro Santos (Membro Externo)

_____

Dra. Ana Claudia do Amaral Melo (Suplente Interno)

_____

Dr. Thiago Estevam Parente Martins (Suplente Externo)

Dedico este trabalho a minha noiva Regina Rafaela, por ela todas as coisas ganharam sentido.

AGRADECIMENTOS

Devo agradecer primeiramente a Deus, por ter sido sempre um Pai para mim, por ter traçado meus caminhos até aqui e ter estado sempre comigo. Em segundo lugar devo agradecer a meus compatriotas, pelo voto de confiança, por terem sustentado meus estudos desde minha juventude, espero um dia poder retribuir a enorme generosidade de um povo tão sofrido.

Agradeço a minha família por ter sempre esperado o melhor de mim e terem me ajudado a me tornar o que sou hoje.

Devo também minha gratidão aos excelentes professores que tive, ao meu professor de matemática Nilson, por ter acreditado em mim e ter me ajudado a alcançar a conquista mais importante da minha vida, ao professor Rodrigo Bisaggio por ter sido acima de tudo o exemplo de professor e ter me inspirado a também me tornar um professor. Agradeço especialmente ao professor Rafael Dias Mesquita por ter me acompanhado nestes quase dez anos em que me aventuro na bioinformática.

Agradeço também aos meus colegas de laboratório, principalmente pela paciência e por todo apoio emocional que me deram nos últimos dois anos.

Por fim, gostaria de agradecer a minha noiva Regina Rafaela de Souza Santos, por ter me ensinado a sorrir. Esta página é curta demais para agradecê-la como ela merece e mesmo uma vida inteira juntos não me parece suficiente.

Hucusque auxiliatus est nobis Dominus

(Liber Primum Regum, VII. 12)

RESUMO

Triatoma infestans é historicamente o mais importante vetor da doença de chagas, principalmente por causa da sua grande antropofilia.

O objetivo deste trabalho é produzir modelos confiáveis de genes de proteínas para o já sequenciado genoma de Triatoma infestans.

Predizer genes em um genoma de eucariotos é uma tarefa difícil, e modelos estatísticos específicos são necessários para cada espécie. A produção destes modelos é geralmente chamada de treinamento e depende de um conjunto de genes anotados manualmente e de boa qualidade.

Este trabalho utilizou uma metodologia previamente desenvolvida neste grupo de pesquisa para encontrar genes cópia única e de alta qualidade de anotação a partir de sequências de CDS utilizando o software SIM4. A metodologia desenvolvida identifica os genes que codificam cada CDS, alinha o CDS com o genoma e classifica os alinhamentos em categorias relevantes (completos, 5' truncados, 3' truncados, fragmentados ou repetitivos) para o treinamento de programas de predição gênica. Esse procedimento resultou em 1147 alinhamentos completos de CDS que foram utilizados para o treinamento de softwares de predição.

Três programas de predição gênica foram testados, MAKER, GENEID e Augustus. Os resultados foram comparados utilizando o programa BUSCO e demonstraram que o procedimento padrão do Augustus teve a melhor performance. Então este software foi selecionado para passos subsequentes de otimização como uma otimização estatística, do próprio pacote Augustus, e implementação de dados públicos de sequenciamento de cDNA.

A predição final resultou em 24489 produtos gênicos que compreendem 93.9% dos genes universalmente conservados e cópia única de endopterigotos. Os resultados são próximos dos obtidos para as predições mais recentes de genoma de hemípteros, com 91.4% para *Cimex lectularius* e 95.9% para *Halyomorpha halys*.

ABSTRACT

Triatoma infestans is historically the most important vector of Chagas' disease mainly due to it's great antrophily.

Our goal in this work is to produce reliable protein gene models for the already sequenced genome of this vector.

Predicting genes in eukaryotes' genome is a difficult task, and specific statistical models are needed for each species. The production of such statistical models are usually called "training" and relies on a set of manually-annotated good-quality genes.

We used a pipeline previously developed by us to find high quality single copy genes from putative full-length coding sequences (CDS) from transcriptomic data using the SIM4 software. The developed pipeline identifies the gene(s) coding each CDS, classifies them in relevant categories (full-length, 5'-truncated, 3'-truncated, fragmented and repetitive) for training prediction software. This procedure resulted in a set of 1147 full-length CDS-genome alignments used for the training procedure.

We tried three gene prediction approaches, Maker (Cantarel et al., 2008), GENEID (Guigó et al., 1992) and Augustus (Stanke et Waack, 2003). The results were benchmarked using the BUSCO software (Simão et al., 2015) showing that Augustus standard pipeline performed better. Then, this protocol was selected, following to improvement steps with Augustus statistics optimization pipeline and also with the implementation of RNAseq evidence data from public sources.

The final prediction resulted in 24489 gene products whose comprise 93.9% of BUSCO endopterygota universally conserved single copy genes. The results were close to the most recent Hymenoptera's genome predictions as 91.4% BUSCOS found for *Cimex lectularius* and 95.9 % BUSCOS found for *Halyomorpha halys*.

LISTA DE ILUSTRAÇÕES

Figura 16: Comparação do número de BUSCOs econtrados (de um total de 2442) para as predições preliminares, com o AUGUSTUS, o MAKER e o GENEID.

Figura 17: Comparação geral dos BUSCOs (de um total de 2442) encontrados para a montagem do genoma (Assembly), montagem do transcriptoma utilizando o genoma como referência (Transcriptome_ref), a montagem do transcriptoma sem utilizar o genoma de referência Transcriptome_uref, e a predição final do AUGUSTUS.

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

DNA ácido desoxirribonucleico

RNA ácido ribonucleico

m.a.a milhões de anos atrás

NCBI National Center for Biotechnology Information

SRA Sequence Read Archive

CDS (Coding Sequences) Sequências codificantes

cDNA DNA complementar: dupla fita de DNA sintetizada a partir de um molde de RNA.

SUMÁRIO

1. INTRODUÇÃO

1.1 Doença de Chagas e seus vetores

A Doença de Chagas é uma doença infecciosa causada por um protozoário do grupo dos Kinetoplastida, da espécie *Trypanosoma cruzi*, sendo endêmica em 21 países da América, desde o México até o Chile e a Argentina e afetando mais de 5 milhões de pessoas o que leva a um número estimado de 12 mil mortes por ano em 2010 (PÉREZ-MOLINA & MOLINA, 2018). Por ser uma doença que afeta principalmente populações mais pobres ou rurais é considerada uma doença negligenciada e a morbidez associada a doença, que tem por principais sinais cardiopatias, arritmias e magavísceras, a torna um grande problema social, representando uma barreira para o desenvolvimento de diversas regiões da América Latina. Estima-se que as mortes e a morbidez relacionados a doença representam uma perda de mais de 1 bilhão de dólares anuais para os países mais ao sul da América do Sul (CONTEH et al., 2010).

O *Trypanosoma cruzi é* um parasita heteroxeno que possui como hospedeiro intermediário, e vetor, insetos hemípteros da subordem Heteroptera, família Reduviidae e subfamília Triatominae. Estes insetos são comumente referidos em língua portuguesa como triatomíneos ou barbeiros (REZENDE & RASSI, 2008), ou até mesmo “besouros-barbeiros” apesar de não pertencerem à ordem Coleoptera que compreende os verdadeiros besouros. São também de conhecimento popular em outras partes da américa e em língua inglesa são conhecidos como “kissing-bugs” (GARCIA et al., 2015), algo que poderíamos traduzir livremente como percevejo-beijador, ambas as denominações populares se baseiam no fato destes animais se alimentarem do sangue de hospedeiros vertebrados incluindo o ser humano.

Das 151 espécies de triatomíneos reconhecidas até o presente (JUSTI et al., 2016) *Triatoma infestans* é historicamente o principal vetor da Doença de Chagas, porém depois do esforço da Iniciativa Cone Sul, a transmissão da Doença de Chagas por esse vetor foi erradicada do Uruguai em 1997, do Chile em 1999 e do Brasil em 2006 de acordo com a certificação Organização Panamericana de Saúde (COURA & DIAS,

2009). A transmissão da doença de Chagas por *Triatoma infestans* persiste na Argentina e no Paraguai, e residualmente nos estados brasileiros do Rio Grande do Sul (RS) e Bahia (BA) (COURA, 2015).

Com a eliminação da transmissão pelo *Triatoma infestans* outros vetores se tornaram importantes no território brasileiro, dos quais se destacam *Panstrongylus megistus* (BURMEISTER, 1835) como o principal vetor brasileiro (COURA, 2015), devida a sua ampla distribuição pelo país, e *Rhodnius prolixus* (STÅL, 1859) como o principal vetor na Colômbia, Guiana, Guiana Francesa, Suriname e Venezuela, devido a sua alta antropofilia e rápido ciclo de desenvolvimento. Outras espécies também adquiriram importância vetorial como: *Triatoma dimidiata* (LATREILLE, 1811), *Triatoma brasiliensis* (NEIVA, 1911), *Triatoma sordida* (STÅL, 1859), *Triatoma maculata* (ERICKSON, 1848), *Panstrongylus geniculatus* (LATREILLE, 1811), *Rhodnius ecuadoriensis* (LENT & LEON, 1958), *Rhodnius pallescens* (BARBER, 1932) e *Rhodnius brethesi* (MATTA, 1919).

Uma descrição mais detalhada da distribuição e padrão biogeográfico dos triatomíneos pode ser encontrada em Monteiro et al. (2018).

Na presente dissertação o foco será dado na espécie *Triatoma infestans*, objeto deste estudo.

1.2 Sistemática e filogenética de *Triatoma infestans*.

A classificação sistemática de *Triatoma infestans* o inclui na ordem Hemiptera, subordem Heteroptera, família Reduviidae, subfamília Triatominae, tribo Triatomini.

Os primeiros estudos de filogenia da subfamília Triatominae consideravam que ela era monofilética baseando-se principalmente na hematofagia e características biológicas, ecológicas e morfológicas relacionadas a este hábito alimentar (LENT & WYGODZINSKY, 1979). A vasta diversidade de habitats ocupados pelos triatomíneos e as fortes semelhanças morfológicas com outros grupos de reduvídeos levaram a monofilia desta subfamília a ser disputada (SCHAEFER, 2003) e uma primeira análise

de filogenia molecular utilizando um único marcador recuperou a subfamília Triatominae como polifilética (SILVA et al., 2005). Apesar disso, estudos mais recententes utilizando mais caracteres tanto morfológicos (WEIRAUCH, 2008) quanto moleculares (WEIRAUCH & MUNRO, 2009) e mais representantes da família Reduviidae recuperaram a subfamília Triatominae como monofilética suportando a hipótese de Lent & Wygodzinsky (1979). Ainda assim estudos mais recentes (ZHANG et al., 2016; HWANG et al., 2012) parecem indicar que a subfamília Triatominae pode ser parafilética em relação aos gêneros *Zelurus* e *Opisthacidius*, o que mantém dúvidas sobre a monofilia do clado.

Existe forte evidência de que a tribo Triatomini da qual pertence o gênero *Triatoma* seja um clado natural (de PAULA et al., 2005), porém a filogenética do gênero *Triatoma* em si é extremamente complexa e ainda espera por uma solução. Para um resumo da complexidade desta questão recomendo a leitura de Justi & Galvão (2017).

Filogeneticamente a espécie *Triatoma infestans* pertence, dentro da tribo Triatomini, ao complexo de espécies *infestans*, subcomplexo *infestans*, ao qual pertencem também as espécies *Triatoma platensis* e *Triatoma delpontei* (JUSTI & GALVÃO, 2017) suas relativas mais próximas.

A espécie *Triatoma infestans* em si é extremamente diversa e algumas subpopulações são descritas: em particular a população “melanosoma” que foi originalmente descrita como uma subespécie, *Triatoma infestans melanosoma,* e posteriormente elevada à espécie *Triatoma melanosoma.* Porém análises de filogenética molecular comprovaram não se tratar de uma espécie distinta e sim de uma população de *Triatoma infestans*, e por isso *Triatoma melanosoma* se tornou um sinônimo taxonômico de *Triatoma infestans* (MONTEIRO et al., 1999).

Um segundo caso é a população “dark morphs” de *Triatoma infestans* que também recebeu especial investigação pela filogenética molecular e se provou tratar de uma variação intraespecífica de *Triatoma infestans* (BARGES et al., 2006).

Uma descrição detalhada da sistemática de triatomíneos pode ser encontrada na recente revisão taxonômica de Barges, Schofield e Dujardin (2017).

1.3 Biogeografia de *Triatoma infestans*

Populações selvagens de *Triatoma infestans* são encontradas amplamente nas planícies áridas do Chaco e nas florestas montanas secas da face oriental dos Andes bolivianos, e também em populações provavelmente ferais (que se adaptaram ao ambiente silvestre após a introdução pelo homem) no Matorral no centro do Chile (MONTEIRO et al., 2018).

Acredita-se que a radiação entre as espécies *T. infestans*, *T. platensis* e *T. delpontei* tenha sido relativamente recente, ocorrendo no Pleistoceno e que o Chaco é o ponto de origem de dispersão da espécie (LENT AND WYGODZINSKY, 1979; BARRETT, 1991; NOIREAU et al., 2000; CEBALLOS et al., 2009;). Alguns autores consideram a posição de *Triatoma delpontei* basal em relação a *Triatoma infestans*+*Triatoma platensis* (BARGUES et al., 2006; GARCÍA et al., 2001; IBARRA-CERDEÑA et al., 2014; SAINZ et al., 2004) enquanto outros autores, minoritários porém mais recentes consideram *Triatoma platensis* basal em relação às outras duas espécies (JUSTI et al., 2016), o que gera um certa controvérsia quanto as relações filogenéticas entre as três espécies. Apesar disso, como as três espécies são naturais do Chaco e savanas adjacentes, as incertezas filogenéticas não complicam a determinação da região do Chaco boliviano como área de origem de *Triatoma infestans*.

1.4 Biologia e evolução de *Triatoma infestans*

Segundo Lehane (2005) existem três principais modelos para a adaptação de insetos à hematofagia, um deles se refere a insetos com aparelhos bucais originalmente mastigadoras ou laceradoras, o que não é o caso dos triatomíneos, já os outros dois modelos se referem a insetos com peças bucais picadoras ou sugadoras.

A adaptação de insetos de aparelho bucal picador sugador a hematofagia poderia ocorrer de duas formas, diretamente a partir de insetos fitófagos, ou a partir de insetos

predadores que se alimentavam de outros invertebrados.

Acredita-se que no caso dos Triatomíneos tenha ocorrido dois importantes eventos: primeiro a adaptação de um hemiptera ancestral fitófago a um reduvídeo ancestral entomófago[1] aproximadamente no final do Cretáceo (HWANG & WEIRAUCH, 2012) após o grande aumento na diversidade de insetos fitófagos com a radiação das Angiospermas (BARBA-MONTOYA et al., 2018; MISOF et al., 2014; NEL et al., 2013; DOYLE & JAMES, 2012); seguido então da adaptação de um reduvídeo entomófago à hematofagia dando origem aos triatomíneos.

A adaptação a hematofagia não tem uma data de origem precisa devido a falta de registros fósseis. Otálora-Luna et al. (2015) propõe 3 hipóteses para o momento de ocorrência desta adaptação. A primeira hipótese, gerada a partir dos resultados de Barges et al. (2000) propõe que a sub-família Triatominae é polifilética, ou seja, que a hematofagia surgiu mais de uma vez entre os reduvídeos, e que a separação entre as tribos Triatomini e Rhodniini ocorreu no Paleoceno-Eoceno (entre 48,9 e 64.4 m.a.a) ou seja, após a completa separação entre a América do Sul e a África, que ocorreu a cerca de 100 m.a.a. Uma segunda hipótese sugere que a origem dos triatomíneos tenha sido bastante anterior, ocorrendo no Cretáceo, a cerca de 100 m.a.a justamente quando o megacontinente Gondwana sofria sua separação e antes do completo isolamento da América do Sul (PATTERSON & GAUNT, 2010). A terceira hipótese sugere uma origem muito mais recente, no Oligoceno (33 m.a.a) (JUSTI et al., 2016; HWANG & WEIRAUCH, 2012) quando a América do Sul já estava completamente isolada dos outros continentes da Gondwana mas ainda sofria grandes mudanças geológicas.

A segunda hipótese, que propõe a origem de Triatominae no Cretáceo recebeu recentemente o suporte do importante achado do fóssil de um provável triatomíneo primitivo do Cretáceo (POINAR, 2019), o que pode revolucionar a compreensão da

[1] Entomófago: “que se alimenta de insetos” no caso dos hemipteras isso pode ocorrer, por exemplo, através da ingestão de hemolinfa.

evolução e biogeografia do grupo.

As principais dificuldades encontradas atualmente no estudo da evolução dos reduvídeos e triatomíneos podem ser resumidas a dois temas: a deficiência em dados moleculares que podem ser utilizados na estimativa dos tempos de divergência; e a escassez de fósseis. Por exemplo, a mais recente e compreensiva reconstrução da filogenia dos reduvídeos, elaborada por Hwang & Weirauch (2012) se valeu de apenas 5 *loci* e 52 fósseis, grande parte preservados em Âmbar Dominicano e a sua maioria recente, datando do Eoceno ao Mioceno (56-23 m.a.a). Trabalhos atuais utilizam uma quantidade muito maior de *loci* Misof et al. (2014), por exemplo, utilizaram 1478 loci, mas para isso são necessárias as sequências completas dos genomas das espécies estudadas.

1.5 A importância do sequenciamento genômico para a biologia molecular

Os genomas dos eucariotos são estruturas extremamente complexas e grande parte do DNA dos eucariotos é composta por sequências repetitivas das quais não compreendemos plenamente a função (ELLIOT et al., 2015). A compreensão da estrutura e da dinâmica do DNA avançou grandemente nos últimos anos devido principalmente ao crescente sequenciamento de novos genomas, que começa a lançar luz nas questões sobre o “DNA lixo” (OHNO, 1972). Estes avanços ocorreram, por exemplo, através de modelos dinâmicos da sequência do DNA (PETROV, 2002; OLIVER et al., 2007; WOODHOUSE et al., 2010; SUN et al., 2012; WOODHOUSE et al., 2014; FREELING et al., 2015) e de técnicas que elucidam a estrutura tridimensional do DNA como o Hi-C (van BERKUM et al., 2010; MIRNY, 2011; BLANK & GOODMAN, 2011; BELTON et al., 2012), ambos intimamente relacionados com o crescente número de genomas sequenciados.

Estes novos estudos têm ajudado a compreender como os tamanhos dos genomas de espécies próximas podem ter tamanhos tão diferentes, o que ficou conhecido como o

“paradoxo do valor de C”[2], e as relações entre a estrutura dos cromossomos e a expressão dos genes.

Nesse contexto o estudo do genoma de *Triatoma infestans* tem uma importância especial devido a alta variabilidade de tamanho do genoma entre as diferentes populações desta espécie, onde o valor de C é estimado entre 1.03 picogramas e 1.82 picogramas[3] (PANZERA et al., 2007), além de triatomíneos possuírem cromossomos holocêntricos, ou seja, sem regiões centroméricas bem localizadas (PITA et al., 2017; PANZERA et al., 2012).

A sequência genômica também continua tendo um papel fundamental na compreensão da fisiologia e evolução dos organismos. Apesar dos transcriptomas revelarem informações de grande valor sobre a expressão gênica e a diversidade de proteínas em um organismo o sequenciamento de genomas ainda é essencial para estudo de ortólogos como os do OrthoDB (ZDOBNOV et al., 2017), Gene Ontology (ASHBURNER et al., 2000; CARBON et al., 2017) e de expansões e contrações de famílias gênicas (DE BIE et al., 2006; HAN et al., 2013), onde o número de ortólogos em cada genoma é uma informação essencial. Estas ferramentas têm sido de grande valor nos estudos mais recentes de evolução e fisiologia comparada, estando presentes em muitos trabalhos recentes sobre novos genomas (ZHOU et al., 2014; CHEN et al., 2015; MENDOZA et al., 2018; CUNNINGHAM et al., 2015; XIA et al., 2017; XU et al., 2016; NEAFSEY et al., 2015).

O conhecimento da sequência genômica é ainda importante para a compreensão da expressão gênica, a visão clássica de “um gene uma enzima” se prova ultrapassada pois as técnicas de sequenciamento de nova geração revelaram que uma porção muito maior do genoma é transcrita (AMARAL et al., 2008) e que eventos de *splicing* alternativo são muito mais frequentes do que se imaginava (RAMANOUSKAYA, 2017). O recente artigo de Deveson et al. (2018) mostra evidências de que todo o

[2] O valor de C é uma estimativa do tamanho de um genoma, sendo definido como a massa de DNA de um núcleo haploide da espécie estudada.
[3] 1 picograma de DNA possui entre 977,0317 e 978,6005 milhões pares de bases de DNA.

cromossomo 21 humano, com exceção da região centromérica, é transcrito e que o *splicing* alternativo é universal entre os exons de genes de RNA não codificante. A compreensão do transcriptoma de um organismo se prova um problema muito complexo, e o uso de um genoma de referência ainda é aconselhado nas análises de sequenciamento de cDNA (CONESA et al., 2016).

Uma nova atenção tem sido dada também à função dos introns (CHOREV et al., 2012; SWINBURNE et al., 2008), e dado que a maioria dos genes parece sofrer *splicing* ainda durante a transcrição (TILGNER, 2012), o sequenciamento do genoma é a única ferramenta viável para se estudar os introns.

1.6 Métodos de predição de proteínas em genomas.

Apesar dos grandes avanços na compreensão da função de sequências repetitivas de DNA e dos RNA não codificantes (PHEASANT & MATTICK, 2007) as proteínas continuam sendo as principais moléculas responsáveis pelo metabolismo da célula e por isso a predição dos genes codificantes das proteínas continua sendo um passo chave na genômica.

Durante a década de 1990 muitos esforços foram unidos para sequenciamento dos primeiros genomas completos de eucariotos e consequentemente para a predição das proteínas que estes genomas codificam. Inicialmente a maioria dos genes codificantes era descoberta através de esforços de sequenciamento de cDNA concomitantes aos esforços do sequenciamento dos genomas, e do alinhamento destes cDNAs com o genoma através de ferramentas de alinhamento local como o BLAST, porém o fato dos mRNAs de eucariotos sofrerem o processo de *splicing* torna essa tarefa complexa pois o BLAST pode errar o posicionamento dos *introns*. Em 1997 foi desenvolvido um programa, o EST_GENOME (MOTT, 1997) que incluía um modelo de introns começados com GT e terminados com AG, dinucleotídeos que delimitam a grande maioria dos introns, permitindo um alinhamento mais preciso entre as sequências de cDNA e genoma. Um procedimento similar é alinhar sequências de aminoácidos ao invés de cDNA contra o genoma, o que permite que proteínas de outros organismos

sejam utilizadas para a predição de ortólogos em um novo genoma, o software GeneWise (BIRNEY & DURBIN, 2000; BIRNEY et al., 2004b) se tornou uma referência nessa tarefa por ser ter sido utilizado na metodologia do *Ensembl* um influente banco de dados genômicos.

Ainda em meados da década de 1990 (STORMO & HAUSSLER, 1994; KULP et al., 1996) modelos matemáticos, principalmente modelos de Hidden Markov, foram desenvolvidos para predizer regiões codificantes dos genomas a partir de características intrínsecas dos genes de um organismo, como as diferenças de composição entre exons e introns e peculiaridades das extremidades dos exons, como o códon inicial (ATG) para o início do primeiro exon, os dinucleotídeos (geralmente GT e AG) que limitam os exons internos e os códons de parada (TAA e TAG e TGA) para o fim do último exon. Essa abordagem foi denominada predição *de novo*, ou *ab initio* e nesta categoria de predição o primeiro software a se destacar foi o Genscan (BURGE & KARLIN, 1997), que utilizava modelos de Hidden Markov generalizados (GHMM) e fundou uma grande linha de preditores gênicos baseados nesse tipo de modelo ou outros tipos de modelo de Hidden Markov como o GENEID (GUIGÓ et al., 1992), o AUGUSTUS (STANKE & WAAK, 2003), o SNAP (KORF, 2004) e o FGENSH (SOLOVYEV et al., 2006).

No início dos anos 2000 as principais linhas de pesquisa em genômica passaram a combinar modelos produzidos por alinhamentos de proteínas e cDNAs com os modelos gerados por metodologias *ab initio*, gerando procedimentos como o OTTO (VENTER et al., 2001), e a metodologia do *Ensembl* (BIRNEY et al., 2004a; CURWEN et al., 2004), além do GenomeScan (YEH et al., 2001), que foi inicialmente utilizado pelo NCBI e depois foi substituído pelo GNOMON (SOUVOROV et al., 2010).

Outra novidade dos anos 2000 foi a utilização de duas ou mais sequências genômicas ao mesmo tempo durante a predição. Devido ao fenômeno da sintenia (as posições dos genes é conservada entre espécies relativamente próximas) a utilização de dois genomas pelos softwares TWINSCAN (KORF et al., 2001; FLICEK et al., 2003) ou

SGP2 (PARRA et al., 2003), ou até mesmo de vários genomas ao mesmo tempo, como é o caso do software N-SCAN (GROSS & BRENT, 2006) é capaz de melhorar a acurácia das metodologias de predição gênica *ab initio*.

A grande quantidade de softwares de predição gênica disponíveis no início dos anos 2000 levou ao desenvolvimento de ferramentas capazes de gerar resultados consenso a partir de diversas predições gênicas como o COMBINER (ALLEN et al., 2004) e o JIGSAW (ALLEN & SALZBERG, 2005).

Com o advento das metodologias de sequenciamento de nova geração uma grande quantidade de dados de sequenciamento de sequências curtas de cDNA se tornou disponível e isso mudou drasticamente a forma como as regiões codificantes de um genoma são anotadas. A utilização destas sequências curtas por programas de montagem de transcriptomas como o ABySS (SIMPSON et al., 2009) e o Trinity (GRABHERR et al., 2011) gera sequências longas e completas o suficiente para serem utilizadas pelas metodologias tradicionais de alinhamento de cDNA, que passaram a ter uma quantidade de dados muito maior para trabalhar. Preditores *ab inicio* também foram beneficiados pelo maior volume de dados de cDNA, pois alguns deles como o AUGUSTUS (STANKE & WAAK, 2003; HOFF & STANKE, 2018), o SNAP (KORF, 2004) e o GeneMark-ES (LOMSADZE et al., 2005) são capazes de utilizar estas sequências curtas de cDNA, além de diversas outras informações, como evidências para direcionar a predição gênica, o que passou a ser denominado *Evidence-driven gene prediction* (Predição gênica direcionada por evidências) alcançando elevadas acurácias. A implementação destas evidências pode ser bastante complicada e isso tem sido facilitado por softwares como o MAKER (CANTAREL et al., 2008, HOLT & YANDELL, 2011) capazes de implementar este tipo de dado de forma semi-automatizada.

A utilização de evidências nos softwares de predição *ab initio* e de modelos matemáticos para auxiliarem o alinhamento de sequências de cDNA com o genoma têm tornado cada vez mais tênue a fronteira entre as metodologias de alinhamentos de

cDNA e predição *ab initio* e não é difícil imaginar que no futuro próximo as duas abordagens de descoberta de regiões codificantes do genoma se unifiquem.

Por fim é importante lembrar que apesar de todo o avanço dos últimos 30 anos a tarefa de predizer as proteínas em um genoma é bastante complicada e até mesmo no caso do genoma humano ainda permanece uma grande dúvida quanto ao seu número exato de genes funcionais (EZKURDIA et al., 2014).

Vários fatores tornam a predição dos genes que codificam proteínas uma tarefa desafiadora, como, por exemplo: 1) a predição gênica em um novo organismo depende de sequências de estrutura conhecida para o treinamento dos programas de predição, o que impõe um dilema para a técnica pois o que podemos conhecer de um genoma depende sempre do que já conhecemos previamente; 2) os diferentes programas de predição gênica possuem diferentes estratégias e baixa compatibilidade entre si, sendo difícil optar por um programa a priori, ou gerar consensos entre os resultados de diferentes programas; 3) alguns introns podem ser muito longos, por exemplo, um dos exons do gene *human dystrophin gene* possui mais de 100 mil pares de bases (SLEATOR, 2010); 4) exons podem ser bastante curtos e a região codificante de um exon pode possuir apenas um par de bases, predizer exons muito curtos é uma tarefa extremamente complicada, ainda mais se seu comprimento for múltiplo de 3 pb pois nesse caso a perda do exon não causa uma alteração no quadro de leitura do gene. 4) mudanças na montagem do genoma muitas vezes exigem que toda a predição gênica seja atualizada, dificultando o progresso dos esforços de anotação pois as posições de referência deixam de ser compatíveis entre si.

2. OBJETIVO

2.1 Objetivo geral

O objetivo do presente estudo foi predizer os genes codificantes de proteínas pelo recém sequenciado genoma de *Triatoma infestans.*.

2.2 Objetivos específicos

- Estabelecer um conjunto de dados (*reads*) de transcriptoma utilizável como evidência para os softwares de predição.
- Estabelecer montagens de transcriptoma com e sem um genoma de referência para a avaliação da completude das predições gênicas.
- Estabelecer um conjunto de genes de referência para treinamento dos programas de predição.
- Treinar diferentes softwares de predição gênica.
- Realizar as predições sem e com o uso de evidência de dados de transcriptoma.

3. MATERIAIS E MÉTODOS

3.1 Avaliação da montagem do genoma

Para se avaliar a completude do genoma montado o software BUSCO versão 3.0.2 (SIMÃO et al., 2015; WATERHOUSE et al., 2018) foi utilizado no modo genoma, onde ele utiliza o software AUGUSTUS versão 3.2.2 (STANKE & WAAK, 2003; HOLT & YANDELL, 2011) com parâmetros para *Drosophila melanogaster* para detectar os genes conservados em um genoma montado de inseto para o qual não existe predição.

O comando utilizado foi:

python ./scripts/run_BUSCO.py -i triatoma_infestans.supercont.fasta -o triatoma_infestans.supercont.fasta_endopterygota -l endopterygota_odb9/ -m genome -c 1 -sp fly

3.2 Aquisição, preparo e montagem de transcriptomas

Dados públicos de transcriptomas foram obtidos a partir do banco de dados Sequence Read Archive (SRA) para serem utilizados como evidências nos programas de predição gênica ou para se buscar genes que não foram identificados no genoma de *Triatoma infestans*.

Todos os dados públicos provenientes de sequenciamento illumina foram utilizados e estão indicados na Tabela 1.

Após sua aquisição, os dados de sequenciamento passaram por um procedimento de limpeza e controle de qualidade.

Tabela 1: Lista de transcriptomas utilizados para gerar evidências gênicas

Sequenciamento	Experimento	Formato da biblioteca	Plataforma	Nome da Amostra
SRR4427078	SRX2248697	PAIRED	ILLUMINA	T_infestans_San_Silvestre
SRR4427079	SRX2248698	PAIRED	ILLUMINA	T_infestans_Villamontes
SRR4449814	SRX2267957	PAIRED	ILLUMINA	T_infestans_Villamontes
SRR4449815	SRX2267958	PAIRED	ILLUMINA	T_infestans_San_Silvestre
SRR4449939	SRX2268055	PAIRED	ILLUMINA	T_infestans_Villamontes
SRR4449940	SRX2268056	PAIRED	ILLUMINA	T_infestans_San_Silvestre
SRR4449941	SRX2268057	PAIRED	ILLUMINA	T_infestans_Villamontes
SRR1168882	SRX470446	SINGLE	ILLUMINA	Ti_Chile_adult
SRR1168885	SRX470450	SINGLE	ILLUMINA	Ti_Chile_larvae
SRR1168888	SRX470453	SINGLE	ILLUMINA	Ti_Peru_adult
SRR1168889	SRX470454	SINGLE	ILLUMINA	Ti_Peru_nymph
SRR1168890	SRX470455	SINGLE	ILLUMINA	Ti_BolF1_adult
SRR1168891	SRX470456	SINGLE	ILLUMINA	Ti_BolF1_nymph
SRR1168892	SRX470457	SINGLE	ILLUMINA	Ti_BolCol_adult
SRR1168893	SRX470458	SINGLE	ILLUMINA	Ti_BolCol_nymph
SRR1168894	SRX470459	SINGLE	ILLUMINA	Ti_adult_Argentina
SRR1168938	SRX470503	SINGLE	ILLUMINA	Ti_nymph_Argentina

No primeiro passo os adaptadores do sistema illumina foram removidos utilizando-se o programa cutadapt versão 1.16 (MARTIN, 2011), após isso os resultados desta etapa de limpeza eram verificados através do programa FastQC versão 0.11.5, (Babraham Bioinformatics https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Essas etapas foram repetidas até que todas as sequências contaminantes fossem removidas.

Para a limpeza de qualidade de bases o programa trimmomatic version 0.36 (BOLGER et al., 2014) foi utilizado pois o tamanho da janela utilizada para calcular a qualidade média de bases é ajustável , uma vantagem em relação ao cutadapt. Um passo final de controle de qualidade utilizando o Fastqc foi realizado antes de se proceder a montagem do transcriptoma com o programa Trinity (GRABHERR et al., 2011), utilizado com configurações padrão.

Após a montagem dos transcriptomas o software BUSCO foi utilizado com o modo transcriptome para se avaliar a completude do transcriptoma montado.

3.3 Preparo do conjunto de sequências codificantes de referência para o treinamento do programa de predição gênica.

Para o preparo do conjunto de genes de referência para o treinamento dos programas de predição gênica um conjunto inicial de 1986 sequências codificantes aparentemente completas foi alinhada contra a montagem do genoma utilizando-se o programa SIM4 (FLOREA et al., 1998) e os resultados do SIM4 foram analisados de modo automatizado por um software desenvolvido em nosso laboratório.

Apenas alinhamentos completos, com identidade igual ou superior a 99% e únicos foram selecionados para o conjunto de sequências de referência.

3.4 Treinamentos e predições gênicas

Três programas foram utilizados preliminarmente para a predição dos genes: GENEID, MAKER e AUGUSTUS.

3.4.1 O treinamento do GENEID

O procedimento de treinamento do GENEID é explicado no *geneid Training Tutorial* disponível no site http://genome.crg.es/software/geneid/training.html mas para o presente trabalho algumas modificações no protocolo foram feitas.

Três informações são necessárias para o arquivo de parâmetros do GENEID, modelos de Markov para o sensor de sinal, modelos de Hidden Markov para o sensor de composição e transição dos exons e uma estimativa do tamanho máximo dos introns.

Para a geração dos modelos do sensor de sinal são obtidas inicialmente sequências de 60 nucleotídeos para o grupo de referência propriamente dito, centralizadas nas intercessões exon/intron ou de início da região codificante do gene (ATG). A aquisição destas sequências de nucleotídeos foi orientada no arquivo *.gff com as posições de alinhamento das sequências codificantes de referência. Com a extração destas sequências é possível saber quais são os dinucleotídios mais frequentemente utilizados como sítio doador ou aceptor nos eventos de *splicing* da espécie. Essa informação é utilizada para se gerar um arquivo de sinal de fundo (*background*), extraindo-se de todo o genoma sequências de 60 nucleotídeos que circundam os dinucleotídeos identificados acima, assim como todos os ATGs.

Dado o grande volume de dados estes dois procedimentos foram realizados de modo automatizado por um *script* de computador na linguagem perl.

Uma vez em posse das sequências de 60 nucleotídeos de treinamento e do sinal de fundo os modelos de markov são calculados utilizando-se o script, *frequency.awk* do tutorial de treinamento do GENEID. O tutorial fornece também um script para calcular qual intervalo, dentre os 60 nucleotídeos, é de fato informativo (*information.awk*), e

deve ser utilizado para o cálculo do modelo de Markov. Uma vez o intervalo informativo decidido, as matrizes são geradas pelos scripts *aux4.awk* e *files.awk*, gerando os modelos finais no formato necessário ao arquivo de parâmetros do GENEID.

Como os intervalos informativos podem ser diferentes em tamanho e posição em relação a cada sinal é necessário se calcular um *offset,* um valor que informa o quanto a janela do intervalo informativo está deslocada do sinal avaliado, este offset representa sempre a posição da interface da região codificante e não codificante, contando-se a partir do zero, seguindo o seguinte exemplo:

Exemplo de intervalo informativo para o sinal ATG:

```
0  1  2  3  4  5  6 7 8 9 10 11 12
C  T  A  C  A  G  A T G A T  A  G
-  -  -  -  -  -  | | | | |  |  |
Não codificante   Codificante
```

Neste caso o offset será 6 pois é a primeira letra do códon inicial ATG é também a primeira letra da região codificante.

Exemplo de intervalo informativo para sítio doador do *splicing* (GT):

```
0  1  2  3  4  5  6 7 8 9  10
A  G  C  G  T  G  G T A G  C
|  |  |  |  |  |  - - - -  -
Exon              Intron
```

Neste caso o offset será 5 pois está é a última posição do exon anterior ao sinal.

Exemplo de intervalo informativo para sítio aceptor do *splicing* (AG):

```
0  1  2  3  4  5  6  7  8  9  10 11
C  T  A  C  A  G  A  G  A  T  A  G
-  -  -  -  -  -  -  -  |  |  |  |
Intron                  Exon
```

Nesse caso o offset será 8 pois esta é a primeira posição do exon seguinte ao sinal.

Para o sensor de composição e transição dos exons, os modelos de Hidden Markov foram gerado a partir das sequências de exons e introns do conjunto de genes de referência. As sequências codificantes (CDS) do conjunto de genes de referência foram utilizadas por conter todos os exons. Para obtenção dos introns um *script* foi escrito para extraí-los a partir do arquivo de alinhamento entre o conjunto de sequências de referência e o genoma.

O volume de dados determina que ordem de modelo de Hidden Markov deve ser utilizada e o cálculo a seguir, presente no tutorial de treinamento do GENEID foi utilizado para determinar a ordem do modelo.

CDS: *Bases* = $90*4^{(n+1)}$

Introns: *Bases* = $30*4^{(n+1)}$

Onde *Bases* é o número total de bases presentes nos dados e *n* é um número que representa a ordem do modelo de Hidden Markov, quando *n* não for um número inteiro ele deve ser arredondado para baixo.

Com essas informações os scripts do tutorial de treinamento do GENEID foram utilizados para calcular os modelos de Hidden Markov. Os modelos de composição e transição de exons foram calculados utilizando-se o script *MarkovMatrices.awk,* pois o quadro de leitura dos códons afeta as probabilidades nas regiões codificantes. Já para os modelos de composição e transição de introns foram calculados utilizando o script *MarkovMatrices-noframe.awk,* visto que não é necessário levar quadros de leitura em

consideração quando se trata de regiões intrônicas.

Por fim um modelo de composição de exons, utilizando o modelo de composição de introns como contraste, foi gerado utilizando-se o script *pro2log_ini.awk* e os dois modelos de composição anteriores. Um modelo de transição de exons, utilizando o modelo de transição de introns como contraste, também foi gerado utilizando-se o script *pro2log_tran.awk* e os dois modelos de transição anteriores.

Para a estimativa do tamanho máximo de introns um histograma foi gerado e o tamanho máximo de intron foi decidido de modo a compreender pelo menos 99% dos introns do conjunto de genes de referência.

Todas as informações obtidas pelos passos supracitados foram utilizadas para a geração de um arquivo de parâmetros para o GENEID.

Uma vez que o arquivo de parâmetros foi gerado ainda foi necessário realizar um processo de otimização estatística utilizando o script *12.Optimization.sh*. Trinta e seis (36) predições gênicas foram realizadas somente para o já conhecido conjunto de genes de referência, variando-se valores de constantes (oWF e eWF) internas ao programa GENEID. Os valores de sensibilidade e especificidade das predições foram calculados com as fórmulas ([“*Bases totais*” - “*Bases não encontradas*”] / “*Bases totais”*) e ([“*Bases encontradas*” - “*Bases erradas*”] / “*Bases encontradas*”), respectivamente. Os valores das constantes internas foram escolhidos baseado na melhor relação custo-benefício quanto a sensibilidade e especificidade.

O software GENEID normalmente gera uma grande quantidade de falsos positivos e por isso os resultados de predição foram filtrados mantendo-se somente modelos gênicos codificantes de proteínas que apresentassem pelo menos um domínio conservado identificado pelo software hmmscan (http://hmmer.org/) com o banco de dados PFAM (EL-GEBALI et al., 2018) e que não fossem identificados como transposons por um rpsblast contra um banco de dados com as principais classes de elementos transponíveis como no trabalho de Fernández-Medina et al. (2016).

3.4.2 O treinamento do AUGUSTUS

Para o treinamento do AUGUSTUS foi necessário basicamente produzir arquivos com os genes confiáveis no formato Gene Bank.

Para isso o arquivo com o conjunto de genes de referência, foi utilizado para gerar um arquivo no formato Gene Bank contendo estas sequências assim como 5 kb imediatamente anteriores e posteriores a estas sequências, através do comando:

./augustus/scripts/gff2gbSmallDNA.pl Ti_perfect.gff triatoma_infestans.supercont.fasta 5000 Ti_genes.gb

Do conjunto de genes de referência, 100 sequências aleatórias foram separadas para o processo de validação da predição, isso foi realizado através de um script do próprio pacote do Augustus:

./augustus/scripts/randomSplit.pl Ti_genes.gb 100

O treinamento foi realizado e validado utilizando-se scripts do pacote AUGUSTUS, através do seguinte conjunto de comandos:

```
### Configurando o caminho para a pasta de configurações do
AUGUSTUS
export
AUGUSTUS_CONFIG_PATH=/home/src/17092801_install_augustus/
augustus-3.2.2/config
### Criando os arquivos de treinamento
./augustus/scripts/new_species.pl --species=t_infestans
### Treinando o Augustus
./augustus/bin/etraining                    --species=t_infestans
Ti_genes.gb.train
### Validando o treinamento
./augustus/bin/augustus  --species=t_infestans  Ti_genes.gb.test
| tee firsttest.out
```

Um segundo passo de treinamento, que realiza uma otimização estatística dos modelos gênicos através de um sistema de partição do conjunto de genes, foi realizado, para isso foram necessários 3 passos:

1) A otimização estatística através do comando:

optimize_augustus.pl --species=t_infestans Ti_genes.gb.train

2) refazer o treinamento com os parâmetros atualizados

etraining --species=t_infestans Ti_genes.gb.train

3) refazer a predição

augustus --species=t_infestans ../../triatoma_infestans.supercont.fasta

Além do conjunto de genes de referência,, dados de transcriptomas públicos de *Triatoma infestans* foram utilizados para a criação de um arquivo de evidências para auxiliar na predição de introns. O procedimento utilizado foi o descrito no tópico "Incorporating RNAseq data into AUGUSTUS predictions with BLAT (including iterative mapping)" da página do AUGUSTUS porém o programa de alinhamento de *reads* utilizado foi o Hisat2 versão 2.1.0 (KIM et al., 2015) ao invés do BLAT.

O procedimento se baseia em primeiro gerar um alinhamento dos *reads* contra o genoma de modo a se gerar as evidências de introns, da seguinte forma:

O genoma mascarado gerado para a predição com o MAKER (próximo tópico) foi utilizado para gerar um arquivo de índices para o hisat2. O hisat2 alinhou os *reads* contra o genoma mascarado e gerou um arquivo de saída no formato *.sam, este arquivo foi convertido para o formato *.bam com o *script samtools,* comando (samtools view -b -S $filename.sam >$filename.bam) e o script *bam2hints* do pacote AUGUSTUS foi executado com as seguintes configurações para gerar as evidências de introns.

bam2hints --intronsonly --in=triatoma_sorted.bam --out=t_infestans.intron_hints.gff

Uma vez que as evidências de introns foram geradas elas são utilizadas pelo augustus para realizar uma predição no genoma, que irá predizer os possíveis exons, para isso é necessário utilizar o AUGUSTUS com as opções: *--alternatives-from-evidence=true,* que permite a predição de variantes de *splicing* a partir de evidências (como as de introns); *--hintsfile*, para a utilização do arquivo de evidências de introns gerado nos passos anteriores; *allow_hinted_splicesites=atac*, para permitir a predição de sítios de *splicing* não convencionais quando houver evidência para isso; *--introns=on*, para que os introns preditos estejam presentes no arquivo de saída; e *--genemodel=complete,* para que apenas genes completos sejam preditos.

Os exons e introns preditos pelo AUGUSTUS são então utilizados para se gerar sequências de junção exon-exon, ou seja, sequências adjacentes a introns, isso é realizado utilizando-se um arquivo *.gff apenas com os introns da predição e o script *intron2exex.pl* do pacote Augustus.

As sequências de junção exon-exon são utilizadas para gerar índices para o hisat2 e as junções que alinham com algum *read* são utilizadas como evidência para as predições do AUGUSTUS. Isso é importante para que apenas introns de genes provavelmente codificantes e com evidências de sequenciamento de transcriptoma sejam utilizados como evidência para a predição, mas as evidências de exon em si não são utilizadas pelo preditor com esta metodologia.

Todos os comandos utilizados para a implementação de dados de transcriptoma como evidência para a predição gênica estão presentes no ANEXO 3 - IMPLEMENTAÇÃO DE DADOS DE TRANSCRIPTOMAS.

3.4.3 O treinamento do MAKER

O treinamento e predição gênica com o software MAKER foram executados pelo colaborador Thiago Venâncio e seus alunos, segundo o procedimento a seguir.[4]

O software RepeatScout foi utilizado para a identificação de novo de elementos repetitivos no genoma de *Triatoma infestans* gerando uma biblioteca de elementos repetitivos. Os elementos repetitivos foram filtrados utilizando-se os seguintes parâmetros:

1) Comprimento maior do que 50 pares de base
2) Frequência maior do que 10

O genoma montado foi então mascarado paras as sequências repetitivas consenso utilizando o software RepeatMasker.

O genoma mascarado foi utilizado para a predição gênica e o protocolo MAKER foi executado iterativamente quatro vezes para um melhor resultado.

Na primeira execução do programa se utilizou o programa AUGUSTUS através do protocolo de anotação do MAKER utilizando-se as sequências do conjunto de genes de referência, como evidências para o AUGUSTUS.

Na segunda execução o conjunto de genes de referência, foi utilizado como evidência para o AUGUSTUS juntamente aos genes preditos na primeira execução.

Na terceira execução o programa SNAP foi treinado utilizando-se os genes preditos na segunda execução.

Na quarta e última execução se utilizou os genes produzidos pela terceira execução, o conjunto de genes de referência, e os genes preditos para *Rhodnius prolixus* como evidências para o programa AUGUSTUS.

---

[4] Tradução minha, texto original no Anexo 2.

A espécie *Rhodnius prolixus* foi utilizada como espécie parâmetro para o software AUGUSTUS.

3.5 Comparação dos resultados das predições gênicas

As predições gênicas foram comparadas entre si através de resultados do software BUSCO (SIMÃO et al., 2015), utilizando-se como parâmetro os ortólogos de Endopterygota (endopterygota_odb9[5]) e o modo *protein*.

O BUSCO também foi utilizado, para gerar dados sobre as outras predições gênicas disponíveis para hemípteros de modo a possibilitar a comparação da qualidade final da predição gênica de *Triatoma infestans*, com a predição de outros genomas, a maioria delas já publicada. As predições para *Acyrthosiphon pisum, Aphis glycines, Diuraphis noxia, Myzus cerasi, Myzus persicae e Myzus persicae* foram obtidas a partir do AphidBase, um banco de dados da plataforma Bioinformatic Platform for Agroecosystem Arthropods (BIPAA), sendo sempre utilizadas as versões mais recentes das anotações. A predição para *Rhodnius prolixus* utilizada foi a versão 1.3 por ter sido utilizada como base para o artigo do genoma (MESQUITA, et al., 2015). A predição para *Cimex lectularius* utilizada foi a versão 1.3 por ser a mais recente disponível no banco de dados VectorBase que hospeda os dados genômicos deste organismo, mas os dados de todas as proteínas do NCBI também foram utilizados para se ter uma ideia do limite máximo de sensibilidade do BUSCO (em outras palavras, quantos genes universalmente conservados de endopterigotos podem estar presentes em hemípteros). Para Oncopeltus fasciatus foi utilizada a anotação v0.5.3 da National Agricultural Library, atualizada em dezembro de 2018. Não existe uma predição publicada para *Halyomorpha halys* por isso no caso deste organismo foram utilizadas as proteínas disponíveis para este organismo no banco de dados de proteínas do NCBI.

[5] Disponível em https://busco.ezlab.org/datasets/endopterygota_odb9.tar.gz

4. RESULTADOS

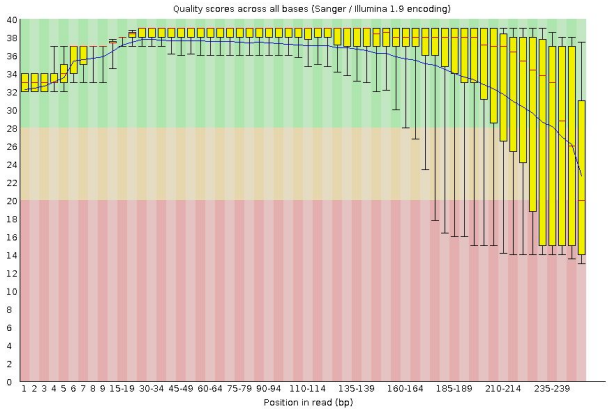
4.1 Montagem e avaliação dos transcriptomas

O controle de qualidade (FastQC) prévio à limpeza das bibliotecas de transcriptomas indicou dois tipos de problema, baixa qualidade de bases na extremidade 3' e contaminação com sequências poli-A ou poli-T, a Tabela 2, indica as bibliotecas nas quais foi detectado algum problema, assim como o problema propriamente dito. A Figura 1 ilustra problemas encontrados quanto a qualidade de bases na extremidade 3' das bibliotecas: SRR4449814_2, SRR4449815_2, SRR4449939_2, SRR4449940_2 e SRR4449941_2.

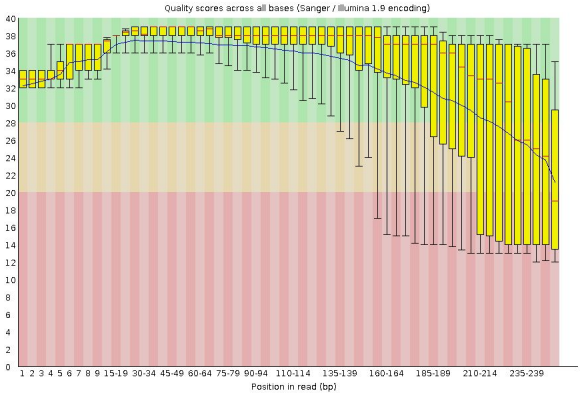
Tabela 2: Resumo dos resultados de contaminação para as bibliotecas utilizadas

Biblioteca	Número de *reads* brutos	Contaminante	Contagem	Percentual %
SRR1168885_1	9191270	poli-T	20548	0.22%
SRR1168885_2	9191270	poli-A	13837	0.15%
SRR1168889_1	9289024	poli-A	33429	0.36%
		poli-T	117193	1.26%
SRR1168889_2	9289024	poli-A	91228	0.98%
		poli-T	28071	0.30%
SRR1168891_1	27704351	poli-A	244181	0.88%
		poli-T	954116	3.44%
SRR1168891_2	27704351	poli-A	954116	3.44%
		poli-T	217652	0.79%
SRR1168893_1	10065144	poli-A	18495	0.18%
		poli-T	40128	0.40%
SRR1168893_2	10065144	poli-A	25582	0.25%
		poli-T	12320	0.12%
SRR1168894_1	10914994	poli-A	11980	0.11%
SRR1168938_1	22189724	poli-A	1044076	4.71%
		poli-T	1845185	8.32%
SRR1168938_2	22189724	poli-A	1590304	7.17%
		poli-T	970896	4.38%

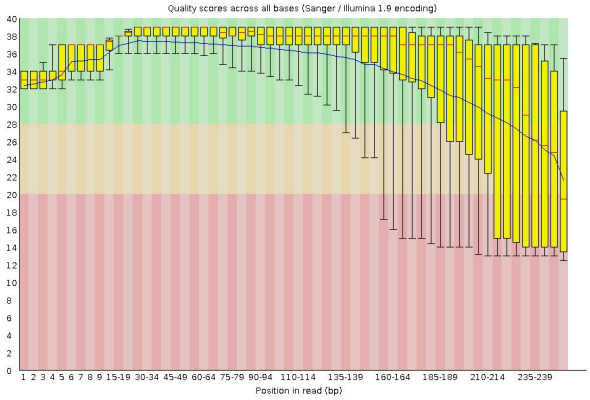
A) SRR4449814_2



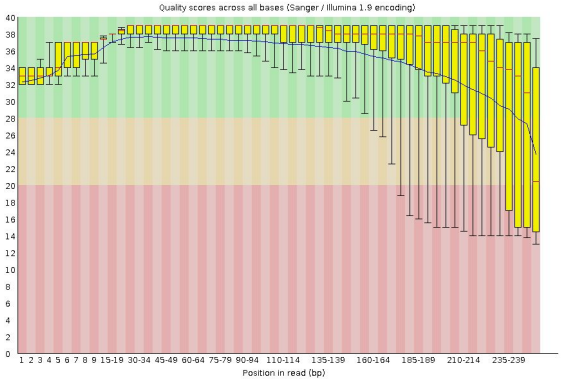
B) SRR4449815_2



C) SRR4449939_2



D) SRR4449940_2



E) SRR4449941_2

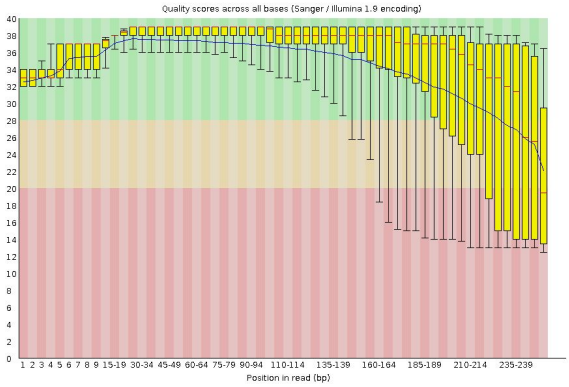
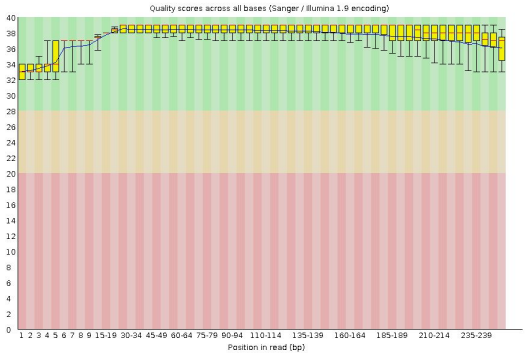


Figura 1: resultados do FASTQC exemplificando bibliotecas com bases de baixa qualidade da extremidade 3'. A faixa verde representa qualidades de 28 a 40, a faixa amarela qualidade entre 20 e 27 e a faixa vermelha qualidades entre 0 e 19. As barras representam o primeiro e o quarto quartis e a faixa amarela representa o segundo e terceiro quartis separados por uma linha vermelha.

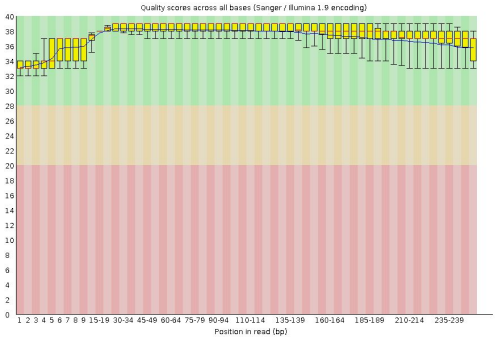
Após os procedimentos de limpeza as sequências contaminantes (poli-A ou T) foram eliminadas, assim como as bases de baixa qualidade da extremidade 3', a Figura 2 ilustra a melhora na qualidade de bases das bibliotecas SRR4449814_2, SRR4449815_2, SRR4449939_2, SRR4449940_2 e SRR4449941_2.

A Tabela 3 sumariza a quantidade de *reads* perdidos por biblioteca pelo processo de limpeza; e os resultados gerais para as montagens dos transcriptomas estão resumidos na Tabela 4.

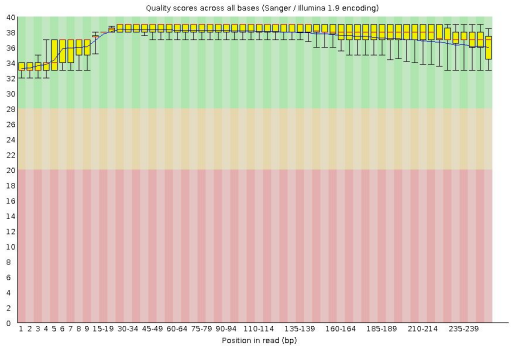
A) SRR4449814_2



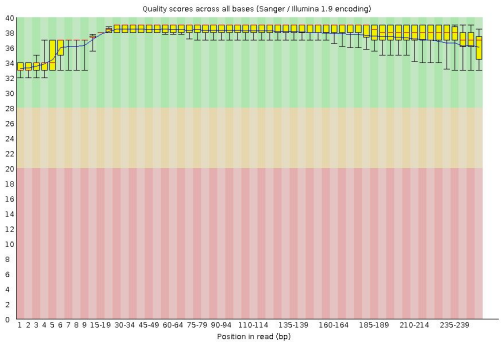
B) SRR4449815_2



C) SRR4449939_2



D) SRR4449940_2



E) SRR4449941_2

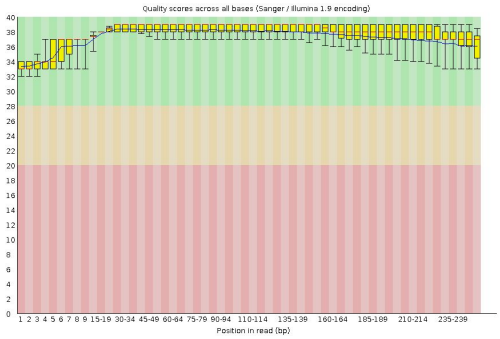


Figura 2: resultados do FASTQC exemplificando a melhora na qualidade de bases para as bibliotecas com bases de baixa qualidade da extremidade 3'. A faixa verde representa qualidades de 28 a 40, a faixa amarela qualidade entre 20 e 27 e a faixa vermelha qualidades entre 0 e 19. As barras representam o primeiro e o quarto quartis e a faixa amarela representa o segundo e terceiro quartis separados por uma linha vermelha.

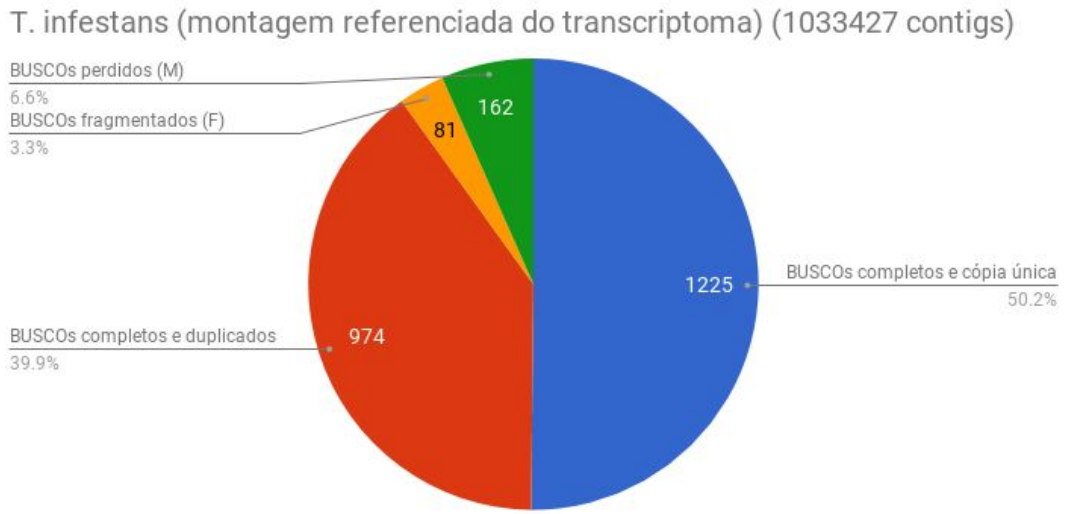
Tabela 3: Quantidade de *reads* inicial, antes da limpeza, e final, após as etapas de limpeza, para cada uma das bibliotecas utilizadas para as montagens do transcriptoma de *Triatoma infestans*.

Biblioteca	*Reads* brutos	Total de *reads* limpos	Perda percentual
SRR1168882_1	10957686	9854271	10.1%
SRR1168882_2	10957686	9474179	13.5%
SRR1168885_1	9191270	6634038	27.8%
SRR1168885_2	9191270	6084728	33.8%
SRR1168888_1	12025475	10908739	9.3%
SRR1168888_2	12025475	10537928	12.4%
SRR1168889_1	9289024	7085331	23.7%
SRR1168889_2	9289024	6431814	30.8%
SRR1168890_1	11789324	10650925	9.7%
SRR1168890_2	11789324	10271196	12.9%
SRR1168891_1	27704351	18378801	33.7%
SRR1168891_2	27704351	16663281	39.9%
SRR1168892_1	10735306	9281742	13.5%
SRR1168892_2	10735306	8716421	18.8%
SRR1168893_1	10065144	8009754	20.4%
SRR1168893_2	10065144	7404053	26.4%
SRR1168894_1	10914994	9009957	17.5%
SRR1168894_2	10914994	8369968	23.3%
SRR1168938_1	22189724	14733427	33.6%
SRR1168938_2	22189724	13392088	39.6%
SRR4427078_1	4698102	4044237	13.9%
SRR4427078_2	4698102	3789205	19.3%
SRR4427079_1	7033068	6064931	13.8%
SRR4427079_2	7033068	5735888	18.4%
SRR4449814_1	923329	815809	11.6%
SRR4449814_2	923329	757176	18.0%
SRR4449815_1	2407024	2050332	14.8%
SRR4449815_2	2407024	1927778	19.9%
SRR4449939_1	2178856	1866796	14.3%
SRR4449939_2	2178856	1750939	19.6%
SRR4449940_1	1954565	1706373	12.7%
SRR4449940_2	1954565	1598122	18.2%
SRR4449941_1	2200700	1897446	13.8%
SRR4449941_2	2200700	1818631	17.4%

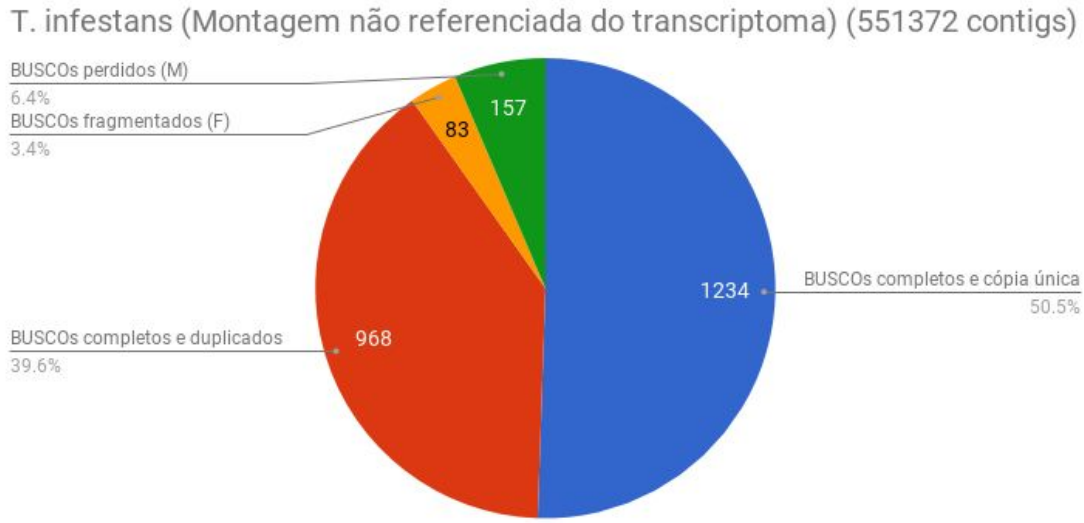
Tabela 4: Dados gerais das montagens de transcriptoma de *Triatoma infestans*.

Montagem	Comprimento total	Número de contigs	N50	N50n
Montagem referenciada	540358474	1033426	597	198144
Montagem não-referenciada	315113485	551372	617	108536

A avaliação da completude das montagens dos transcriptomas pelo software BUSCO indicou que cerca de 93% dos genes cópia simples universalmente conservados entre os endopterigotos foram detectados nas montagens do transcriptoma, porém com um elevado nível de duplicação, 968 genes na montagem não referenciada e 974 genes na montagem que utilizou um genoma de referência. Os resultados da avaliação da completude das montagens pelo software BUSCO estão presentes na Figura 3.



A



B

Figura 3: Resultados do BUSCO para os transcriptomas montados.

4.2 Avaliação da montagem do genoma

Mesmo utilizando parâmetros de *Drosophila melanogaster* para a predição com o AUGUSTUS, o BUSCO foi capaz de detectar 93% dos genes de endopterigotos universalmente conservados e cópia simples. Esse resultado é de grande valor pois representa um padrão de qualidade a ser obtido pela predição gênica, é de se esperar que a predição utilizando parâmetros gerados especificamente para *Triatoma infestans* tenha um resultado melhor ou igual ao obtido utilizando-se parâmetros gerados para outro organismo. O resultado do BUSCO para a montagem do genoma é apresentado na Figura 4.

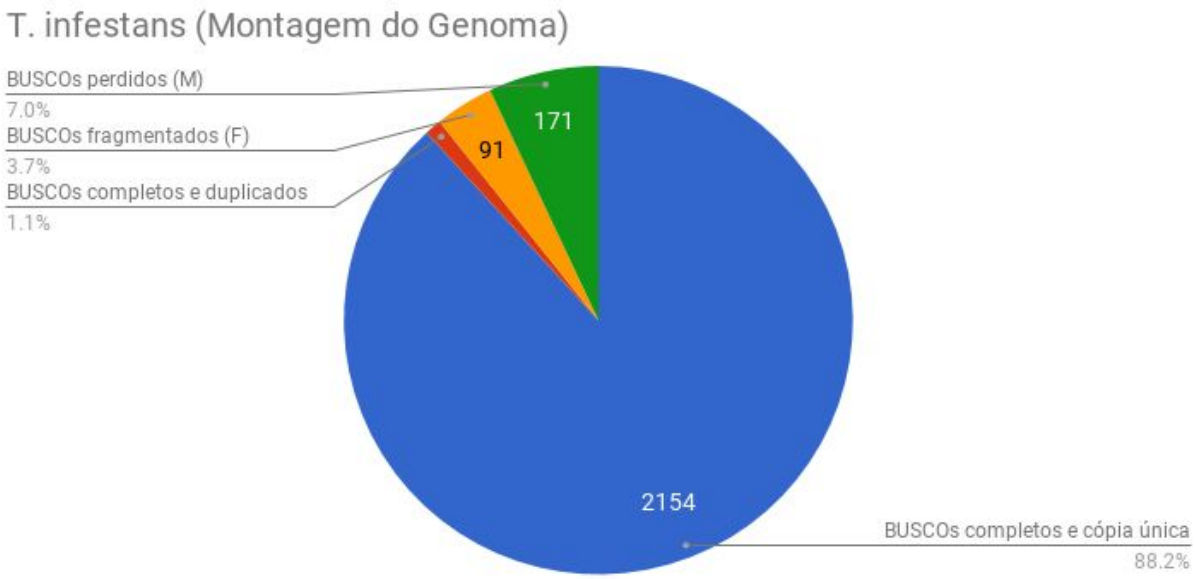


Figura 4: Resultado do BUSCO para o genoma montado ainda sem a predição dos genes. O número de genes conservados que foram encontrados completos e duplicados foi de 26.

4.2 Conjunto de genes de referência

O procedimento de alinhamento dos transcritos contra o genoma gerou um total de 1986 alinhamentos e destes: 1147 foram alinhamentos perfeitos contra o genoma; 595 tiveram o alinhamento truncado na extremidade 3'; 48 alinhamentos estavam fragmentados ou truncados na extremidade 5'; e 196 sequências alinharam em mais de uma região do genoma. Um esquema exemplificando alinhamentos completo, truncado na extremidade 3', truncado na extremidade 5' e fragmentado é ilustrado na Figura 5.

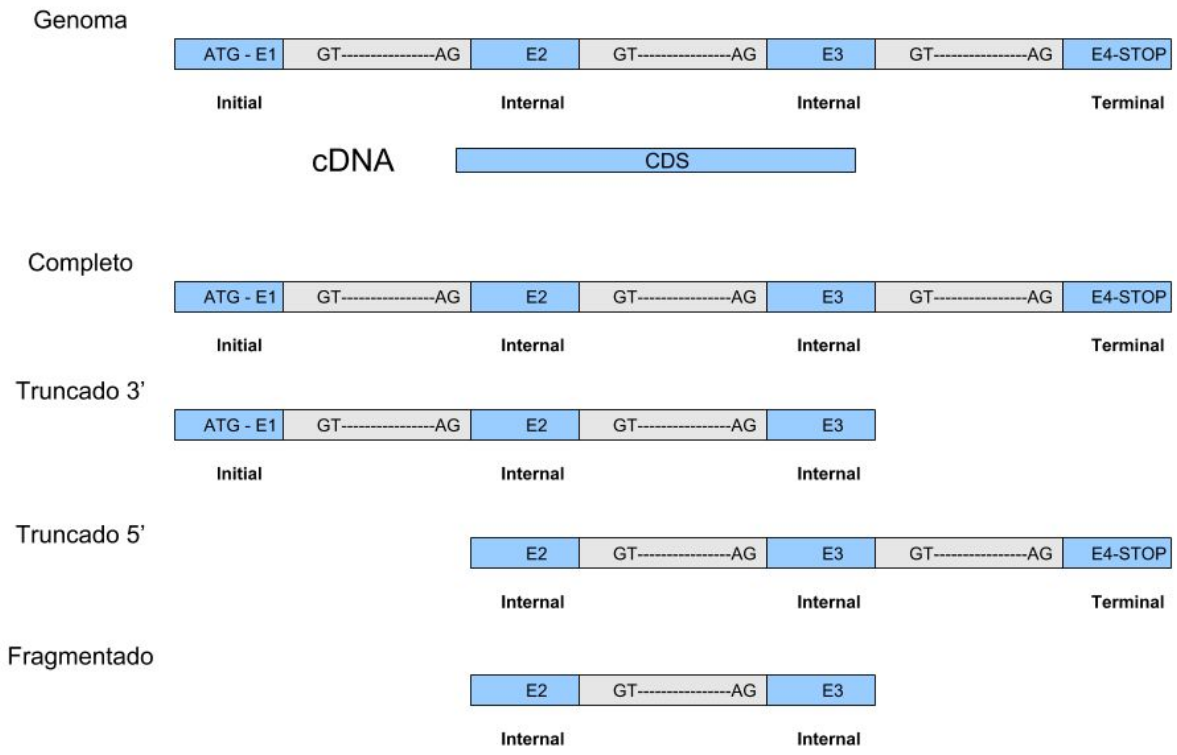


Figura 5: Representações dos quatro tipos de alinhamento de CDS reportados pelo software utilizado. "Genoma" representa a sequência genômica anotada, e "cDNA" representa o CDS proveniente dos dados de sequenciamento de cDNA que deve ser alinhado contra o genoma.

Apenas os 1147 alinhamentos considerados perfeitos foram incluídos no grupo de genes de treinamento.

4.3 Predições gênicas preliminares

4.3.1 Treinamento do GENEID

Os números de sequências de 60 nucleotídeos obtidos a partir do conjunto de genes de referência para cada tipo de sinal estão relacionados na Tabela 5.

Tabela 5: Número de sequências de 60 nucleotídeos obtidas para cada um dos sinais encontrados entre as sequências do conjunto de genes de referência.

Sequência	Número	Tipo de Sinal
AC	2	Aceptor de *splicing*
AG	6797	Aceptor de *splicing*
ATG	1146	Início da tradução (Codon inicial - Metionina)
AT	6	Aceptor de *splicing*
GC	26	Doador de *splicing*
GT	6770	Doador de *splicing*
STOP	1146	Término de tradução (Codon de terminação)

Os aceptores de sequência AG e os doadores de sequência GT foram os mais abundantes no genoma de *Triatoma infestans*. A quantidade de sítios de *splicing* com outras sequências (AC, AT e GC) não seria suficiente para se gerar os modelos estatísticos necessários, portanto estes sinais não foram utilizados.

Os intervalos das sequências de 60 nucleotídeos que foram consideradas informativas e os seus offsets estão presentes na Tabela 6. Estes intervalos foram utilizados para gerar modelos de hidden markov de ordens 3, 2 e 3, respectivamente para os sinais AG, ATG e GT.

Tabela 6: Fatores de correção (*offset*) para os sinais utilizados.

Sítio	Início	Fim	Comprimento	Offset
AG	10	29	19	18
ATG	26	34	8	3
GT	28	36	8	0

Para os sensores de composição foram obtidos um total de 7.596.159 bp para a os dados de exons e 33.082.096 para os dados de introns. Isso nos permitiria utilizar um modelo de Hidden Markov de ordem 8 para os dados de introns e de ordem 7 para os dados de exons. Entretanto, como é necessário que ambos os modelos sejam da mesma ordem, modelos de ordem 7 foram utilizados.

A escolha do tamanho máximo de intron a ser utilizado nas configurações do programa GENEID foi feita considerando os introns mapeados no conjunto gênico de referência. Um gráfico de histograma (Figura 6.A) e um gráfico de soma acumulada da frequência dos introns (Figura 6.B) ilustram os tamanhos dos introns. Ambos os gráficos mostraram que a grande maioria (99,25%) dos introns era menor do que 51.200 bp o que justificou a escolha de 60.000 como o valor de tamanho máximo dos introns.

A

B

Figura 6: Histograma e gráfico de soma acumulada do tamanhos dos introns do conjunto de genes de referência.

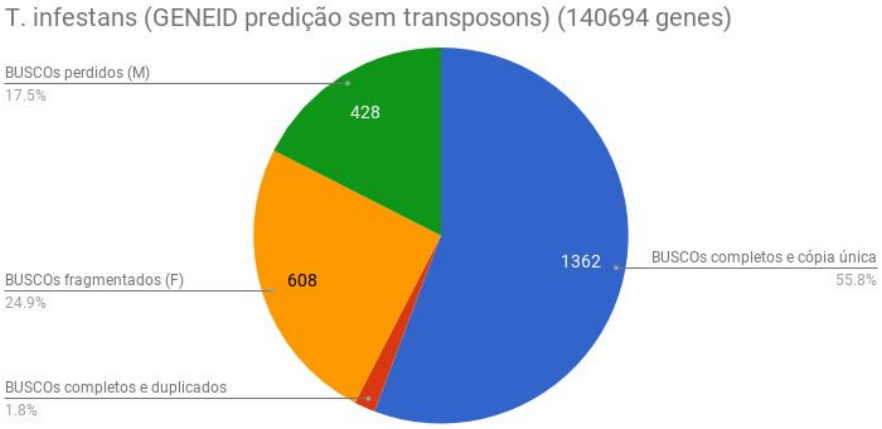
Após a obtenção dos modelos de Hidden Markov para os sinais e composição gênicos e da escolha do tamanho máximo de íntron, trinta e seis predições gênicas foram feitas no conjunto de genes de referência, variando-se duas constantes internas (oWf e eWf). Os resultados de CC, SNSPe, SNSPg, SATG, SACC e SDonor obtidos para predições são ilustrados na Tabela 7. Neste procedimento de otimização os valores de oWF e eWF escolhidos para a predição foram respectivamente 0.7 e -5 pois o melhor valor de "sensibilidade x especificidade" (SNSPg) para genes foi obtido com estes valores para oWF e eWF.

A predição inicial gerou 175842 modelos gênicos, dos quais 35147 foram classificados como transposons e removidos e 17246 possuíam pelo menos um domínio conservado do banco de dados Pfam-A e foram incluídos na predição final. Os resultados do BUSCO para a predição com o GENEID antes e depois dos filtros estão ilustrados na Figura 7, a utilização dos filtros causou a perda de 235 genes conservados, porém reduziu em 10 vezes o número total de genes preditos.

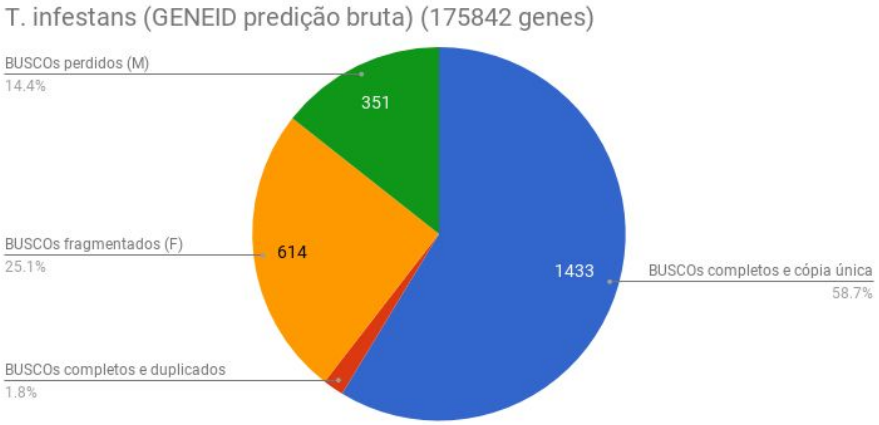
Tabela 7: Resumo dos resultados das 36 predições utilizadas para otimizar os parâmetros estatísticos do GENEID.

oWF	eWF	CC	SNSPe	SNSPg	SATG	SAcc	SDonor
0.55	-5	0.614	0.431	0.326	0.669	0.955	0.959
0.55	-4.5	0.574	0.37	0.294	0.643	0.951	0.955
0.55	-4	0.534	0.316	0.262	0.625	0.945	0.947
0.55	-3.5	0.492	0.271	0.241	0.599	0.94	0.94
0.6	-5	0.624	0.481	0.346	0.674	0.956	0.965
0.6	-4.5	0.591	0.415	0.314	0.651	0.953	0.962
0.6	-4	0.55	0.353	0.279	0.632	0.949	0.957
0.6	-3.5	0.51	0.299	0.251	0.606	0.944	0.949
0.65	-5	0.629	0.527	0.366	0.676	0.955	0.968
0.65	-4.5	0.593	0.459	0.334	0.656	0.951	0.965
0.65	-4	0.559	0.394	0.301	0.635	0.952	0.962
0.65	-3.5	0.517	0.331	0.265	0.61	0.946	0.955
0.7	**-5**	**0.625**	**0.556**	**0.376**	**0.659**	**0.95**	**0.967**
0.7	-4.5	0.595	0.494	0.344	0.646	0.948	0.965
0.7	-4	0.558	0.427	0.312	0.628	0.948	0.961
0.7	-3.5	0.518	0.359	0.275	0.604	0.943	0.958
0.75	-5	0.612	0.566	0.37	0.622	0.937	0.96
0.75	-4.5	0.584	0.508	0.343	0.618	0.938	0.96
0.75	-4	0.551	0.445	0.314	0.599	0.934	0.957
0.75	-3.5	0.512	0.378	0.284	0.584	0.93	0.953
0.8	-5	0.592	0.552	0.356	0.582	0.922	0.951
0.8	-4.5	0.566	0.502	0.337	0.572	0.921	0.949
0.8	-4	0.534	0.442	0.313	0.563	0.917	0.944
0.8	-3.5	0.5	0.38	0.283	0.559	0.913	0.941
0.85	-5	0.566	0.522	0.332	0.533	0.895	0.929
0.85	-4.5	0.54	0.478	0.324	0.527	0.894	0.929
0.85	-4	0.513	0.425	0.299	0.525	0.889	0.924
0.85	-3.5	0.48	0.369	0.274	0.52	0.886	0.919

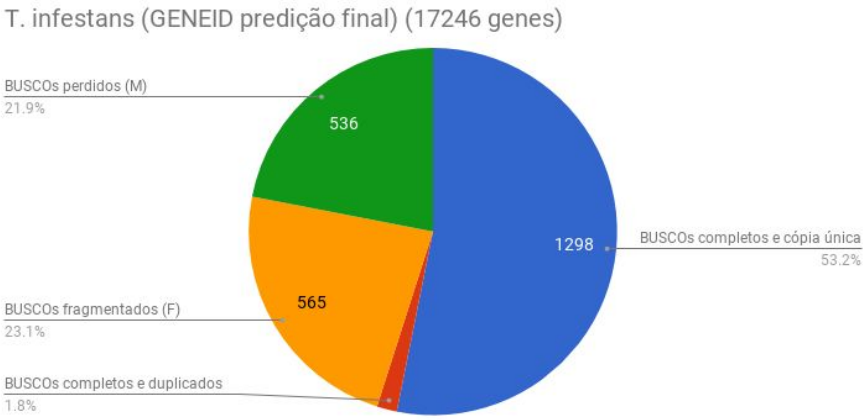
Onde: CC é o produto entre a sensibilidade e a especificidade da predição de bases; SNSPe é o produto entre a sensibilidade e a especificidade da predição de exons; SNSPg é o produto entre a sensibilidade e a especificidade da predição de genes; e SATG, SAcc e SDonor, são as especificidades para os sítios ATG, aceptor do *splicing* e doador do *splicing* respectivamente.



A



B



C

Figura 7: Resultados do BUSCO para a predição com o GENEID antes e após os filtros de transposons e domínios conservados. O número de genes encontrados completos e duplicados foi de 44 para os gráficos A e B e 43 para o gráfico C.

4.3.2 Predição preliminar com o MAKER

A predição procedida por colaboradores utilizando a metodologia MAKER gerou um total de 16414 modelos gênicos, porém apenas cerca de 57% dos genes universalmente conservados entre os endopterigotos foram encontrados nesta predição (Figura 8).

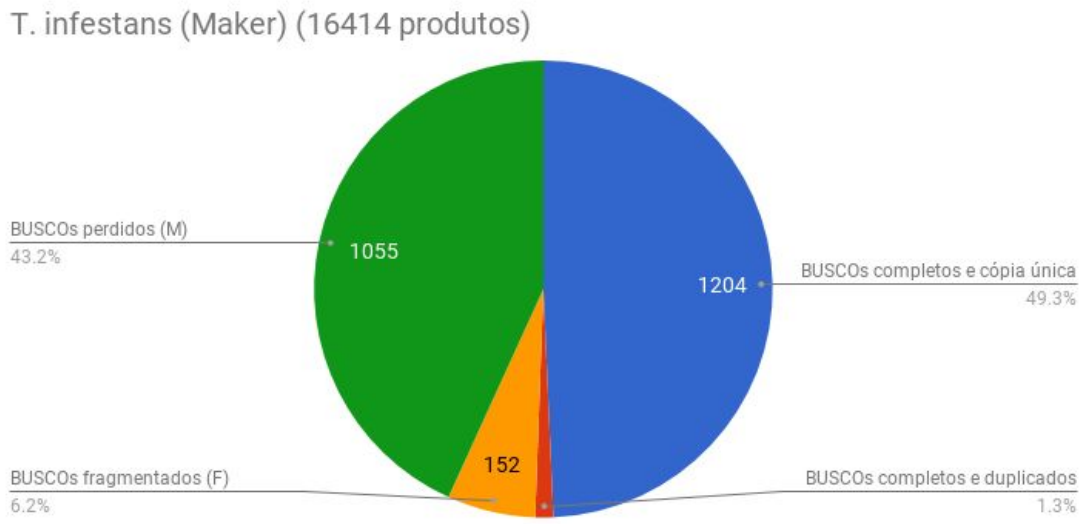


Figura 8: Resultados do BUSCO para a predição preliminar com o MAKER. Trinta e um genes conservados foram encontrados completos e duplicados.

4.3.3 Predições preliminares com o AUGUSTUS

A primeira predição teve a seguinte estimativa de sensibilidade e especificidade (Figura 9).

```
*******      Evaluation of gene prediction     *******

----------------------------------------------\
                 | sensitivity | specificity |
----------------------------------------------|
nucleotide level |       0.932 |       0.626 |
----------------------------------------------/

-----------------------------------------------------------------------------------------------------
--\
           |  #pred |  #anno |      |    FP = false pos. |    FN = false neg. |             |
|
           | total/ | total/ |   TP |--------------------|--------------------| sensitivity |
specificity |
           | unique | unique |      | part | ovlp | wrng | part | ovlp | wrng |             |
|
-----------------------------------------------------------------------------------------------------
--|
           |        |        |      |                386 |                135 |             |
|
exon level |    918 |    667 |  532 | ------------------ | ------------------ |       0.798 |
0.58 |
           |    918 |    667 |      |   56 |    3 |  327 |   56 |    3 |   76 |             |
|
-----------------------------------------------------------------------------------------------------
--/

-------------------------------------------------------------------------\
transcript | #pred | #anno |   TP |   FP |   FN | sensitivity | specificity |
-------------------------------------------------------------------------|
gene level |   200 |   100 |   34 |  166 |   66 |        0.34 |        0.17 |
-------------------------------------------------------------------------/

-------------------------------------------------------------------\
            UTR | total pred | CDS bnd. corr. |   meanDiff | medianDiff |
-------------------------------------------------------------------|
            TSS |         20 |              0 |         -1 |         -1 |
            TTS |         24 |              0 |         -1 |         -1 |
-------------------------------------------------------------------|
            UTR | uniq. pred |    unique anno |      sens. |      spec. |
-------------------------------------------------------------------|
                |  true positive = 1 bound. exact, 1 bound. <= 20bp off |
 UTR exon level |          0 |              0 |       -nan |       -nan |
-------------------------------------------------------------------|
 UTR base level |          0 |              0 |       -nan |       -nan |
-------------------------------------------------------------------/
nucUTP= 0 nucUFP=0 nucUFPinside= 0 nucUFN=0
# total time: 181
# command line:
# ./augustus/bin/augustus --species=t_infestans Ti_genes.gb.test
```

Figura 9: Resultado do software de avaliação de sensibilidade e especificidade do AUGUSTUS para a predição preliminar.

Os resultados do BUSCO para esta predição preliminar estão ilustrados na Figura 10.

Um grande número de produtos foram preditos pelo *software*, porém apenas 11% dos genes conservados não foram encontrados, o que indicou uma maior sensibilidade do AUGUSTUS.

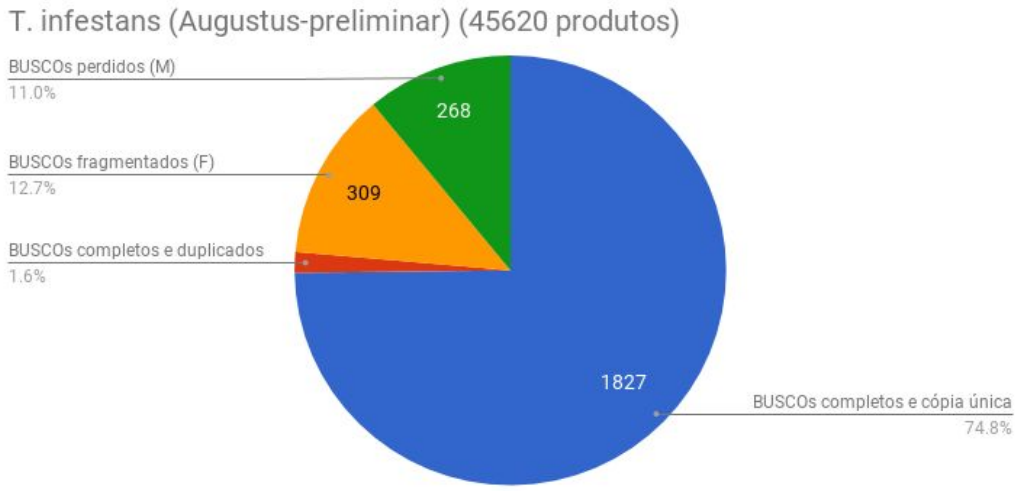


Figura 10: Resultados do BUSCO para a predição preliminar com o AUGUSTUS. O número de genes encontrados completos e duplicados foi de 38.

4.4 Otimização do AUGUSTUS

O procedimento de otimização estatística obteve um aumento na sensibilidade no nível de gene de 0.34 para 0.49 e um aumento na especificidade de 0.17 para 0.266. Os resultados da validação do próprio AUGUSTUS estão na Figura 11.

```
*******      Evaluation of gene prediction     *******

----------------------------------------------\
                  | sensitivity | specificity |
----------------------------------------------|
nucleotide level |       0.954 |       0.655 |
----------------------------------------------/

-----------------------------------------------------------------------------------------------------------\
           |  #pred |  #anno |      |    FP = false pos. |    FN = false neg. |             |             |
           | total/ | total/ |   TP |--------------------|--------------------| sensitivity | specificity |
           | unique | unique |      | part | ovlp | wrng | part | ovlp | wrng |             |             |
-----------------------------------------------------------------------------------------------------------|
           |        |        |      |                345 |                 89 |             |             |
exon level |    923 |    667 |  578 | ------------------ | ------------------ |       0.867 |       0.626 |
           |    923 |    667 |      |   34 |    1 |  310 |   34 |    1 |   54 |             |             |
-----------------------------------------------------------------------------------------------------------/

-------------------------------------------------------------------------\
transcript | #pred | #anno |   TP |   FP |   FN | sensitivity | specificity |
-------------------------------------------------------------------------|
gene level |   184 |   100 |   49 |  135 |   51 |        0.49 |       0.266 |
-------------------------------------------------------------------------/

-----------------------------------------------------------------------\
            UTR | total pred | CDS bnd. corr. |   meanDiff | medianDiff |
-----------------------------------------------------------------------|
            TSS |         13 |              0 |         -1 |         -1 |
            TTS |         18 |              0 |         -1 |         -1 |
-----------------------------------------------------------------------|
            UTR | uniq. pred |    unique anno |      sens. |      spec. |
-----------------------------------------------------------------------|
                |  true positive = 1 bound. exact, 1 bound. <= 20bp off |
 UTR exon level |          0 |              0 |       -nan |       -nan |
-----------------------------------------------------------------------|
 UTR base level |          0 |              0 |       -nan |       -nan |
-----------------------------------------------------------------------/
nucUTP= 0 nucUFP=0 nucUFPinside= 0 nucUFN=0
# total time: 196
# command line:
# ./augustus/bin/augustus --species=t_infestans Ti_genes.gb.test
```

Figura 11: Resultado do software de avaliação de sensibilidade e especificidade do AUGUSTUS para a predição otimizada.

Para as evidências de intron os dados de sequenciamento de transcriptomas geraram 258154 alinhamentos de intron, destes, 235261 foram considerados introns de regiões provavelmente codificantes pelo AUGUSTUS, porém apenas 104879 junções exon-exon foram confirmadas por dados de transcriptoma e foram utilizadas como evidências para a predição final com o AUGUSTUS.

Os resultados dos dois processos de otimização do AUGUSTUS foram avaliados

utilizando-se o BUSCO e os resultados são ilustrados na Figura 12 e 13. A maior sensibilidade do AUGUSTUS após a otimização estatística permitiu que um número menor de genes conservados fosse perdido (9.6% no otimizado contra 11% no preliminar) e que um menor número total de genes fosse predito (34628 no otimizado contra 45620 no preliminar). A utilização de evidências de transcriptoma reduziu (melhorou) ainda mais o número total de genes preditos, para 24489, e de genes conservados perdidos, para 6.1%. A Tabela 8 apresenta os resultados numéricos das avaliações do BUSCO para as três predições.

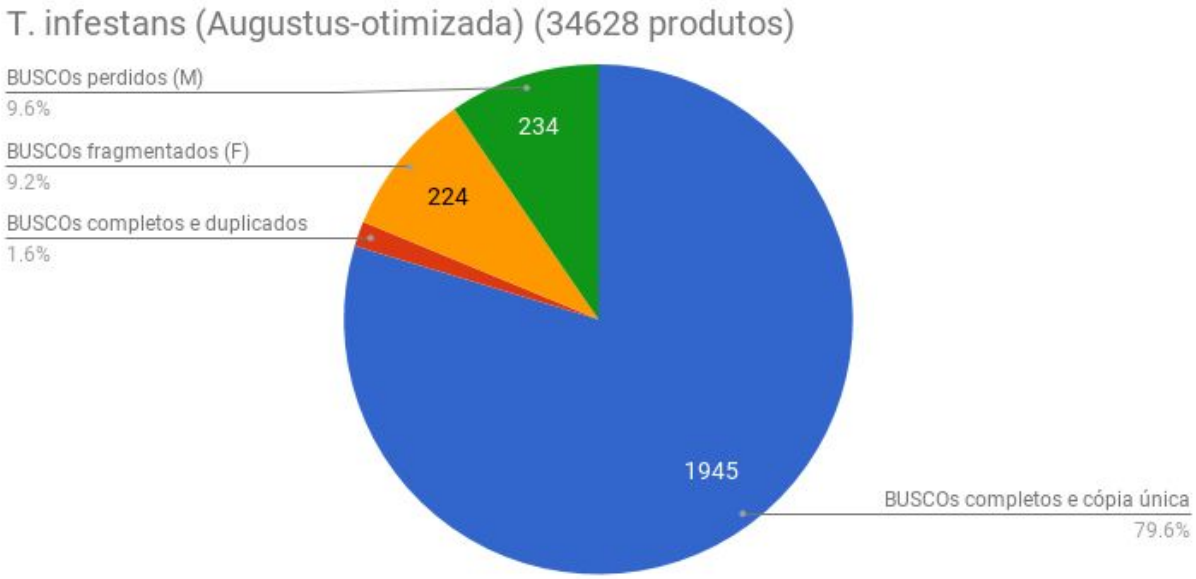


Figura 12: Resultados do AUGUSTUS após a otimização estatística. O número de genes encontrados completos e duplicados foi de 39.

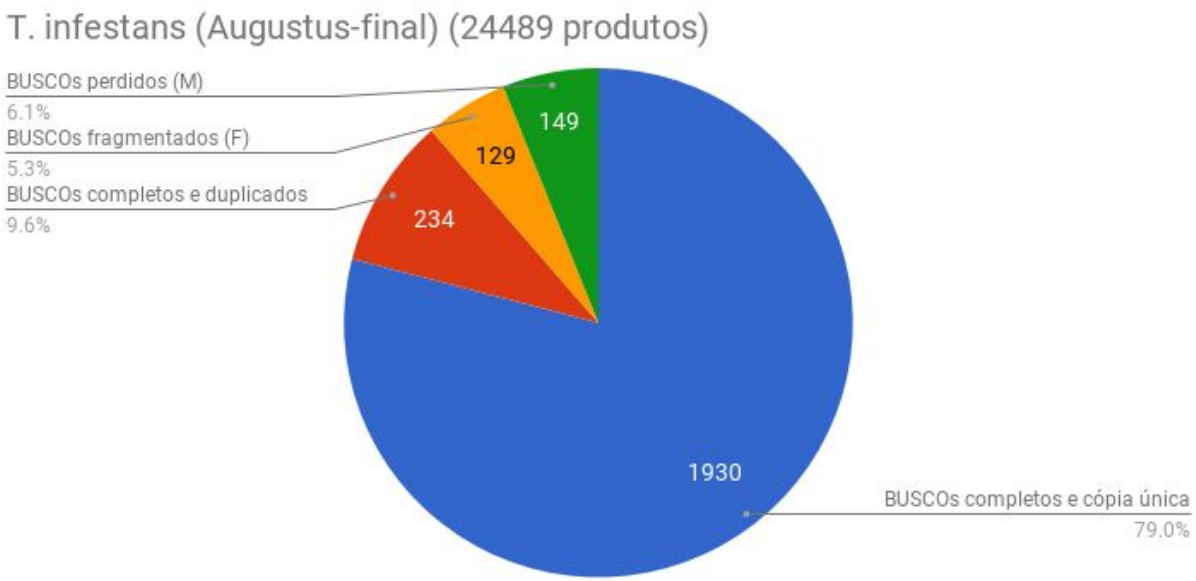


Figura 13: Resultado final do AUGUSTUS após a otimização estatística e adição de evidências de transcriptoma.

Além de melhorar a qualidade geral da predição a utilização de dados de transcriptoma permitiu a predição de variantes de *splicing*, que são interpretados pelo programa BUSCO como "genes duplicados". Desta forma para uma correta avaliação foi necessário considerar apenas um transcrito para cada locus gênico, assim o resultado da classe completos e duplicados no BUSCO não foi de 234 (Figura 13) mas sim de 41. O gráfico a seguir (Figura 14) representa os resultados do BUSCO quando apenas um transcrito é utilizado para cada locus gênico.

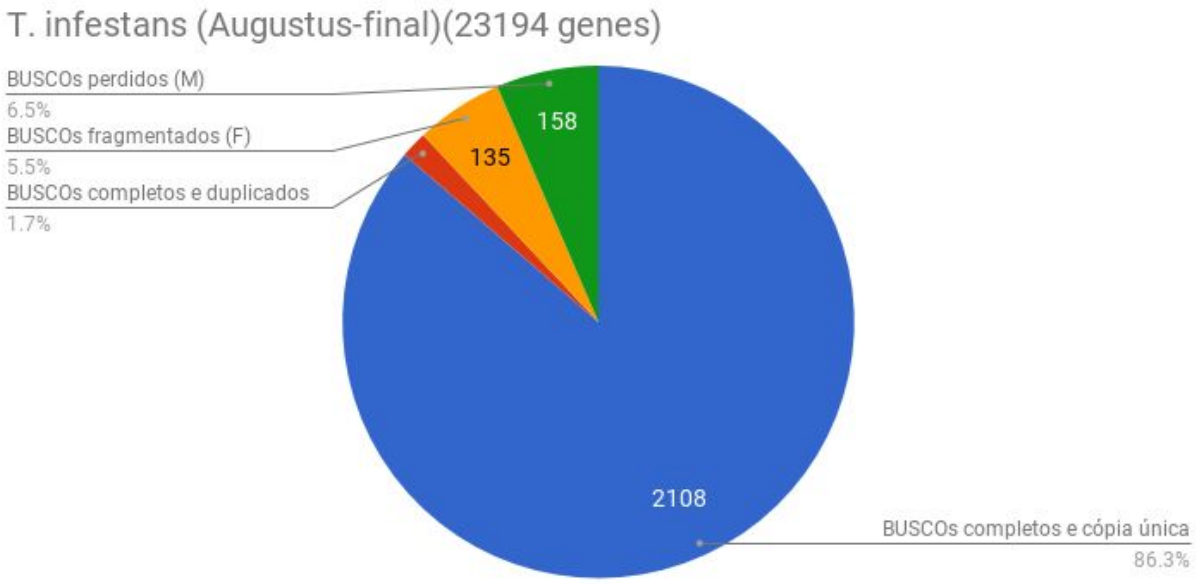


Figura 14: Resultado final da predição com AUGUSTUS. Esta análise foi feita com apenas um transcrito para cada locus gênico. O número de genes encontrados completos e duplicados foi de 41.

Tabela 8: Tabela comparativa das três predições realizadas com o AUGUSTUS.

Predição	BUSCOs completos e cópia única (S)	BUSCOs completos e duplicados (D)	BUSCOs fragmentados (F)	BUSCOs perdidos (M)	Total de preditos
Augustus (preliminar)	1827	38	309	268	45620
Augustus (otimizada)	1945	39	224	234	34628
Augustus (final)	1930	234	129	149	24489
Augustus (final) (um produto por locus)	2108	41	135	158	23194

4.5 Comparação com outras predições gênicas de hemípteros.

Os resultados do BUSCO para os genomas de hemípteros utilizados para comparação estão presentes nos gráfico e tabelas abaixo. A predição de *Acyrthosiphon pisum* é a que possui o maior número total de preditos (36195), a de *Oncopeltus fasciatus* é a que apresenta o maior número de genes conservados fragmentados (672) e perdidos (445), o menor número de genes preditos é o da predição para *Cimex lectularius* (14210), tendo o conjunto de genes de *Halyomorpha halys*[6], Figura 15 o maior número de genes duplicados (752) e o menor número de genes perdidos (100). A Tabela 9 apresenta um resumo dos resultados do BUSCO para todas as predições gênicas avaliadas.

Figura 15: Resultados do BUSCO para o genoma de *Halyomorpha halys*

[6] O conjunto de genes de *Halyomorpha halys* é constituído de todos dados de proteínas depositados para esta espécie no banco de dados de proteínas do NCBI.

Tabela 9: Resumo dos resultados do BUSCO para os genomas de hemípteros avaliados e seus respectivos totais de preditos.

Nome	BUSCOs completos (C)	BUSCOs fragmentados (F)	BUSCOs perdidos (M)	Total de preditos
Acyrthosiphon pisum	2152	88	202	36195
Aphis glycines	2041	103	298	19182
Cimex lectularius	2003	228	211	14210
Diuraphis noxia	1797	284	361	25987
Myzus cerasi	1868	218	356	28688
Myzus persicae	2171	69	202	30127
Oncopeltus fasciatus	1325	672	445	19615
Rhodnius prolixus	1849	233	360	15075
Rhopalosiphum padi	2119	114	209	27817
Triatoma infestans	**2164**	**129**	**149**	**23194**
Média ± Desvio Padrão	1949±259	214±177	279±97	24009±7034

5. DISCUSSÃO

5.1 Montagem do genoma

O BUSCO foi capaz de detectar no genoma de *T. infestans* 93% dos ortólogos universalmente conservados para insetos endopterigotos, mesmo utilizando-se parâmetros de outra espécie (*Drosophila melanogaster*) para identificar os genes com o programa AUGUSTUS (internamente e automaticamente). Isso foi possível também pela utilização de ortólogos de outras espécies como evidência para o BUSCO, porém esse tipo de estratégia se limita a identificar os genes conservados.

Os resultados obtidos pelo BUSCO para a montagem (2154 BUSCOs completos e cópia única (S), 26 BUSCOs completos e duplicados (D), 91 BUSCOs fragmentados (F), 171 BUSCOs perdidos (M)) são comparáveis com os obtidos pelas predições gênicas dos outros organismos avaliados neste trabalho, o que indica que o genoma foi montado com uma qualidade compatível com o presente estado da arte em genômica.

A utilização de análises como BUSCO para a avaliação da montagem do genoma é de grande importância também serve como uma meta de qualidade para as predições gênicas, uma predição gênica só pode ser considerada adequada quando se detecta um número de genes conservados próximo ou maior do que os encontrados pelo BUSCO no genoma.

5.2 Transcriptomas

É muito comum que as bibliotecas reversas de um sequenciamento Illumina pareado tenham baixa qualidade na extremidade 3' pois o grande número de ciclos efetuados prejudica a qualidade da leitura de bases, o procedimento de limpeza foi capaz de remover estas bases de baixa qualidade de modo que isso não interferisse nos procedimentos posteriores.

Algumas bibliotecas tiveram uma considerável perda de sequências devido a elevada exigência de qualidade de base (qualidade média mínima de 30 em um janela de 4 nucleotídeos) e em alguns casos devido a contaminação com sequências repetitivas, porém devido a grande quantidade de dados disponíveis se optou por favorecer a qualidade das sequências brutas ao invés de sua quantidade de modo a elevar a confiabilidade das evidências para a predição.

A montagem dos transcriptomas sozinha foi capaz de detectar 93.6% dos genes universalmente conservados de insetos endopterigotos, porém gera uma grande quantidade de BUSCOs duplicados pois é rica em variantes de splicing além de ser pouco específica para genes codificantes. A montagem sem referência, que gerou o menor número de isotigs, gerou 551372 sequências, o que é pelo menos 20 vezes maior do que número de genes encontrado para outros hemípteros.

Assim como a avaliação da montagem do genoma, a avaliação pelo BUSCO das montagens de transcriptomas ajuda a definir uma meta para a qualidade da predição gênica. Espera-se detectar um número de genes conservados próximo ou maior do que as montagens de transcriptoma porém com um número total de sequências preditas

mais próximo do número de genes da espécie.

5.3 Geração do conjunto de sequências codificantes de referência

O laboratório possui um procedimento para encontrar genes cópia única de alta qualidade em alinhamentos contra o genoma a partir de sequências de cDNA utilizando o software SIM4 (FLOREA et al., 1998). Esse procedimento foi utilizado no genoma de *Rhodnius prolixus* (MESQUITA et al., 2015 [Apêndice]), e separa os alinhamentos em categorias relevantes para o treinamento dos programas de predição, baseado na completude do alinhamento, número de cópias no genoma, integridade das interseções exon/intron, início e fim das regiões codificantes. Para o genoma de *Triatoma infestans* o conjunto de genes de referência conteve somente as sequências que alinharam completamente (da extremidade 5' até a extremidade 3').

5.4 Predições preliminares

Inicialmente, o resultado total de genes preditos, por ser numericamente próximo nas predições com o GENEID (17.246) e o MAKER (16.414), fazia parecer que estas duas predições eram equivalentes. Entretanto quando foram analisadas para o conteúdo de genes conservados pelo BUSCO (Figura 16) foi observado que um grande número de preditos identificados com o GENEID estavam ausentes na predição realizada pelo MAKER (695 BUSCOs) e que uma considerável parte dos preditos do MAKER também estava ausente na predição pelo GENEID (176 BUSCOs). A opção mais parcimoniosa foi utilizar a predição mais sensível (isto é, aquela que predisse o maior número de genes corretamente), que foi a do AUGUSTUS, pois apenas 89 BUSCOs preditos pelo GENEID e MAKER não foram encontrados pelo AUGUSTUS, enquanto 180 BUSCOS foram encontrados apenas pelo AUGUSTUS, e utilizar abordagens de otimização capazes de melhorar a especificidade desta predição.

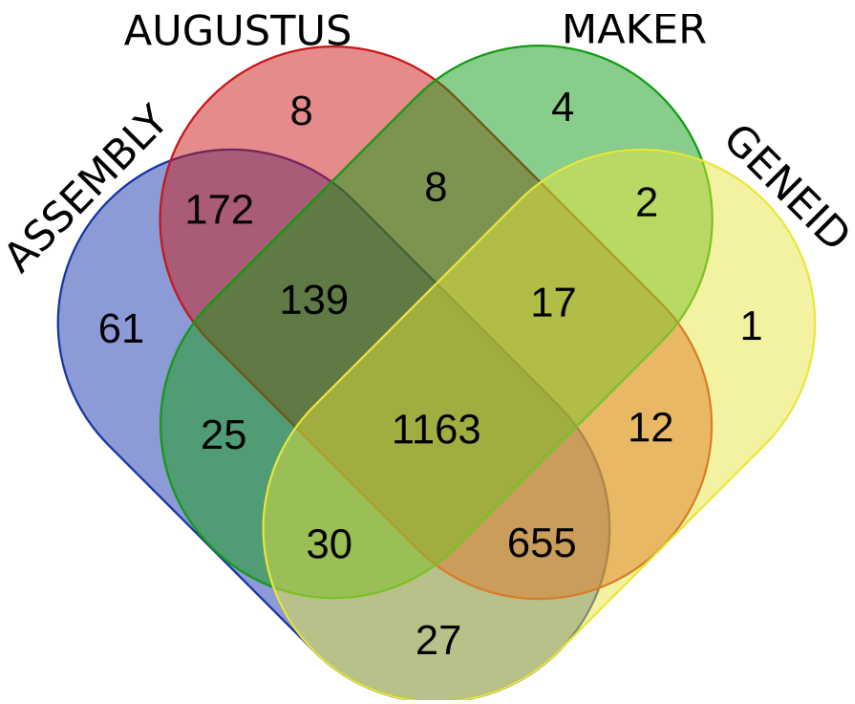


Figura 16: Comparação do número de BUSCOs econtrados (de um total de 2442) para as predições preliminares, com o AUGUSTUS, o MAKER e o GENEID. Os resultados do BUSCO para a montagem do genoma sem predição ("ASSEMBLY") são apresentados como uma referência para a comparação.

As três metodologias de predição obtiveram resultados bastante distintos, com um destaque para a qualidade e simplicidade da utilização do AUGUSTUS.

5.5 Otimização do AUGUSTUS

Os dois passos de otimização executados melhoraram profundamente a especificidade do software augustus sem comprometer a sua sensibilidade. A otimização estatística é um processo simples de implementar e foi capaz de melhorar tanto a sensibilidade quanto a especificidade da predição gênica, já a adição de evidências de transcriptomas é bem mais complexa mas foi capaz de gerar uma grande melhora na qualidade da predição.

Removida a redundância gerada pelos variantes de *splicing*, a utilização de evidências de RNA foi capaz de melhorar tanto a sensibilidade (modelos gênicos perdidos de 234 para 149), quanto especificidade da predição, reduzindo o número total de modelos gênicos de 34628 para 23194 produtos quando a evidência de transcriptoma é utilizada, sem gerar problemas como a falsa duplicação de genes.

Recentemente uma metodologia foi publicada com detalhes sobre a utilização do AUGUSTUS em genomas novos (Hoff & Stanke, 2018), onde se sugere a utilização de 4 tipos de evidências: 1) sequências curtas de RNA; 2) proteínas de organismos próximos; 3) sequências longas de RNA; 4) predições interativas. Dos quatro tipos de evidência apresentados no artigo 3 foram utilizadas neste trabalho (sequências curtas de RNA, sequências longas de RNA e predições interativas) a execução de predições interativas teve seu uso limitado a otimização estatística dos parâmetros do AUGUSTUS, duas predições interativas foram tentadas utilizando genes preditos como evidência para novos treinamentos, mas essa abordagem foi abandonada por ser tendenciosa (pois o conjunto de genes de referência passa a ter uma super-representação de genes com estruturas similares que são mais facilmente preditos pelo AUGUSTUS) e não terem apresentado melhora nos resultados.

5.6 Comparação da predição gênica com as montagens do genoma e transcriptomas.

Uma comparação geral da predição final, com a montagem do genoma e as duas montagens de transcriptoma produzidas, mostra que a predição compreende a grande maioria dos BUSCOS detectáveis para *Triatoma infestans*, mantendo ainda assim uma boa especificidade se comparada aos outros tipos de dados, como podemos ver no diagrama da Figura 17, apenas 52 BUSCOs (2.1%) dos BUSCOs detectados através de todas as metodologias juntas não foram detectados na predição final, destes, 13 BUSCOs foram detectados apenas nos dados de transcriptomas e isso pode ser consequência de problemas no sequenciamento do genoma, como a ausência de suas sequências no genoma, a fragmentação das sequências em mais de um contig ou a presença de introns demasiadamente longos. A montagem do genoma sozinha possui 7 genes exclusivos, que não foram encontrados nem mesmo nos transcriptomas, pode-se tratar de genes muito raramente expressos ou pseudogenes ou algum motivo desconhecido, porém estes 7 genes representam apenas 0.3% ds genes universalmente conservados de endopterigotos e os 6.1% de BUSCOs perdidos na predição final do AUGUSTUS já levam estes 7 genes em consideração

5.7 Comparação da predição final de *Triatoma infestans* com os demais genomas de hemípteros

Quanto ao número de preditos a predição para *Triatoma infestans* possui um número próximo ao observado para os outros genomas avaliadas (média de 24.009 genes, desvio padrão de 7.034 genes).

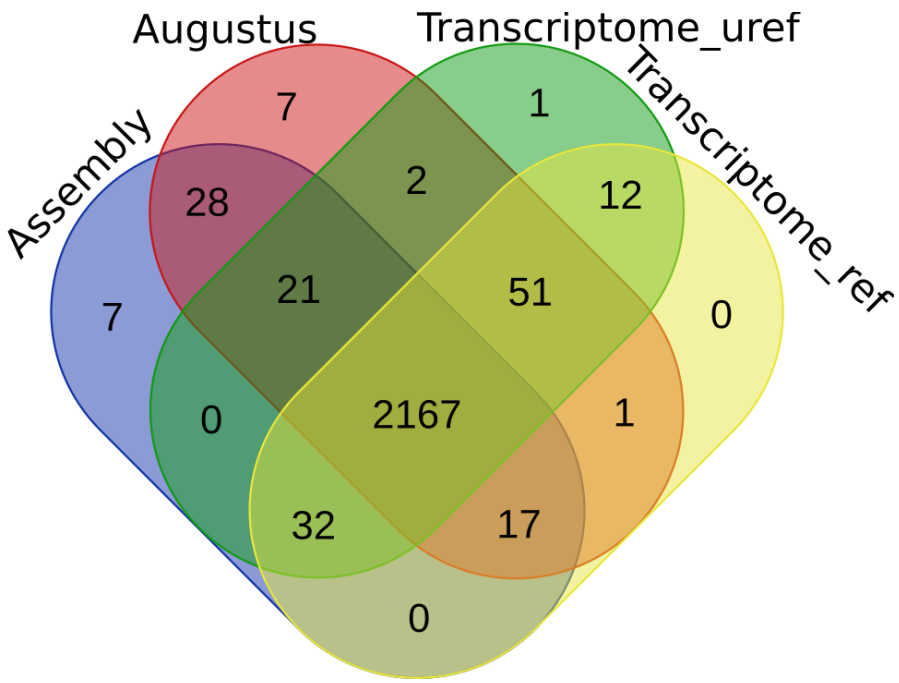


Figura 17: Comparação geral dos BUSCOs (de um total de 2442) encontrados para a montagem do genoma (Assembly), montagem do transcriptoma utilizando o genoma como referência (Transcriptome_ref), a montagem do transcriptoma sem utilizar o genoma de referência Transcriptome_uref, e a predição final do AUGUSTUS (Augustus).

Quando os resultados do BUSCO para a predição de *Triatoma infestans* são comparados com as predições de genomas de hemípteros mais recentemente publicados percebe-se que a predição de *Triatoma infestans* apresentou uma quantidade de genes totais e modelos BUSCO perdidos similar aos de outras predições. A qualidade final da predição de *Triatoma infestans* se mostrou inclusive superior aos resultados obtidos para *Acyrtosiphom pisum* e *Rhodnius prolixus*, é provável que a menor qualidade destas duas predições seja um reflexo do fato delas terem sido feitas a mais tempo quando a tecnologia de softwares de predição gênica era menos desenvolvida.

A grande quantidade de genes duplicados em *Halyomorpha halys* se deve ao fato de grande parte das proteínas depositadas no NCBI serem provenientes de dados de sequenciamento de transcriptomas e apresentarem duplicações.

6. CONCLUSÃO

Podemos separar as conclusões obtidas neste trabalho em dois grupos, conclusões técnicas e conclusões biológicas.

As principais conclusões técnicas são que o software AUGUSTUS possui um excelente desempenho quando devidamente treinado, e que um treinamento específico para a espécie de estudo é de grande importância para a predição gênica. Outros dois pontos técnicos são de grande importância: o treinamento de um software de predição gênica deve se fundamentar em um conjunto de genes de referência não necessariamente grande, mas de grande qualidade. Foi possível perceber também que evidências de transcriptomas são a principal fonte de informação para a predição de genes em novos genomas atualmente, e isso se deve principalmente a redução nos custos de sequenciamento nos sequenciadores de última geração, comumente chamados NGS (Next-Generation Sequencing), o que tem revolucionado os campos da genômica e transcriptômica.

A principal conclusão biológica é que apesar das grandes diferenças de tamanho do genoma nos hemípteros o número de genes parece ser próximo a 25 mil na maioria deles, com a exceção e *Acyrthosiphon pisum*.

A publicação do genoma de Triatoma infestans pode ajudar na compreensão da filogenética de hemípteros, permitindo que este novo genoma seja utilizado em trabalhos de filogenômica como o de Misof (2014).

7. REFERÊNCIAS

ALLEN, J. E.; PERTEA, M.; SALZBERG, S. L. Computational gene prediction using multiple sources of evidence. **Genome research**, v. 14, n. 1, p. 142–8, jan. 2004.

ALLEN, J. E.; SALZBERG, S. L. JIGSAW: integration of multiple sources of evidence for gene prediction. **Bioinformatics (Oxford, England)**, v. 21, n. 18, p. 3596–603, 15 set. 2005.

AMARAL, P. P. Et Al. The eukaryotic genome as an RNA machine. **Science (New York, N.Y.)**, v. 319, n. 5871, p. 1787–9, 28 mar. 2008.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature genetics**, v. 25, n. 1, p. 25–9, maio 2000.

BARBA-MONTOYA, J. et al. Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. **New Phytologist**, v. 218, n. 2, p. 819–834, 2018.

BARGUES, M. D. et al. Origin and phylogeography of the Chagas disease main vector Triatoma infestans based on nuclear rDNA sequences and genome size. **Infection, Genetics and Evolution**, v. 6, n. 1, p. 46–62, 2006.

BARGUES, M. D. et al. Nuclear rDNA-based Molecular Clock of the Evolution of Triatominae (Hemiptera: Reduviidae), Vectors of Chagas Disease. **Memorias do Instituto Oswaldo Cruz**, v. 95, n. 4, p. 567–573, 2000.

BARGUES, M. D.; SCHOFIELD, C.; DUJARDIN, J. P. **Classification and systematics of the Triatominae**. Second Edi ed. [s.l.] Elsevier Inc., 2017.

BARRETT, T.V. Advances in triatomine bug ecology in relation to Chagas disease. In: HARRIS, K.H. (Ed.), **Advances in Disease Vector Research, vol. 8**. Springer-Verlag, New York, pp. 143–176. 1991

BELTON, J.-M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. **Methods (San Diego, Calif.)**, v. 58, n. 3, p. 268–76, nov. 2012.

BIRNEY, E.; DURBIN, R. Using GeneWise in the Drosophila annotation experiment. **Genome research**, v. 10, n. 4, p. 547–8, abr. 2000.

BIRNEY, E. et al. An overview of Ensembl. **Genome research**, v. 14, n. 5, p. 925–8, maio 2004a.

BIRNEY, E.; CLAMP, M.; DURBIN, R. GeneWise and Genomewise. **Genome research**, v. 14, n. 5, p. 988–95, maio 2004b.

BLANK, M.; GOODMAN, R. DNA is a fractal antenna in electromagnetic fields.

International journal of radiation biology, v. 87, n. 4, p. 409–15, abr. 2011.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

BURGE, C.; KARLIN, S. Prediction of complete gene structures in human genomic DNA. **Journal of molecular biology**, v. 268, n. 1, p. 78–94, 25 abr. 1997.

BRENT, M. R. Brent MR. Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res. 2005;15:1777--86. p. 1777–1786, 2005.

CANTAREL, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. **Genome research**, v. 18, n. 1, p. 188–96, jan. 2008.

CARBON, S. et al. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. **Nucleic Acids Research**, v. 45, n. D1, p. D331–D338, 2017.

CEBALLOS, L. A. et al. First finding of melanic sylvatic Triatoma infestans (Hemiptera: Reduviidae) colonies in the Argentine Chaco. **Journal of medical entomology**, v. 46, n. 5, p. 1195–202, set. 2009.

CHEN, X.-G. et al. Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. **Proceedings of the National Academy of Sciences of the United States of America**, v. 112, n. 44, p. E5907-15, 3 nov. 2015.

CHOREV, M. et al. The function of introns. **Frontiers in Genetics**, v. 3, n. April, p. 1–15, 2012.

CONESA, A. et al. A survey of best practices for RNA-seq data analysis. **Genome biology**, v. 17, n. 4, p. 13, 26 jan. 2016.

CONTEH, L.; ENGELS, T.; MOLYNEUX, D. H. Neglected Tropical Diseases 4 Socioeconomic aspects of neglected tropical diseases. **The Lancet**, v. 375, n. 9710, p. 239–247, 2010.

COURA, J. R. The main sceneries of chagas disease transmission. The vectors, blood and oral transmissions - A comprehensive review. **Memorias do Instituto Oswaldo Cruz**, v. 110, n. 3, p. 277–282, 2015.

CUNNINGHAM, C. B. et al. The genome and methylome of a beetle with complex social behavior, nicrophorus vespilloides (coleoptera: Silphidae). **Genome Biology and Evolution**, v. 7, n. 12, p. 3383–3396, 2015.

CURWEN, V. et al. The Ensembl automatic gene annotation system. **Genome research**, v. 14, n. 5, p. 942–50, maio 2004.

DE BIE, T. et al. CAFE: a computational tool for the study of gene family evolution. **Bioinformatics**, v. 22, n. 10, p. 1269–1271, 15 maio 2006.

DE PAULA, A. S.; DIOTAIUTI, L.; SCHOFIELD, C. J. Testing the sister-group relationship of the Rhodniini and Triatomini (Insecta: Hemiptera: Reduviidae: Triatominae). **Molecular Phylogenetics and Evolution**, v. 35, n. 3, p. 712–718, 2005.

DEVESON, I. W. et al. Universal Alternative Splicing of Noncoding Exons. **Cell Systems**, v. 6, n. 2, p. 245–255.e5, 2018.

DOYLE, J. A. Molecular and Fossil Evidence on the Origin of Angiosperms. **Annual Review of Earth and Planetary Sciences**, v. 40, n. 1, p. 301–326, 2012.

DUJARDIN, J.; PANZERA, P.; SCHOFIELD, C. Triatominae as a model of morphological plasticity under ecological pressure. **Memórias do Instituto Oswaldo Cruz**, v. 94, n. suppl 1, p. 223–228, set. 1999.

EDDY, S. R. What is a hidden Markov model? **Nat Biotech**, v. 22, n. 10, p. 1315–1316, 2004.

EL-GEBALI, S. et al. The Pfam protein families database in 2019. **Nucleic Acids Research**, v. 47, n. October 2018, p. 427–432, 2018.

ELLIOTT, T. A.; GREGORY, T. R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 370, n. 1678, 2015.

EZKURDIA, I. et al. Multiple evidence strands suggest that theremay be as few as 19 000 human protein-coding genes. **Human Molecular Genetics**, v. 23, n. 22, p. 5866–5878, 2014.

FERNÁNDEZ-MEDINA, R. D. et al. Transposition burst of mariner-like elements in the sequenced genome of Rhodnius prolixus. Insect Biochemistry and Molecular Biology, v. 69, p. 14–24, 2016.

FERREIRA, I. DE L. M.; SILVA, T. P. T. Eliminação da transmissão da doença de Chagas pelo Triatoma infestans no Brasil : um fato histórico Transmission elimination of Chagas ' disease by Triatoma infestans in Brazil : an historical fact. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 39, n. 5, p. 507–509, 2006.

FLICEK, P. et al. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. **Genome research**, v. 13, n. 1, p. 46–54, jan. 2003.

FLOREA, L. et al. A computer program for aligning a cDNA sequence with a genomic

DNA sequence. **Genome research**, v. 8, n. 9, p. 967–74, set. 1998.

FORATTINI, O. P. Biogeography, origin, and distribution of triatominae domiciliarity in Brazil. **Revista de saude publica**, v. 40, n. 6, p. 964–998, 2006.

FREELING, M. et al. A solution to the c-value paradox and the function of junk DNA: The genome balance hypothesis. **Molecular Plant**, v. 8, n. 6, p. 899–910, 2015.

GALVÃO, C. et al. A checklist of the current valid species of the subfamily Triatominae Jeannel, 1919 (Hemiptera, Reduviidae) and their geographical distribution, with nomenclatural and taxonomic notes. **Zootaxa**, v. 202, n. 1, p. 1, 2003.

GARCIA, M. N. et al. The 1899 United States Kissing Bug Epidemic. **PLoS Neglected Tropical Diseases**, v. 9, n. 12, p. 2–5, 2015.

GORLA, D. E.; DUJARDIN, J. P.; SCHOFIELD, C. J. Biosystematics of Old World triatominae. **Acta Tropica**, v. 63, n. 2–3, p. 127–140, 1997.

GRABHERR, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. **Nature biotechnology**, v. 29, n. 7, p. 644–52, 15 maio 2011.

GROSS, S. S.; BRENT, M. R. Using multiple alignments to improve gene prediction. **Journal of computational biology : a journal of computational molecular cell biology**, v. 13, n. 2, p. 379–93, mar. 2006.

GUHL, F. Chagas disease in Andean countries. **Memorias do Instituto Oswaldo Cruz**, v. 102, n. SUPPL. 1, p. 29–37, 2007.

GUIGÓ, R. et al. Prediction of gene structure. **Journal of Molecular Biology**, v. 226, n. 1, p. 141–157, 1992.

HAN, M. V. et al. Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. **Molecular Biology and Evolution**, v. 30, n. 8, p. 1987–1997, ago. 2013.

HARROW, J. et al. Identifying protein-coding genes in genomic sequences. **Genome Biology**, v. 10, n. 1, p. 201, 2009.

HOFF, K. J.; STANKE, M. Predicting Genes in Single Genomes with AUGUSTUS. **Current Protocols in Bioinformatics**, p. e57, 2018.

HOLT, C.; YANDELL, M. MAKER2 : an annotation pipeline and genome- database management tool for second- generation genome projects. **BMC Bioinformatics**, v. 12, n. 1, p. 491, 2011.

HWANG, W. S.; WEIRAUCH, C. Evolutionary History of Assassin Bugs (Insecta: Hemiptera: Reduviidae): Insights from Divergence Dating and Ancestral State

Reconstruction. **PLoS ONE**, v. 7, n. 9, 2012.

IBARRA-CERDEÑA, C. N. et al. Phylogeny and niche conservatism in North and Central American triatomine bugs (Hemiptera: Reduviidae: Triatominae), vectors of Chagas' disease. **PLoS neglected tropical diseases**, v. 8, n. 10, p. e3266, out. 2014.

JONES, S. J. M. Prediction of genomic functional elements. **Annual review of genomics and human genetics**, v. 7, p. 315–38, 2006.

JUSTI, S. A.; GALVÃO, C. The Evolutionary Origin of Diversity in Chagas Disease Vectors. **Trends in Parasitology**, v. 33, n. 1, p. 42–52, 2017.

JUSTI, S. A.; GALVÃO, C.; SCHRAGO, C. G. Geological Changes of the Americas and their Influence on the Diversification of the Neotropical Kissing Bugs (Hemiptera: Reduviidae: Triatominae). **PLoS Neglected Tropical Diseases**, v. 10, n. 4, p. 1–22, 2016.

JUSTI, S. A. et al. Molecular phylogeny of Triatomini (Hemiptera: Reduviidae: Triatominae). **Parasites and Vectors**, v. 7, n. 1, p. 1–12, 2014.

KIM, D.; LANGMEAD, B.; SALZBERG, S. L. HISAT: a fast spliced aligner with low memory requirements. **Nature methods**, v. 12, n. 4, p. 357–60, 9 abr. 2015.

KORF, I. et al. Integrating genomic homology into gene structure prediction. **Bioinformatics (Oxford, England)**, v. 17 Suppl 1, p. S140-8, 2001.

KORF, I. Gene finding in novel genomes. **BMC bioinformatics**, v. 5, p. 59, 14 maio 2004.

KULP, D. et al. A generalized hidden Markov model for the recognition of human genes in DNA. **Proceedings. International Conference on Intelligent Systems for Molecular Biology**, v. 4, p. 134–42, 1996.

LEHANE, M. J. The evolution of the blood-sucking habit. In: **The Biology of Blood-Sucking in Insects**. Second Edi ed. Cambridge: Cambridge University Press, 2005. 7–13.

LENT, H.; WYGODZINSKY, P. Revision Of The Triatominae (Hemiptera, Reduviidae), And Their Significance As Vectors Of Chagas' Disease. **Bulletin Of The American Museum Of Natural History**, v. 163, 1979.

LOMSADZE, A. et al. Gene identification in novel eukaryotic genomes by self-training algorithm. **Nucleic Acids Research**, v. 33, n. 20, p. 6494–6506, 2005.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, v. 17, n. 1, p. 10, 2 maio 2011.

MESQUITA, R. D. et al. Genome of Rhodnius prolixus, an insect vector of Chagas

disease, reveals unique adaptations to hematophagy and parasite infection. **Proceedings of the National Academy of Sciences**, v. 112, n. 48, p. 14936–14941, 16 nov. 2015.

MIRNY, L. A. The fractal globule as a model of chromatin architecture in the cell. **Chromosome Research**, v. 19, n. 1, p. 37–51, 2011.

MISOF, B. et al. Phylogenomics resolves the timing and pattern of insect evolution. **Science**, v. 346, n. 6210, p. 763–767, 7 nov. 2014.

MONTEIRO, F. A. et al. Mitochondrial DNA Variation of Triatoma infestans.pdf. v. 94, p. 229–238, 1999.

MONTEIRO, F. A. et al. Evolution, Systematics, and Biogeography of the Triatominae, Vectors of Chagas Disease. **Advances in Parasitology**, n. September, 2018.

MOTT, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Computer applications in the biosciences : CABIOS, v. 13, n. 4, p. 477–8, ago. 1997.

NEAFSEY, D. E. et al. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. **Science**, v. 347, n. 6217, 2015.

NEL, A. et al. The earliest known holometabolous insects. **Nature**, v. 503, n. 7475, p. 257–261, 2013.

NOIREAU, F. et al. Sylvatic population of Triatoma infestans from the Bolivian Chaco: from field collection to characterization. **Memorias do Instituto Oswaldo Cruz**, v. 95 Suppl 1, n. 11, p. 119–22, 2000.

OHNO, S. So much “junk” DNA in our genome. **Brookhaven symposia in biology**, v. 23, p. 366–70, 1972.

OLIVER, M. J. et al. The mode and tempo of genome size evolution in eukaryotes. **Genome Research**, v. 17, p. 594–601, 2008.

OTÁLORA-LUNA, F. et al. Evolution of hematophagous habit in Triatominae (Heteroptera: Reduviidae). **Revista Chilena de Historia Natural**, v. 88, n. February, 2015.

PANZERA, F. et al. Genomic Changes of Chagas Disease Vector, South America. **Emerging Infectious Diseases**, v. 10, n. 3, p. 438–446, 2004.

PANZERA, F. et al. Genome size determination in chagas disease transmitting bugs (hemiptera-triatominae) by flow cytometry. **The American journal of tropical medicine and hygiene**, v. 76, n. 3, p. 516–21, mar. 2007.

PANZERA, Y. et al. High dynamics of rDNA cluster location in kissing bug holocentric chromosomes (Triatominae, Heteroptera). **Cytogenetic and genome research**, v.

138, n. 1, p. 56–67, 2012.

PARRA, G. et al. Comparative gene prediction in human and mouse. **Genome research**, v. 13, n. 1, p. 108–17, jan. 2003.

PATTERSON, J. S.; GAUNT, M. W. Phylogenetic multi-locus codon models and molecular clocks reveal the monophyly of haematophagous reduviid bugs and their evolution at the formation of South America. **Molecular Phylogenetics and Evolution**, v. 56, n. 2, p. 608–621, 2010.

PÉREZ-MOLINA, J. A.; MOLINA, I. Chagas disease. **Lancet (London, England)**, v. 391, n. 10115, p. 82–94, 2018.

PETROV, D. A. Mutational Equilibrium Model of Genome Size Evolution. **Theoretical Population Biology**, v. 61, n. 4, p. 533–546, 2002.

PHEASANT, M.; MATTICK, J. S. Raising the estimate of functional human sequences. **Genome Research**, v. 17, n. 9, p. 1245–1253, 2007.

PITA, S. et al. Holocentric chromosome evolution in kissing bugs (Hemiptera: Reduviidae: Triatominae): Diversification of repeated sequences. **Parasites and Vectors**, v. 10, n. 1, p. 1–8, 2017.

POINAR, G. A primitive triatomine bug, Paleotriatoma metaxytaxa gen. et sp. nov. (Hemiptera: Reduviidae: Triatominae), in mid-Cretaceous amber from northern Myanmar. **Cretaceous Research**, v. 93, p. 90–97, 2019.

RAMANOUSKAYA, T. V; GRINEV, V. V. The determinants of alternative RNA splicing in human cells. **Molecular genetics and genomics : MGG**, v. 292, n. 6, p. 1175–1195, dez. 2017.

RAMSEY, J. M. et al. Atlas of Mexican Triatominae (Reduviidae: Hemiptera) and vector transmission of Chagas disease. **Memorias do Instituto Oswaldo Cruz**, v. 110, n. 3, p. 339–352, 2015.

REZENDE, J. M. DE; RASSI, A. R. Por Que Os Triatomíneos São Chamados De “Barbeiros”? **Revista de Patologia Tropical**, v. 37, n. 1, p. 75–83, 2008.

SAINZ, A. C. et al. Phylogeny of triatomine vectors of Trypanosoma cruzi suggested by mitochondrial DNA sequences. **Genetica**, v. 121, n. 3, p. 229–40, jul. 2004.

SALERNO, R. et al. A regional fight against Chagas disease: lessons learned from a successful collaborative partnership. **Revista Panamericana de Salud Publica**, v. 37, n. 1, p. 38–43, 2015.

SCHAEFER, W. C. Triatominae (Hemiptera : Reduviidae): Systematic Questions and Some Others. **Neotropical Entomology**, v. 32, n. 1, p. 1–10, 2003.

SCHOFIELD, C. J.; GALVÃO, C. Classification, evolution, and species groups within

the Triatominae. **Acta tropica**, v. 110, n. 2–3, p. 88–100, 2009.

SILVA, L. J. DA. Desbravamento, agricultura e doença: a doença de Chagas no Estado de São Paulo. **Cad. Saúde Pública**, v. 2, n. 2, p. 124–140, 1986.

SIMÃO, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics (Oxford, England)**, v. 31, n. 19, p. 3210–2, 1 out. 2015.

SIMPSON, J. T. et al. ABySS: A parallel assembler for short read sequence data. **Genome Research**, v. 19, n. 6, p. 1117–1123, 2009.

SLEATOR, R. D. An overview of the current status of eukaryote gene prediction strategies. **Gene**, v. 461, n. 1–2, p. 1–4, 2010.

SOLOVYEV, V. et al. Automatic annotation of eukaryotic genes, pseudogenes and promoters. **Genome Biology**, v. 7, n. Suppl 1, p. S10, 2006.

SOUVOROV, A., KAPUSTIN, Y., KIRYUTIN, B., CHETVERNIN, V.; TATUSOVA, T., AND LIPMAN, D. Gnomon – NCBI eukaryotic gene prediction tool. **Ncbi**, p. 1–24, 2010.

STANKE, M.; WAACK, S. Gene prediction with a hidden Markov model and a new intron submodel. **Bioinformatics**, v. 19, n. Suppl 2, p. ii215-ii225, 8 out. 2003.

STORMO, G. D.; HAUSSLER, D. Optimally parsing a sequence into different classes based on multiple types of evidence. **Proceedings. International Conference on Intelligent Systems for Molecular Biology**, v. 2, p. 369–75, 1994.

SUN, C.; LÓPEZ ARRIAZA, J. R.; MUELLER, R. L. Slow DNA loss in the gigantic genomes of salamanders. **Genome biology and evolution**, v. 4, n. 12, p. 1340–8, 2012.

SWINBURNE, I. A. et al. Intron length increases oscillatory periods of gene expression in animal cells. **Genes & development**, v. 22, n. 17, p. 2342–6, 1 set. 2008.

TARTAROTTI, E.; AZEREDO-OLIVEIRA, M. T. V.; CERON, C. R. Phylogenetic approach to the study of Triatomines (Triatominae, Heteroptera). **Brazilian Journal of Biology**, v. 66, n. 2b, p. 703–708, maio 2006.

TILGNER, H. et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. **Genome research**, v. 22, n. 9, p. 1616–25, set. 2012.

TRESS, M. L.; ABASCAL, F.; VALENCIA, A. Alternative Splicing May Not Be the Key to Proteome Complexity. **Trends in biochemical sciences**, v. 42, n. 2, p. 98–110, 2017.

VAN BERKUM, N. L. et al. Hi-C: A Method to Study the Three-dimensional Architecture

of Genomes. **Journal of Visualized Experiments**, n. 39, p. 1–7, 2010.

VENTER, J. C. et al. The sequence of the human genome. Science (New York, N.Y.), v. 291, n. 5507, p. 1304–51, 16 fev. 2001.

WATERHOUSE, R. M. et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. **Molecular Biology and Evolution**, v. 35, n. 3, p. 543–548, 1 mar. 2018.

WEIRAUCH, C. Cladistic analysis of Reduviidae (Heteroptera: Cimicomorpha) based on morphological characters. **Systematic Entomology**, v. 33, n. 2, p. 229–274, 2008.

WEIRAUCH, C.; MUNRO, J. B. Molecular phylogeny of the assassin bugs (Hemiptera: Reduviidae), based on mitochondrial and nuclear ribosomal genes. **Molecular Phylogenetics and Evolution**, v. 53, n. 1, p. 287–299, 2009.

WOODHOUSE, M. R. et al. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. **PLoS Biology**, v. 8, n. 6, 2010.

WOODHOUSE, M. R. et al. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. **Proceedings of the National Academy of Sciences**, v. 111, n. 14, p. 5283–5288, 8 abr. 2014.

XIA, E. H. et al. The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis. **Molecular Plant**, v. 10, n. 6, p. 866–877, 2017.

XU, T. et al. The genome of the miiuy croaker reveals well-developed innate immune and sensory systems. **Scientific Reports**, v. 6, p. 1–9, 2016.

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

YEH, R. F.; LIM, L. P.; BURGE, C. B. Computational inference of homologous gene structures in the human genome. **Genome research**, v. 11, n. 5, p. 803–16, maio 2001.

ZDOBNOV, E. M. et al. OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. **Nucleic Acids Research**, v. 45, n. D1, p. D744–D749, 2017.

ZEPEDA MENDOZA, M. L. et al. Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat. **Nature Ecology & Evolution**, v. 2, n. 4, p. 659–668, 19 abr. 2018.

ZHANG, J. et al. Evolution of the assassin's arms: Insights from a phylogeny of combined transcriptomic and ribosomal DNA data (Heteroptera: Reduvioidea). **Scientific Reports**, v. 6, n. February, p. 1–8, 2016.

ZHOU, D. et al. Genome sequence of Anopheles sinensis provides insight into genetics basis of mosquito competence for malaria parasites. **BMC genomics**, v. 15, n. 1, p. 42, 18 jan. 2014.

8. ANEXOS

ANEXO 1 - Comandos utilizados para a limpeza e montagem do transcriptoma

Limpeza padrão

```
cutadapt -e 0.2 --overlap=1 -g CTACACGACGCTCTTCCGATCT -a GATCGGAAGAGCACACGTCTGA -n 2 --trim-n -o ../1cutadapt/$filename.trimed.fastq $file >>../1cutadapt/$filename.log 2>>../1cutadapt/$filename.log &

cutadapt -e 0.2 --overlap=1 -g TCAGACGTGTGCTCTTCCGATC -a AGATCGGAAGAGCGTCGTGTAG -n 2 --trim-n -o ../1cutadapt/$filename.trimed.fastq $file >>../1cutadapt/$filename.log 2>>../1cutadapt/$filename.log &
```

Limpeza de contaminação por poly-A ou poly-T (utilizado apenas quando esse tipo de contaminação é encontrada)

```
cutadapt -e 0.2 --overlap=1 -n 2 -m 50 -a "A{12}" -a "T{12}" -o ../1cutadapt/$filename.fastq $file >>../1cutadapt/$filename.log 2>>../1cutadapt/$filename.log &
```

Limpeza de qualidade de base e sincronização

```
trimmomatic PE $filename\_1.trimed.fastq $filename\_2.trimed.fastq ../3trimmomatic/$filename\_p1.fastq ../3trimmomatic/$filename\_s1.fastq ../3trimmomatic/$filename\_p2.fastq ../3trimmomatic/$filename\_s2.fastq SLIDINGWINDOW:4:30 MINLEN:50 >>../3trimmomatic/$filename.log 2>>../3trimmomatic/$filename.log &
```

Controle de Qualidade

```
fastqc -o ../2fastqc --contaminants ../illumina_contaminants.tabular $file >>../2fastqc/$filename.log 2>> ../2fastqc/$filename.log &
```

Montagem do transcriptoma

```
trinity --seqType fq --max_memory 246G \
--single ../SRA/0_illumina/5clean/*s1.fastq \
--single ../SRA/0_illumina/5clean/*s2.fastq \
--left ../SRA/0_illumina/5clean/*p1.fastq \
--right ../SRA/0_illumina/5clean/*p2.fastq \
--CPU 20
```

O comando utilizado para a avaliação de ambas as montagens de transcriptoma foi:

```
python ./scripts/run_BUSCO.py -i transcriptome.fasta -o transcriptome.fasta_endopterygota -l endopterygota_odb9/ -m transcriptome -c 1
```

ANEXO 2 - Texto original do treinamento pelo script MAKER

Methodology:

RepeatScout was used to identify *de novo* repetitive elements in the *Triatoma infestans* genome. It generated a library of 3286 repetitive sequences with l-mer size 16. This library was then filtered using below parameters:

1) Length of predicted repeats should be >50 bp;

2) Repeats with frequency in the genome should be > 10.

The resultant 1945 consensus sequences were classified using TEclass.

Repeat statistics: Teclass result

DNA transposons	458
LTRs	250
LINEs	832
SINEs	62
Unclear	343

Then genome assembly was masked with 1945 consensus sequences using RepeatMasker and 1945 distinct families of repetitive elements masked 28.78% of the genome.

Gene prediction:

The masked genome was used for annotation. Annotation for the *T. infestans* genome assembly were generated using genome annotation pipeline MAKER. MAKER was run four times to improve gene prediction result.

First run: *Ab initio* gene predictions were produced inside of MAKER by the programs Augustus. CDS sequences of *T. infestans* was submitted as hint.

Second run: *Ab initio* gene predictions were produced inside of MAKER by the programs Augustus. This time CDS sequences of *T. infestans,* maker derived gff3 (from the first run) and annotation of *Rhodnius prolixus* in form of GFF3 format was provided as hints.

Third run: gene prediction program SNAP (Semi-HMM-based Nucleic Acid Parser) was trained using derived gff3 (obtained from second run). Then gene prediction was carried out using MAKER by the program Augustus and SNAP. Derived gff3 (from second run), CDS sequences of *T. infestans,* and annotation of *Rhodnius prolixus* were used as evidence.

Fourth run: gene prediction was performed again using MAKER by the program Augustus and GFF3 file (generated from third run), CDS of *T. infestans* and annotation of *Rhodnius prolixus* were passed for evidence. A total of **16,414** genes were predicted from *T. infestans genome* using MAKER.

Note: In Augustus, *Rhodnius* was selected as a model species. MAKER was rerun many times to refine gene prediction.

Gene prediction result:

	Number of predicted genes
First run	12835
Second run	13748
Third run	20407
Fourth run	**16,414**

ANEXO 3 - Implementação De Dados De Transcriptomas

```
#!/bin/bash


### Routines


function download {
# download SRA files
cd $1
nohup fastq-dump --split-files SRR4449941 >SRR4449941.log &
nohup fastq-dump --split-files SRR4449940 >SRR4449940.log &
nohup fastq-dump --split-files SRR4449939 >SRR4449939.log &
nohup fastq-dump --split-files SRR4449815 >SRR4449815.log &
nohup fastq-dump --split-files SRR4449814 >SRR4449814.log &
nohup fastq-dump --split-files SRR4427079 >SRR4427079.log &
nohup fastq-dump --split-files SRR4427078 >SRR4427078.log &
nohup fastq-dump --split-files SRR1168938 >SRR1168938.log &
nohup fastq-dump --split-files SRR1168894 >SRR1168894.log &
nohup fastq-dump --split-files SRR1168893 >SRR1168893.log &
nohup fastq-dump --split-files SRR1168892 >SRR1168892.log &
nohup fastq-dump --split-files SRR1168891 >SRR1168891.log &
nohup fastq-dump --split-files SRR1168890 >SRR1168890.log &
nohup fastq-dump --split-files SRR1168889 >SRR1168889.log &
nohup fastq-dump --split-files SRR1168888 >SRR1168888.log &
nohup fastq-dump --split-files SRR1168885 >SRR1168885.log &
nohup fastq-dump --split-files SRR1168882 >SRR1168882.log &
cd -
}


function hisat {
```

```
#### Standard cleaning of R1

cd $1

mkdir $3

for file in *_p1.fastq

do

filename=${file%_p1.fastq}

echo $filename

hisat2 -x $2 -1 $filename\_p1.fastq -2 $filename\_p2.fastq -S
$3/$filename\_paired.sam |tee $3/$filename\_paired.log &

hisat2 -x $2 -U $filename\_s1.fastq,$filename\_s1.fastq -S
$3/$filename\_single.sam |tee $3/$filename\_single.log &

done

cd -

}


function sam2bam {

mkdir $2

cd $1

for file in *.sam

do

filename=${file%.sam}

echo $filename

echo "samtools view -b -S $filename.sam >$2/$filename.bam"

samtools view -b -S $filename.sam >$2/$filename.bam

done

cd -

}


function discard_fail {

mkdir $2/08b_SAM
```

```
cd $1


for file in *.sam

do

filename=${file%.sam}

echo $filename

cat $2/header.txt >$2/08b_SAM/$filename.F.sam

samtools view -S -F 4 $file >> $2/08b_SAM/$filename.F.sam

done

cd -

}

function samMap {

mkdir $2/08c_SAM

cd $1

for file in *.sam

do

filename=${file%.sam}

echo "   $filename"

cat $2/header.txt >$2/08c_SAM/$filename.global.sam

$2/augustus/scripts/samMap.pl $file $2/map.psl
>>$2/08c_SAM/$filename.global.sam

done

cd -

}




echo ""

#1# Genome preparation (Masking)

echo "1. Masking genome"
```

```
    # I have a masked genome already


#2# Read preparation

    #2.1# Rename Headers

    #2.2# Quality trimming

    # In our experiments, we obeserved a decrease in gene prediction accuracy
if the reads were

    # quality trimmed prior alignment (we tested accepted error rates of 1%
and 5%),

    # although Bowtie/Tophat aligns more reads if they were quality trimmed.
Therefore, we recommend

    # to work with untrimmed fastq files. (Augustus team)

echo "2. Downloading SRA files" #for now I will use the trimmed fastqs I have
already

#mkdir ./0raw

#download ./0raw



#3a# Align Reads using bowtie

#3a.1# Build Indexes

echo "3a.1 Build Indexes"

#hisat2-build ../../Tinf.genome.fa.masked
/home/double1/Triatoma_infestans_genome/hisat2/triatoma_infestans_masked

#3a.2# Align Reads

echo "3a.2 Align Reads"

#hisat ./5clean ../../hisat2/triatoma_infestans_masked



#4# Filtering raw alignments (step 1) optional

#samtools sort SAM/*.sam both.ssf

# filter alignments with filterBam

#filterBam --uniq --paired --in output_directory/accepted_hits.s.bam --out
output_directory/accepted_hits.sf.bam
```

```
#samtools view -H output_directory/accepted_hits.sf.bam > header.txt


#5# Creating intron hints (step 1)

echo "5. Creating intron hints (step 1)"

echo "converting SAM files to BAM format"

#sam2bam

echo "joining BAM files"

#samtools cat BAM/*.bam >all_libraries.bam

echo "processing BAM headers"

#samtools view -H all_libraries.bam > header.txt

echo "sorting BAM file"

#samtools sort all_libraries.bam triatoma_sorted

echo "processing intron hints"

#augustus/auxprogs/bam2hints/bam2hints --intronsonly --in=triatoma_sorted.bam
--out=t_infestans.intron_hints.gff


#6# Run Augustus (step 1)

echo "6. Run Augustus to interactively generate exon hints"

#6.1# Set hint parameters (step 1)

echo "6.1 Creating extinsic information configuration file"

#cp augustus/config/extrinsic/extrinsic.M.RM.E.W.cfg
augustus/config/species/t_infestans/extrinsic.cfg

#6.2# Run augustus

echo "Running augustus"

#augustus/bin/augustus --species=t_infestans
--extrinsicCfgFile=augustus/config/species/t_infestans/extrinsic.cfg
--alternatives-from-evidence=true --hintsfile=t_infestans.intron_hints.gff
--allow_hinted_splicesites=atac --introns=on --genemodel=complete
../../triatoma_infestans.supercont.fasta > aug1.out

#7# Create an exon-exon junction database (step 2)

echo "7. Creating an exon-exon junction database (step 2)"
```

```
echo "extracting intron informations from augustus prediction"

#cat 17112801_results/*.out | tee aug.prelim.gff | grep -P "\tintron\t" >
aug1.introns.gff

# in case hints.gff also contains other hints than "intron", you need to
filter for "intron", first!

#cat t_infestans.intron_hints.gff aug1.introns.gff | perl -ne '@array =
split(/\t/, $_);print "$array[0]:$array[3]-$array[4]\n";'| sort -u >
introns.lst

echo "excising the exon-exon-junctions and write them into fasta format"

#augustus/scripts/intron2exex.pl --introns=introns.lst
--seq=../../Tinf.genome.fa.masked --exex=exex.fa --map=map.psl

# Build hisat2 database for exex junctions

echo "building a bowtie database from the exon-exon junction file"

#hisat2-build exex.fa t_infestans_exex1



#8# Aligning reads with Bowtie (step 2)

echo "8. Aligning reads with Bowtie (step 2)"

echo "a. running alignment"

#mkdir 08_SAM

#hisat 5clean
/home/double1/Triatoma_infestans_genome/augustus/17112202_RNA-seq_incorporatio
n/t_infestans_exex1
/home/double1/Triatoma_infestans_genome/augustus/17112202_RNA-seq_incorporatio
n/08a_SAM

echo "b. Discaring failed alignments from the output:"

#discard_fail 08a_SAM
/home/double1/Triatoma_infestans_genome/augustus/17112202_RNA-seq_incorporatio
n

echo "c. Maping the local exex-alignments to global genome level:"

#samMap 08b_SAM
/home/double1/Triatoma_infestans_genome/augustus/17112202_RNA-seq_incorporatio
n

echo "9. Join data from step 1 and step 2"

echo "a. discarding intron containing alignments from the original bam file"

#bamtools filter -in triatoma_sorted.bam -out triatoma_sorted.noN.bam -script
```

```
augustus/auxprogs/auxBamFilters/operation_N_filter.txt

echo "b. create a bam file with header from the sam files"

#sam2bam 08c_SAM
/home/double1/Triatoma_infestans_genome/augustus/17112202_RNA-seq_incorporatio
n/09b_BAM

echo "c. join bam files"

#samtools merge 09_final.bam 09b_BAM/*.bam triatoma_sorted.noN.bam

echo "d.sorting bam file"

#samtools sort -n 09_final.bam 09d_final.s

echo "10. Filtering raw alignments (step 2)"

#augustus/auxprogs/filterBam/filterBam --uniq --in 09d_final.s.bam --out
10_final.sf.bam

augustus/auxprogs/filterBam/filterBam --paired --uniq --in 09d_final.s.bam
--out 10_final.sfp.bam --pairwiseAlignments

echo "11. Creating intron hints (step 2)"

echo "a. sorting filtered bam"

samtools sort 10_final.sfp.bam 11_final.ssfp

echo "b. generating final intron hints"

augustus/auxprogs/bam2hints/bam2hints --intronsonly --in=11_final.ssfp.bam
--out=11_final.hints.2p.gff
```