UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA E BIOCIÊNCIAS

Guilherme Augusto Maia

Nova versão do genoma de *Trypanosoma rangeli* isolado em Santa Catarina, Brasil: comparação de aspectos estruturais e funcionais com tripanosomatídeos patogênicos a humanos

Florianópolis
2023

Guilherme Augusto Maia

Nova versão do genoma de *Trypanosoma rangeli* isolado em Santa Catarina, Brasil: comparação de aspectos estruturais e funcionais com tripanosomatídeos patogênicos a humanos

Tese submetida ao Programa de Pós-Graduação em Biotecnologia e Biociências da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Doutor em Biotecnologia e Biociências.

Orientador: Prof. Glauber Wagner, Dr.
Coorientador: Prof. Edmundo Carlos Grisard, Dr.

Florianópolis
2023

Maia, Guilherme Augusto
Nova versão do genoma de Trypanosoma rangeli isolado em Santa Catarina, Brasil : comparação de aspectos estruturais e funcionais com tripanosomatídeos patogênicos a humanos / Guilherme Augusto Maia ; orientador, Glauber Wagner, coorientador, Edmundo Carlos Grisard, 2023.
130 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Biotecnologia e Biociências, Florianópolis, 2023.

Inclui referências.

1. Biotecnologia e Biociências. 2. Parasito. 3. Bioinformática. 4. Genômica Comparativa. 5. Proteínas Hipotéticas. I. Wagner, Glauber. II. Grisard, Edmundo Carlos. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Biotecnologia e Biociências. IV. Título.

Guilherme Augusto Maia

Nova versão do genoma de *Trypanosoma rangeli* isolado em Santa Catarina, Brasil: comparação de aspectos estruturais e funcionais com tripanosomatídeos patogênicos a humanos

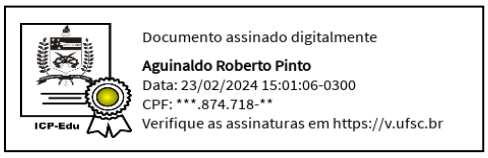
O presente trabalho em nível de Doutorado foi avaliado e aprovado, em 15 de dezembro de 2023, pela banca examinadora composta pelos seguintes membros:

Prof. Diogo Antônio Tschoeke, Dr.
Universidade Federal do Rio de Janeiro

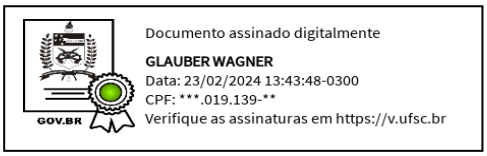
Profª. Patrícia Flávia Quaresma, Drª.
Universidade Federal de Santa Catarina

Prof. Rodrigo de Paula Baptista, Dr.
Houston Methodist Research Institute

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Doutor em Biotecnologia e Biociências.



Coordenação do Programa de Pós-Graduação



Prof. Glauber Wagner, Dr.
Orientador

Florianópolis, 2023

Dedico este trabalho a todos os cientistas brasileiros,
que mesmo diante das adversidades,
continuarão a fazer ciência de qualidade.

AGRADECIMENTOS

À minha família, Edísio, Rosana e Juliana, por todo o amor, apoio e incentivo de correr atrás dos meus sonhos. Obrigado por sempre investirem na minha educação, tanto dentro quanto fora do ambiente acadêmico. Eu não estaria onde estou hoje se não fosse por vocês e, com certeza, todo o carinho de vocês tornou essa jornada mais fácil.

Ao meu orientador, Prof. Dr. Glauber Wagner, por ter me recebido de braços abertos há anos atrás e aceitar conduzir o meu trabalho de pesquisa. Obrigado pelo seu tempo, pelas cobranças e por me incentivar a ser um profissional melhor, da maneira mais humana e correta o possível. Obrigado pelo exemplo de que um ambiente de trabalho acadêmico e laboratorial pode ser leve, assim como pela oportunidade de fazer parte de um grupo de pesquisa de excelência.

Ao meu coorientador, Prof. Dr. Edmundo Carlos Grisard, por todas as contribuições para a construção deste trabalho. Obrigado por todas as conversas, horas de discussões e por sempre me estimular a procurar na biologia uma maneira de explorar algo além da luz do poste. Obrigado pelo exemplo de liderança e organização, bem como por ter me dado um voto de confiança nas portas que abristes durante a minha formação profissional e pessoal.

Ao Prof. Dr. Björn Andersson, pela oportunidade de ter feito parte do seu grupo de pesquisa durante o período do meu estágio sanduíche. Obrigado por ter me recebido, por me ensinar e por compartilhar tanto em tão pouco tempo. O ano passou voando, mas espero carregar seus ensinamentos durante toda a minha carreira científica.

Às equipes do Laboratório de Bioinformática e do Laboratório de Protozoologia: aos professores, Prof. Dr. Mário Steindel, Prof.ª Dr.ª Patrícia Flávia Quaresma e, em especial, à Prof.ª Dr.ª Patrícia Hermes Stoco, pelas discussões, pelos ensinamentos, pelas conversas, pelos conselhos e pelo tratamento de forma amigável e profissional ao longo de todos esses anos; também aos demais colegas que fazem ou fizeram parte do grupo de pesquisa.

Aos amigos imbatíveis e tristes, Beatriz, Carolina, Dayane, Eric, Fernando, Karin, Renato, Tatiany e Vilmar, por todas as conversas, todas as risadas, todos os choros, todos os rolês e, acima de tudo, por todo o incentivo dentro e fora do ambiente acadêmico.

Aos amigos do corredor C9 do Karolinska Institutet, Isabel, Julie, Karoline, Nuno, Pryscilla, Taís e, em especial, ao Gabriel e à Veronika. Obrigado por todo o carinho, todo o amparo e pela troca de experiências culturais. Gabriel, meu sanduíche não teria sido a mesma coisa sem a tua parceria e sem a tua ajuda, serei eternamente grato pela nossa amizade, nossas conversas, risadas e rolês! Veronika, obrigado por todo o carinho, companheirismo, por ter aceitado conhecer um pouco da Europa comigo e por ter estado do meu lado! *Milujem ťa, miláčik*.

Aos amigos de longa data, Aline, Felipe, Laís, Leandra, Rodrigo e Vinícius, pela amizade e companheirismos, tanto nos momentos bons quanto nos momentos ruins, ao longo desses anos. Por mais que a vida nos leve à lugares e situações diferentes, vocês sempre estiveram presentes e eu sou muito grato por isso.

À Universidade Federal de Santa Catarina, que foi minha segunda casa durante 11 anos, e ao Programa de Pós-Graduação em Biotecnologia e Biociências, por toda a infraestrutura, serviços e investimentos, que possibilitaram minha formação acadêmica e pessoal.

À CAPES, pelo suporte financeiro e a possibilidade de ter realizado todo o meu período de doutoramento com bolsa. Bolsa essa que possibilitou, inclusive, minha estadia no exterior pelo Programa Institucional de Internalização da CAPES (CAPES PrInt).

A todos que influenciariam e ajudaram na realização deste trabalho, direta ou indiretamente, uma vez que a construção científica acontece de forma colaborativa.

Shocking but we're nothing, we're just moments
We're clever, but we're clueless, we're just human
Amusing, confusing, but the truth is
All we got is questions we'll never know
- "Never Know", por Jack Jonhson

RESUMO

Trypanosoma rangeli é um protozoário que ocorre nas Américas, que compartilha aspectos genéticos, morfológicos e antigênicos com *T. cruzi*, agente etiológico da Doença de Chagas. Apesar das características compartilhadas e da proximidade filogenética com o *T. cruzi*, o *T. rangeli* não é considerado patogênico para seu hospedeiro mamífero. Em comparação com o *T. cruzi*, o genoma do *T. rangeli* apresenta um número menor de cópias de genes que codificam para proteínas de superfície, as quais são possíveis fatores de virulência associados aos processos de invasão celular e evasão do sistema imune. Neste contexto, o objetivo do presente estudo foi gerar uma nova versão do genoma de *T. rangeli* (cepa SC58), combinando dados obtidos através de diferentes tecnologias de sequenciamento, e realizar um estudo comparativo de aspectos estruturais e funcionais com outras espécies de tripanosomatídeos. Para tal, foi realizada uma nova montagem genômica de *T. rangeli* utilizando dados de sequenciamento do tipo *long reads* e *short reads*. As etapas de predição gênica e anotação gênica foram realizadas utilizando-se o programa AnnotaPipeline, que utilizou dados experimentais de transcriptômica e proteômica para validar as predições gênicas de *T. rangeli*. Para a anotação gênica automática por análise de similaridade, foram utilizadas sequências depositadas nos bancos de dados do SwissProt e do TriTrypDB, também sendo realizadas anotação funcional, análise de evidência de transcrição e de evidência de tradução. Em seguida, foi realizada uma análise de homologia entre as CDS descritas na versão 2 do genoma de *T. rangeli*, as CDS de *T. brucei* (cepa TREU927) e de *T. cruzi* (cepa CL Brener Esmeraldo-like), assim como as CDS preditas e anotadas na montagem do genoma de *T. conorhini* (cepa 025E). Os resultados mostram que a versão 2 do genoma de *T. rangeli* SC58 é mais contígua, menos fragmentada e abrange 99,95% de toda a informação contida na descrição original do genoma do parasito. Nesta versão, o conteúdo genômico repetitivo de *T. rangeli* é representado por uma fração de 13,53% de elementos repetitivos estruturais e por apenas 2,74% de elementos repetitivos funcionais. Os elementos estruturais mais proeminentes são as repetições intercaladas (compostas principalmente por retroelementos do tipo LTRs e LINEs) e as repetições simples (com uma maior repetição de motivos de di-nucleotídeos). Os elementos funcionais que mais se destacam são os genes multicópias das famílias das sialidases e GP63, representando 44,67% e 38,42% do total identificado, respectivamente. A devida identificação destes elementos repetitivos no genoma possibilita um melhor entendimento sobre os processos de adaptação e evolução desses organismos. A atribuição de função por meio de anotação automática foi obtida para 73,12% das CDS preditas, restando apenas 26,88% das CDS anotadas como hipotéticas. Os resultados obtidos da análise de homologia reforçam a existência de um genoma central compartilhado entre tripanosomatídeos, representado por um agrupamento de 4.156 grupos de CDS ortólogas das quatro espécies estudadas. A versão 2 do genoma de *T. rangeli* SC58 apresenta um conjunto de CDS anotadas como hipotéticas a serem caracterizadas, principalmente as ortólogas com as demais espécies de tripanosomatídeos.

Palavras-chave: Parasito; Genômica; Bioinformática; Genômica Comparativa; Proteínas Hipotéticas.

ABSTRACT

Trypanosoma rangeli is a protozoan that occurs in the Americas, sharing genetic, morphological, and antigenic aspects with *T. cruzi*, the etiological agent of Chagas disease. Despite the shared characteristics and phylogenetic proximity to *T. cruzi*, *T. rangeli* is not considered pathogenic to its mammalian host. Compared to *T. cruzi*, the genome of *T. rangeli* presents a smaller number of copies of genes encoding surface proteins, which are potential virulence factors associated with cellular invasion and immune system evasion processes. In this context, the aim of this study was to generate a new version of the *T. rangeli* genome (SC58 strain), combining data obtained through different sequencing technologies, and to conduct a comparative study of structural and functional aspects with other trypanosomatid species. To achieve this, a new genomic assembly of *T. rangeli* was performed using long-read and short-read sequencing data. Gene prediction and gene annotation steps were carried out using the AnnotaPipeline software, which used experimental transcriptomic and proteomic data to validate gene predictions of the genome. For automatic gene annotation by similarity analysis, sequences deposited in the SwissProt and TriTrypDB databases were used, along with functional annotation, transcription evidence analysis, and translation evidence analysis. Subsequently, a homology analysis was conducted among the CDS described in version 2 of the *T. rangeli* genome, the CDS of *T. brucei* (TREU927 strain) and *T. cruzi* (CL Brener Esmeraldo-like strain), as well as predicted and annotated CDS of the *T. conorhini* (025E strain) genome assembly. The results show that the version 2 of the *T. rangeli* SC58 genome is more contiguous, less fragmented, and covers 99.95% of all information contained in the original description of the parasite genome. In this version, the repetitive genomic content of *T. rangeli* is represented by a fraction of 13.53% structural repetitive elements and only 2.74% functional repetitive elements. The most prominent structural elements are interspersed repeats (composed mainly of LTRs and LINEs retroelements) and simple repeats (with a higher repetition of di-nucleotide motifs). The most representative functional elements are the multicopy genes of the sialidase and GP63 families, representing 44.67% and 38.42% of the total identified, respectively. The proper identification of these repetitive elements in the genome allows a better understanding of the processes of adaptation and evolution of these organisms. Function assignment by means of automatic annotation was obtained for 73.12% of the predicted CDS, leaving only 26.88% of the CDS annotated as hypothetical. Results obtained from the homology analysis reinforce the existence of a shared core genome among trypanosomatids, represented by a cluster of 4,156 groups of orthologous CDS from the four species studied. Version 2 of the *T. rangeli* SC58 genome presents a set of CDS annotated as hypothetical to be characterized, mainly the orthologs with the other trypanosomatid species.

Keywords: Parasite; Genomics; Bioinformatics; Comparative Genomics; Hypothetical Proteins.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

aCGH	Hibridização genômica comparativa baseada em arranjo de DNA
CDC	Centros de Controle e Prevenção de Doenças dos Estados Unidos
CDS	Sequência gênica que codifica proteína
DGF-1	Proteína de família gênica dispersa 1
EST	Etiqueta de sequência expressa
gDNA	DNA genômico
GPI	Glicosilfosfatidilinusitol
iRNA	RNA de interferência
kDNA	DNA extranuclear presente no cinetoplasto
KMP-11	Proteína de membrana de Kinetoplastida 11
LINE	Elemento nuclear intercalado longo
MASP	Proteína de superfície associada a mucina
MLST	Tipagem de sequência multilocus
ORF	Janela abertura de leitura
RHS	*Hot spot* de retrotransposon
rRNA	RNA ribossômico
SINE	Elemento nuclear intercalado curto
tRNA	RNA transportador
VSG	Glicoproteína variável de superfície

SUMÁRIO

1 INTRODUÇÃO

1.1 TRIPANOSOMATÍDEOS

Os tripanosomatídeos são protozoários pertencentes ao Filo Euglenozoa, Ordem Trypanosomatida e Família Trypanosomatidae, caracterizados por apresentarem um único flagelo e serem parasitos obrigatórios, que infectam uma grande variedade de espécies de animais e plantas (D'ALESSANDRO, 1976; DOLLET, 1984). Esta Família faz parte da Classe Kinetoplastea, caracterizada pela presença de uma região rica em DNA extranuclear chamada cinetoplasto (em inglês, *kinetoplast* DNA – kDNA), localizada em sua mitocôndria (LUKEŠ et al., 2014).

O gênero *Trypanosoma* contém agentes etiológicos que causam doenças em humanos, como a espécie *Trypanosoma cruzi* (causador da Doença de Chagas) nas Américas (CHAGAS, 1909a) e a espécie *Trypanosoma brucei* (causador da Doença do Sono) no continente Africano (BRUCE, 1914). As doenças causadas por tripanosomatídeos são conhecidas como tripanosomíases, as quais fazem parte de um grupo maior de doenças conhecidas como doenças tropicais negligenciadas (OMS, 2022). Dentre as doenças tropicais negligenciadas, destacam-se também as enfermidades causadas pelas espécies do gênero *Leishmania*, o qual também faz parte da Família Trypanosomatidae, que causam diferentes doenças com manifestações clínicas distintas em humanos e outros animais no mundo inteiro (ALVAR et al., 2012; RUIZ-POSTIGO et al., 2021).

Alguns autores ainda classificam os tripanosomatídeos em duas seções, baseadas nas diferentes formas de transmissão dos parasitos por seus vetores: a seção Stercoraria, que inclui espécies em que as formas tripomastigotas metacíclicas, que são as formas infectivas, se desenvolvem na porção medial do intestino do vetor, sendo sua infecção do tipo contaminativa ou posterior; e a seção Salivaria, que inclui espécies em que as formas tripomastigotas metacíclicas estão presentes nas glândulas salivares do vetor, sendo sua infecção do tipo inoculativa ou anterior. Exemplos de espécie pertencentes à seção Stercoraria são *T. conorhini* e *T. cruzi*, enquanto exemplos de espécies pertencentes à seção Salivaria são os tripanosomatídeos africanos *T. brucei*, *T. congolense* e *T. vivax* (STEVENS; GIBSON, 1999; SILVA et al., 2004; KAUFER et al., 2017), assim como o objeto de estudo deste trabalho: a espécie *T. rangeli*.

1.2 *Trypanosoma rangeli*: TRIPANOSOMATÍDEO NÃO PATOGÊNICO PARA MAMÍFEROS

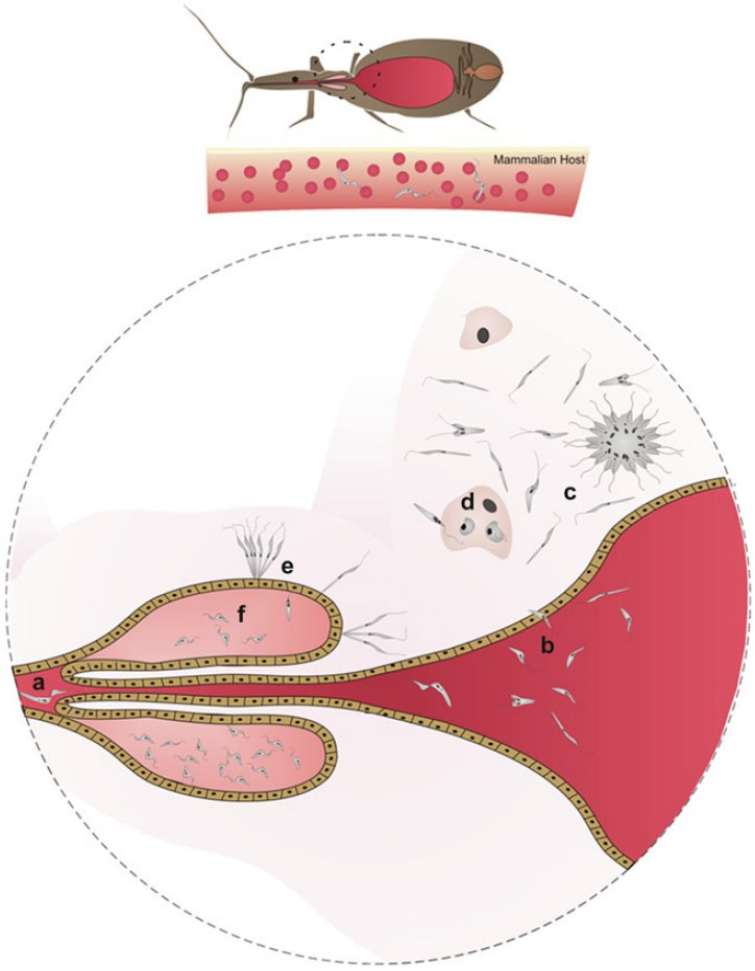
Outra espécie do gênero *Trypanosoma* que é encontrada nas Américas é o *Trypanosoma* (*Herpetosoma*) *rangeli*, descrito por Tejera em 1920. Este parasito possui um ciclo heteroxênico que envolve mamíferos de diversas ordens, como Carnivora, Edentata, Marsupialia, Primata e Rodentia, sendo primariamente transmitido por triatomíneos do gênero *Rhodnius* (D'ALESSANDRO, 1976; GRISARD et al., 1999; GUHL; VALLEJO, 2003). O *T. rangeli* apresenta alta similaridade genética e antigênica com o *T. cruzi (*AFCHAIN et al., 1979; STOCO et al., 2014), sendo que estas espécies se distribuem de maneira simpátrica, compartilhando reservatórios e vetores em diversas regiões da América Central e da América do Sul (GRISARD et al., 1999; MORAES et al., 2008). Porém, ainda que seja filogeneticamente próximo ao *T. cruzi* e que compartilhe aspectos epidemiológicos, genéticos e morfológicos com este parasito, o *T. rangeli* não é considerado patogênico para seus hospedeiros mamíferos (GUHL; VALLEJO, 2003). Neste sentido, a ocorrência simpátrica destes parasitos na América e a existência de casos de infecção humana pelo *T. rangeli* são de suma importância para o diagnóstico específico da Doença de Chagas.

As diferentes cepas de *T. rangeli* são classificadas filogeneticamente em cinco grupos, de acordo com análises de variabilidade genética realizadas com isolados de diferentes vetores e espécies de hospedeiros vertebrados: TrA, TrB, TrC, TrD e TrE (MAIA DA SILVA et al., 2007; BRADWELL et al., 2018; ESPINOSA-ÁLVAREZ et al., 2018). O estudo filogenético das cepas de *T. rangeli* pode ser feito por meio de métodos clássicos, como a amplificação aleatória de DNA polimórfico (do inglês, *Random Amplification of Polymorphic* DNA – RAPD), que envolve a amplificação do DNA genômico (gDNA) por meio da PCR, utilizando iniciadores (do inglês, *primers*) de sequências aleatórias provenientes de diversas regiões do genoma do parasito (STEINDEL et al., 1994). Nesse método, as relações filogenéticas são estimadas pela contagem do número de fragmentos amplificados em comum entre as amostras. A principal vantagem da utilização da técnica de RAPD é que a técnica não requer nenhuma informação prévia sobre as sequências nucleotídicas do genoma da espécie (HADRYS; BALICK; SCHIERWATER, 1992). Da mesma forma, também podem ser empregadas análises de marcadores moleculares, tais como o espaçador interno transcrito do gene ribossômico (do inglês, *Internal Transcribed Spacer of Ribosomal*

Gene – ITS rDNA), gliceraldeído-3-fosfato desidrogenase glicossomal (do inglês, *Glycosomal Glyceraldehyde-3-Phosphate Dehydrogenase* – gGAPDH) e polimorfismos na região variável V7 e V8 do RNA da subunidade menor do ribossômico (do inglês, *Small Subunit Ribosomal* RNA – SSU rRNA) (MAIA DA SILVA et al., 2004; ESPINOSA-ÁLVAREZ et al., 2018). Fragmentos de RNA da sequência *Spliced-Leader* (SL) também são comumente utilizados para refinar a filogenia molecular das cepas (MAIA DA SILVA et al., 2007). Ainda, considerando aspectos genômicos dos parasitos, as cepas de *T. rangeli* podem ser classificadas em duas grandes linhagens, KP1(+) e KP1(-). Esta classificação é baseada no arranjo – ou seja, na presença ou ausência – dos três tipos de sequências de kDNA minicírculo identificadas na espécie: KP1 (minicírculo que apresenta uma região conservada); KP2 (que contém duas regiões conservadas); ou KP3 (com quatro regiões conservadas) (VALLEJO et al. 2002). Biologicamente, estudos correlacionam diferentes linhagens das cepas de *T. rangeli* e suas respectivas capacidades de infectar diferentes espécies do gênero *Rhodnius* (VALLEJO et al. 2002; CUERVO; LÓPEZ; PUERTA, 2006; MAIA DA SILVA et al., 2007).

Resumidamente, o ciclo biológico do *T. rangeli* ocorre quando um inseto triatomíneo ingere as formas tripomastigotas do parasito, presentes na corrente sanguínea de um mamífero infectado. Então, os protozoários se desenvolvem no intestino do inseto, alterando sua forma celular de tripomastigotas para epimastigotas. Em seguida, as formas epimastigotas, que são as formas proliferativas, penetram as paredes do intestino do inseto, passam à hemocele, alcançam a hemolinfa e, por fim, migram em direção às glândulas salivares onde se diferenciam em formas tripomastigotas metacíclicas, que são as formas infectivas ao hospedeiro mamífero (Figura 1). Pelo fato de que as formas tripomastigotas são inoculadas junto com a saliva do inseto vetor diretamente na corrente sanguínea do hospedeiro mamífero, esta forma de transmissão é classificada como transmissão anterior ou inoculativa (MELLO et al., 1995). É interessante destacar que, além da mudança morfológica celular do parasito, a posição do cinetoplasto em relação ao núcleo e ao flagelo também varia, sendo característica para cada etapa de ciclo de vida do parasito (HOARE; WALLACE, 1966).

Figura 1. Representação do ciclo de vida de *Trypanosoma rangeli* no hospedeiro invertebrado e no hospedeiro mamífero.



Legenda: (a) ingestão das formas tripomastigotas de *T. rangeli* durante o repasto sanguíneo do triatomíneo em um hospedeiro infectado; (b) formas tripomastigotas desenvolvem-se em epimastigotas no trato intestinal do inseto, que começam a se multiplicar; (c) epimastigotas já penetraram as paredes do intestino do triatomíneo, alcançando a hemocele e a hemolinfa do inseto; (d) invasão das glândulas salivares do inseto pelas formas epimastigotas; (e) formas epimastigotas começam a se diferenciar em formas tripomastigotas metacíclicas, que são as formas infectivas; (f) inoculação das formas tripomastigotas metacíclicas através da saliva do triatomíneo durante seu repasto sanguíneo. Fonte: STOCO et al., 2016.

Apesar do *T. rangeli* compartilhar inúmeras características com o *T. cruzi*, pouco se conhece respeito a sua capacidade de sobrevivência após a infecção no hospedeiro mamífero. Até então, a divisão celular de *T. rangeli* está descrita apenas nas formas epimastigotas, que ocorrem no inseto vetor (PRESTES et al., 2019). Ainda, dados na literatura descrevem a infecção pelo *T. rangeli* como sendo de curta duração em mamíferos, atingindo o pico de máxima parasitemia no quinto dia após a

infecção, a qual começa a diminuir gradativamente no início da segunda semana pós-infecção, até tornar-se indetectável por microscopia convencional (CUBA, 1998; GRISARD; ROMANHA; STEINDEL, 2016). Após esta curta fase aguda, a presença de parasitos no mamífero só pode ser detectada por métodos indiretos, como a hemocultura e o xenodiagnóstico (GRISARD; ROMANHA; STEINDEL, 2016). De fato, foi detectada a presença de *T. rangeli* em hemocultura de camundongos infectados experimentalmente até 7 meses após a infecção (STEINDEL, 1993).

1.3 ASPECTOS GENÔMICOS DE *Trypanosoma* spp.

O ano de 2005 foi um momento muito importante nos estudos de tripanosomatídeos, com a publicação simultânea das primeiras versões dos genomas dos TriTryps: *T. cruzi*, *T. brucei* e *Leishmania major* (EL-SAYED et al., 2005a; BERRIMAN et al., 2005; IVENS et al., 2005). Ao mesmo tempo, também foi apresentada uma análise comparativa da estrutura e conteúdo genômico destas espécies (EL-SAYED et al., 2005b). Da montagem do *T. cruzi* (cepa CL Brener) foram obtidos 8.780 *scaffolds*, totalizando um tamanho de genoma haploide de aproximadamente 67 milhões de pares de bases (Mpb). O principal ponto de discussão acerca deste resultado foram os problemas na contiguidade do genoma montado, devido principalmente ao colapso ou a montagem incorreta das diversas regiões repetitivas do genoma (EL-SAYED et al., 2005a). Chama a atenção o fato de que quase 20 anos depois da publicação do primeiro genoma deste parasito, ainda não há uma versão de seu genoma à nível cromossômico. Para diferentes cepas do *T. cruzi*, diversos são os trabalhos que selecionam *scaffolds* para que sejam chamados de cromossomos artificiais ou pseudo-cromossomos, baseando-se em parâmetros como tamanho do *scaffold*, número de genes ou quantidade de repetições (WEATHERLY; BOEHLKE; TARLETON, 2009; BERNÁ et al., 2018; REIS-CUNHA; BARTHOLOMEU, 2019; CALLEJAS-HERNÁNDEZ et al., 2019; WANG et al., 2021; TALAVERA-LÓPEZ et al., 2021).

Em 2014 foi publicada a primeira versão do genoma da cepa SC58 de *T. rangeli* (STOCO et al., 2014). Nesta publicação foram obtidos 9.066 *scaffolds* e o tamanho do genoma haploide da espécie foi estimado em cerca de 24 Mpb. Posteriormente, em 2018, o genoma da cepa AM80 do *T. rangeli* foi publicado, fracionado em 3.436 *scaffolds* e com um tamanho haploide do genoma de

aproximadamente 21 Mb (BRADWELL et al., 2018). Estes estudos também indicam que o cariótipo de *T. rangeli* pode conter entre 32 (STOCO et al., 2014) ou 40 cromossomos (BRADWELL et al., 2018), com a maior banda cromossômica observada em torno de aproximadamente 3,3 Mpb. De maneira independente, ambos estudos demonstram que o genoma da espécie *T. rangeli* é menor do que o genoma do *T. cruzi*, destacando também a significativa diminuição no número de genes que codificam fatores de virulência, que são as proteínas envolvidas nos processos necessários para o estabelecimento de infecção em mamíferos (STOCO et al., 2014; BRADWELL et al., 2018).

Apesar da grande divergência de sequência entre os conteúdos gênicos das diferente espécies e cepas de tripanosomatídeos, está bem estabelecido na literatura que os genomas destes organismos exibem alta sintenia, ou seja, uma alta conservação na ordem dos genes (EL-SAYED et al., 2005b; STOCO et al., 2014; BERNÁ et al., 2018). Esta observação ressalta a atuação de uma pressão seletiva para manter a ordem dos genes e preservar agrupamentos de genes. Tais agrupamento de genes são uma característica marcante do genoma de tripanosomatídeos, cuja organização é estruturada em conjuntos de genes direcionais (do inglês, *Directional Gene Clusters* – DGCs). Em algumas espécies, estes conjuntos podem conter centenas ou até milhares de genes não sobrepostos, presentes na mesma orientação da fita de DNA, sendo separados apenas por regiões de troca de fita (REIS-CUNHA et al., 2022).

Os DGCs são transcritos de maneira policistrônica, semelhante aos operons de organismos procariotos. Os RNA mensageiros (mRNA) são processados no núcleo para gerar transcritos maduros após duas reações acopladas: trans-splicing e poliadenilação (HERREROS-CABELLO et al., 2020). Em tripanosomatídeos, estes dois processos são importantes devido à transcrição contínua de sequências gênicas que codificam proteínas (do inglês, *Coding Sequences* – CDS), caracterizando um tipo de regulação pós-transcricional (MASLOV et al., 2018). O trans-splicing leva à adição de uma sequência SL de 39 nucleotídeos, que é espécie-específica, à extremidade 5' de cada mRNA. A inserção da sequência SL confere estabilidade ao mRNA e causa a excisão das diferentes unidades de transcrição policistrônicas, o que permite suas poliadenilações e formação do respectivo mRNA maduro. A principal diferença destes DGCs para os operons de procariotos é que genes pertencentes à uma mesma unidade transcricional não são relacionados funcionalmente e têm níveis distintos de

expressão, destacando o fato de que o controle da expressão gênica em tripanosomatídeos ocorre principalmente no nível pós-transcricional (ARAÚJO; TEIXEIRA, 2011; HERREROS-CABELLO et al., 2020).

Tripanosomatídeos também realizam regulações pós-traducionais do seu conteúdo gênico, de modo que a abundância de uma determinada proteína muitas vezes não é correlacionada diretamente com os níveis detectados de seu respectivo mRNA (MASLOV et al., 2018). As modificações pós-traducionais mais comuns em tripanosomatídeos incluem: adições de glicosilfosfatidilinositol (do inglês, *glycosylphosphatidylinositol* – GPI), para ancoramento proteico na membrana celular (NAKAYASU et al., 2009); glicosilações, que auxiliam nos processos de adesão celular e na proteção do parasito contra enzimas digestivas no inseto vetor (ACOSTA-SERRANO et al., 2001) ou contra anticorpos anti-alfa-galactosil em humanos (PEREIRA-CHIOCCOLA et al., 2000); miristoilações (ROBERTS; FAIRLAMB, 2016); e palmitoilações (EMMER et al., 2011).

Por fim, com relação à maquinaria celular necessária para o processo de tradução, foram identificados ao menos 10 genes parálogos em *T. brucei* que codificam proteínas que constituem o complexo proteico eIF (do inglês, *Eukaryotic translation Initiation Factor*), como eIF4E e eIF4G, os quais fazem parte da via canônica de controle da tradução em eucariotos (FREIRE et al., 2014). Em *T. rangeli*, estão descritos três genes que codificam para eIF5A, assim como outras proteínas do processo de tradução, como fatores de alongamento de tradução eucariótica 1-alfa e 1-beta (STOCO et al., 2014).

1.4 PROTEÍNAS DE SUPERFÍCIE DE TRIPANOSOMATÍDEOS

O intricado ciclo de vida desses protozoários – que inclui mudanças severas na bioquímica e na morfologia da célula – demonstra a complexidade biológica destes organismos. Mecanismos regulatórios refinados são essenciais para que os parasitos consigam sobreviver e se multiplicar em ambientes dinâmicos, nos quais eles precisam se adaptar rapidamente a diferentes condições (HERREROS-CABELLO et al., 2020). Um dos fatores que contribuem para a adaptabilidade desses organismos é justamente sua plasticidade genômica, que engloba processos de controle da expressão gênica por meio de controles pós-transcricionais e modificações pós-

traducionais do seu repertório de proteínas de superfície, entre outros (ARAÚJO; TEIXEIRA, 2011; MUCCI et al., 2017; HERREROS-CABELLO et al., 2020).

Dados na literatura apontam que as espécies do gênero *Trypanosoma* que infectam humanos apresentam, em média, 10 mil genes em seu genoma (BERRIMAN et al., 2005; EL-SAYED et al., 2005a; STOCO et al., 2014). Ainda, espécies como *T. rangeli* compartilham até 93% do seu conteúdo gênico com outros tripanosomatídeos patogênicos (STOCO et al., 2014). Desse conjunto de genes, cerca de 50% compõem sequências repetitivas, particularmente de algumas famílias multigênicas que codificam proteínas de superfície, como: mucinas, proteínas de superfície associadas a mucinas (do inglês, *Mucin-Associated Surface Proteins* – MASP), sialidases, leishmanolisinas ou GP63, cruzipaina e amastinas (PABLOS; OSUNA, 2012; PECH-CANUL; MONTEÓN; SOLÍS-OVIEDO, 2017). Além disso, boa parte das proteínas de superfície de tripanosomatídeos são ancoradas na membrana celular através de âncoras de GPI, as quais são modificações pós-traducionais que ocorrem no retículo endoplasmático, típicas de organismos eucariotos (MUCCI et al., 2017).

Um estudo sobre as proteínas de superfície de formas tripomastigotas de *T. rangeli* confirmou a presença de proteínas do tipo GP63, glicoproteínas de adesão flagelar, mucina-*like* e sialidases, assim como destacou a importância de investigar as proteínas de superfície deste parasito para auxiliar em métodos de diagnóstico diferencial de *T. cruzi* (WAGNER et al., 2013). A expressão dessa grande variedade de proteínas na superfície da membrana celular desses parasitos, assim como suas modificações pós-traducionais, tem implicação direta nos processos de interação, de invasão, de internalização e de replicação nas células do hospedeiro, ou ainda estando ligadas à evasão do sistema imune dos mesmos (BARRIAS et al., 2012; COSTA; SILVEIRA; BAHIA, 2016). Porém, ao se analisar o número de cópias gênicas de famílias relacionadas a proteínas de superfície descritas no genoma de *T. rangeli* o que se observa é uma redução no número de genes codificantes para MASP, mucinas e sialidases, em comparação ao *T. cruzi* (STOCO et al., 2014). Além disto, famílias gênicas que codificam para proteínas de superfície estão contidas em regiões repetitivas do genoma de tripanosomatídeos, sendo melhor caracterizadas na espécie *T. cruzi* (REIS-CUNHA; BARTHOLOMEU, 2019; HERREROS-CABELLO et al., 2020). Em uma versão mais recente do genoma da cepa Sylvio X10/1 de *T. cruzi*, foi proposto que as regiões repetitivas do genoma, que contêm famílias gênicas de moléculas de superfície, são responsáveis por gerar diversidade antigênica em *T. cruzi* por meio de

eventos de recombinação e conversão gênica (TALAVERA-LÓPEZ et al., 2021). Este mesmo estudo aponta que regiões de baixa complexidade e de repetições simples costumam flanquear as sequências que codificam para proteínas de superfície. Esta configuração permite que processos de micro-homologia facilitem a recombinação entre essas sequências repetitivas e semelhantes, resultando em novas variantes de moléculas de superfície, como descrito em outros parasitos (HALL; WANG; BARRY, 2013; CLAESSENS et al., 2014; TALAVERA-LÓPEZ et al., 2021). Brevemente, a micro-homologia é um tipo de recombinação homóloga que pode acontecer entre sequências de DNA repetitivas e muito semelhantes, de tal forma que a recombinação entre essas repetições é facilitada, uma vez que enzimas que realizam a recombinação podem se ligar facilmente às sequências semelhantes em pontas livres (SFEIR; SYMINGTON, 2015).

Atualmente, a baixa contiguidade no genoma de *T. rangeli* (STOCO et al., 2014; BRADWELL et al., 2018) pode representar uma subestimativa no número total genes que codificam proteínas, principalmente de genes multicópias que codificam para proteínas de superfície. Outro desafio é a porção significativa de genes que são anotados como hipotéticos nestes genomas, que dificultam a correta identificação de características funcionais e importância metabólica de seus produtos (WAGNER et al., 2013). Genes hipotéticos são aqueles preditos por ferramentas computacionais, cujo produto gênico não apresenta evidência experimental de sua expressão. Dados do haplótipo Esmeraldo da cepa CLBrener de *T. cruzi* apontam que 47,15% do seu conteúdo gênico são de sequências hipotéticas (EL-SAYED et al., 2005), enquanto as cepas AM80 e SC58 de *T. rangeli* apresentam 56,72% e 66,25% de suas sequências anotadas como hipotéticas, respectivamente (STOCO et al., 2014; BRADWELL et al., 2018). Dito isto, muitas das proteínas anotadas como hipotéticas apresentam características de proteínas de superfície ancoradas na membrana ou secretadas, como peptídeo sinal ou âncora de GPI, assim como domínios proteicos conhecidos (COSTA et al., 2020).

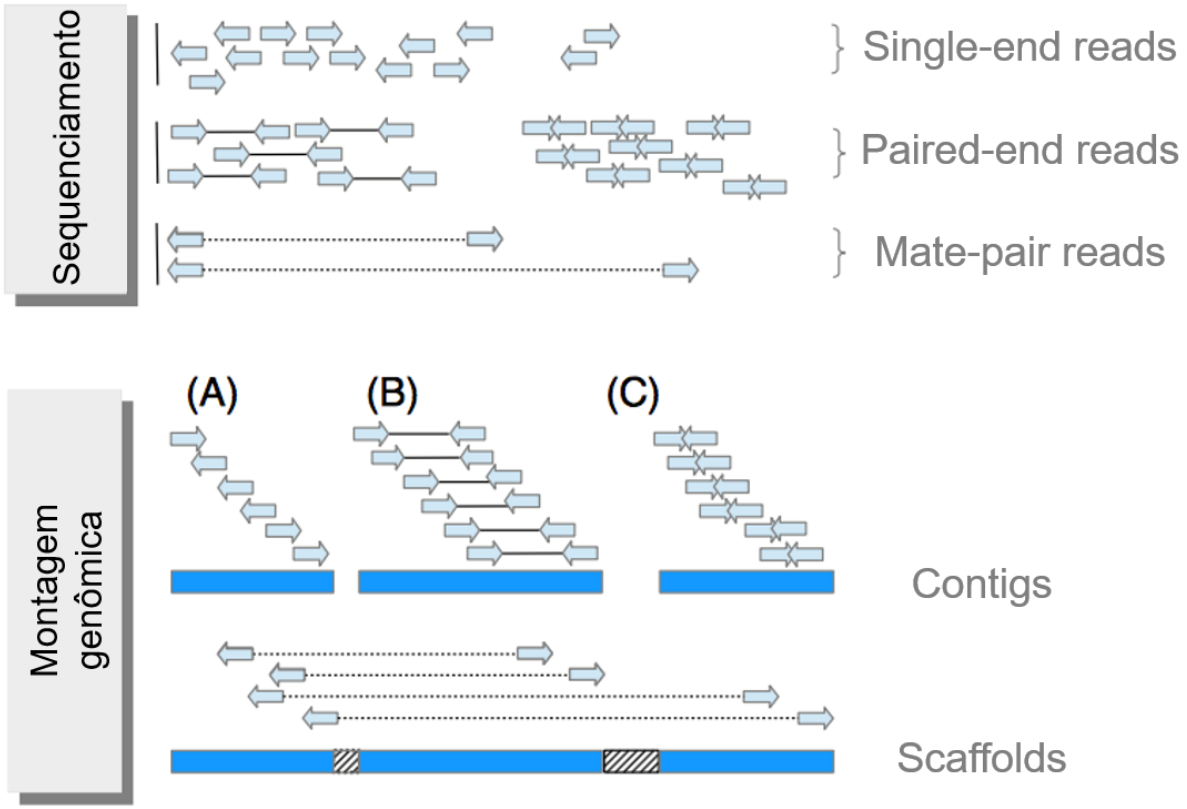
Este cenário sugere que proteínas de superfície de *T. rangeli* podem ser melhor conhecidas a partir de um conjunto de dados genômico mais completo e mais contíguo. Neste contexto, a melhor compreensão das diferentes famílias proteicas envolve a resolução destas regiões repetitivas presentes nas versões atuais dos genomas de *T. rangeli*.

1.5 SEQUENCIAMENTO E MONTAGEM DE GENOMAS

De maneira geral, o processo de sequenciamento de DNA começa com a fragmentação do DNA em segmentos, os quais servem como base para o sequenciador que fará a leitura destes, gerando fragmentos de leitura, denominados *reads*. Sequenciadores diferem no tamanho e na quantidade de *reads* gerados, por exemplo: um sequenciador do tipo Illumina (modelo HiSeq) gera *reads* em torno de 150 pares de bases (também chamados de *short reads*); enquanto um sequenciador do tipo PacBio (modelo Sequel II System) gera *reads* que chegam até 20.000 pares de bases (também chamados de *long reads*) (GOODWIN; MCPHERSON; MCCOMBIE, 2016; ILLUMINA, 2020; PACBIO, 2021).

Uma vez que o DNA tenha sido sequenciado, inicia-se a etapa de montagem do genoma em questão, cujo objetivo é reconstruir *in silico* as sequências de DNA do organismo. Ferramentas computacionais são utilizadas para averiguar a confiabilidade de cada base sequenciada, que os *reads* estejam em suas devidas orientações, assim como fazem o agrupamento dos *reads* baseando-se na identidade de suas contiguidades (YE et al., 2012). Desta forma, em ordem crescente de grandeza e tamanho, os agrupamentos de *reads* são denominados *contigs* que, por sua vez, também são agrupados e ordenados em longas sequências contínuas, passando a ser denominados de *scaffolds*. Quando o número de *scaffolds* aproxima-se ou torna-se igual ao número determinado de cromossomos de uma determinada espécie, com elevada confiabilidade, obtém-se a montagem do genoma completo (Figura 2). Entretanto, inúmeros são os genomas que não chegam ao nível de montagem cromossômica completa (WAJID; SERPEDIN, 2012).

Figura 2. Ilustração do processo de montagem genômica e de suas terminologias.



Legenda: As setas representam os *reads* obtidos de sequenciamento e seus respectivos sentidos (senso e antissenso). *Reads* do tipo *single-end* são obtidos quando apenas um dos extremos de um fragmento de DNA é sequenciado. *Reads* do tipo *paired-end* são gerados quando ambos os extremos do fragmento de DNA são sequenciados. *Reads* do tipo *mate-pair* originam-se de fragmentos de DNA distintos e complementares, que apresentam uma distância maior do que 500 pares de bases entre si. (A) *reads* do tipo *single-end* são ordenados para formação de *contigs*. (B) *reads* do tipo *paired-end* com inserção são ordenados para formação de *contigs*. Em (C) *reads* do tipo *paired-end* sem inserção são ordenados para formação de *contigs*. Após, os *contigs* são ordenados para a formação de *scaffolds*. Fonte: adaptado de EKBLOM; WOLF, 2014.

Ainda, a montagem do genoma pode ser abordada de duas maneiras distintas: (i) montagem *de novo*, quando são utilizados apenas os *reads* oriundos do sequenciamento, sem que seja fornecido nenhum tipo de modelo de comparação, orientação ou outra sequência para guiar o processo de montagem; ou (ii) montagem com referência, quando as sequências geradas são comparadas a um genoma previamente montado da mesma espécie ou de uma espécie filogeneticamente próxima (POP, 2004; YE et al., 2012). Geralmente, é realizada uma montagem com referência quando já existe um genoma de referência pré-estabelecido para um

organismo e o objetivo experimental é determinar variações em relação a ele. Este método é ideal para detecção de polimorfismos de nucleotídeos únicos (em inglês, *Single Nucleotide Polymorphisms* – SNPs), inserções ou deleções nucleotídicas (em inglês, *insertion-deletion* ou *indels*) e variações no número de cópias gênicas (do inglês, *Copy Number Variation* – CNV), mas não é uma abordagem eficaz para montar rearranjos ou quebras em regiões sintênicas. Além disso, uma desvantagem significativa dessa abordagem é que ela não revelará características genômicas que não estão presentes na referência utilizada (SOHN; NAM, 2016; BAPTISTA; KISSINGER, 2019). Por outro lado, montagens *de novo* são a única opção para a primeira montagem do genoma de um novo organismo, permitindo o estabelecimento de sequências de referência, o reconhecimento de variabilidades estruturais e rearranjos no genoma. A principal vantagem deste método é que ele possibilita a descoberta de novas características genômicas de uma determinada espécie. O ponto fraco da montagem de novo é a necessidade de um grande poder computacional para sua realização, o que se torna especialmente problemático para a montagem de genomas de organismos eucariotos (SOHN; NAM, 2016; BAPTISTA; KISSINGER, 2019; DIDA; YI, 2021).

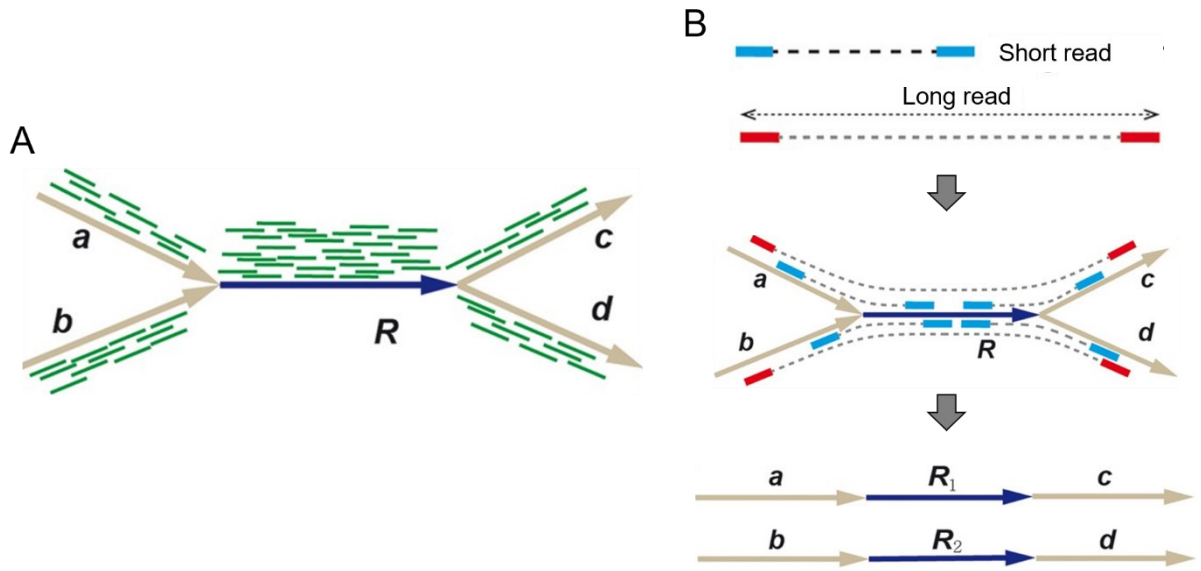
Por fim, mais relevante para este trabalho, existe a possibilidade de utilizar dados de *long reads* e *short reads* durante uma montagem *de novo*. Este tipo de montagem é denominada de montagem híbrida, pois utiliza dados oriundos de diferentes plataformas de sequenciamento. A principal vantagem de uma montagem híbrida é a combinação dos pontos fortes das tecnologias envolvidas. Neste sentido, os *long reads* provêm a base da estrutura do genoma, possibilitando a correta organização de suas variações estruturais ou de regiões repetitivas, enquanto os *short reads* provêm a resolução das regiões genômicas.

1.6 ELEMENTOS REPETITIVOS NO GENOMA DE TRIPANOSOMATÍDEOS

Mesmo com o avanço das tecnologias de sequenciamento e dos protocolos de montagem genômica, um dos maiores desafios para a montagem de genomas de tripanosomatídeos continua a ser o grande número de elementos repetitivos em seus genomas (EL-SAYED et al., 2005b; PITA et al., 2019). No caso destes organismos, o que se observa é um colapso de cobertura em regiões repetitivas do genoma durante a etapa de montagem (Figura 3), porque os programas de montagem não conseguem

lidar com a organização de centenas ou milhares de cópias gênicas e de sequências repetitivas dispostas em tandem (BERNÁ et al., 2018; WANG et al., 2021).

Figura 3. Ilustração do colapso de regiões repetitivas durante uma nova montagem genômica e sua possível resolução através do método de montagem híbrida.



Legenda: (A) A cobertura de *reads* (verde) é maior em uma região repetitiva do genoma (*R*) do que em outras regiões (*a, b, c, d*). (B) A região repetitiva do genoma (*R*) pode ser resolvida utilizando *long reads*, com apoio de dados de *short reads*, o que caracteriza uma montagem híbrida. Fonte: adaptado de SOHN; NAM, 2016.

Dados na literatura apontam que cerca de metade de todo o conteúdo genômico de tripanosomatídeos é composto por elementos repetitivos, os quais estão preferencialmente dispostos em regiões de macro- e microssatélites, regiões subteloméricas, complementados por elementos genômicos móveis e genes multicópias de famílias multigênicas dispersos por todo o seu genoma (EL-SAYED et al., 2005b; PITA et al., 2019). Estes elementos podem ser divididos em dois tipos: (i) elementos repetitivos estruturais; e (ii) elementos repetitivos funcionais (REQUENA; LÓPEZ; ALONSO, 1996).

Repetições do tipo macro- e microssatélites, regiões subteloméricas e elementos genômicos móveis são exemplos de elementos repetitivos estruturais, além de regiões canônicas da estrutura genômica de organismos eucariotos como centrômeros e telômeros (CHARLESWORTH; SNIEGOWSKI; STEPHAN, 1994; OBADO et al., 2007). Outro tipo de elemento repetitivo estrutural bastante representativo no genoma de tripanosomatídeos são as repetições intercaladas (do

inglês, *interspersed repeats*), as quais podem ser curtas (do inglês, *Short Interspersed Nuclear Elements* – SINEs) ou longas (do inglês, *Long Interspersed Nuclear Elements* – LINEs). Este tipo de repetição representa centenas ou milhares de cópias de elementos transponíveis, que se transpõem por meio de uma via baseada em DNA – sendo classificados como transposons de DNA – ou elementos que requerem uma via de transcrição reversa, a partir de um intermediário de RNA – sendo classificados como retroelementos (CHARLESWORTH; SNIEGOWSKI; STEPHAN, 1994; WICKSTEAD; ERSFELD; GULL, 2003). Os elementos repetitivos estruturais desempenham papéis importantes na evolução dos genomas de tripanosomatídeos, ocasionando rearranjos cromossômicos por meio de eventos de translocações, inversões e duplicações gênicas, que garantem a manutenção da diversidade genômica nestes organismos. (SOUZA et al., 2007; LIMA et al., 2013). A literatura também aponta que estes elementos podem: gerar novos genes, que pode acontecer quando um elemento móvel se insere em uma região codificante, causando uma possível alteração de função; e ajudar na sobrevivência destes parasitos, uma vez que um alto conteúdo de repetições e elementos móveis no genoma permite que organismos se adaptem rapidamente a ambientes diferentes (WICKSTEAD; ERSFELD; GULL, 2003; SCHRADER; SCHMITZ, 2018).

No genoma de tripanosomatídeos, elementos repetitivos funcionais correspondem a genes multicópias que fazem parte de famílias multigênicas. Estas famílias gênicas podem estar distribuídas em diferentes cromossomos destes parasitos, na forma de múltiplos grupos de cópias gênicas repetidas, comumente dispostos em tandem. A organização desses agrupamentos é complexa, onde os genes de cada família multigênica podem, ou não, estarem agrupados homogeneamente, com combinações entre as diferentes famílias, de forma intrincada, às vezes alternando até o senso da fita codificadora (REIS-CUNHA et al., 2022). Como exposto anteriormente, as famílias multigênicas mais estudadas são aquelas cujos genes codificam para proteínas de superfície, devido à sua importância nos processos de invasão celular e evasão do sistema imune do hospedeiro. Um exemplo de família multigênica bastante estudada no genoma de *T. cruzi* é a família das mucinas, cujas estimativas apontam a presença de mais de 750 cópias de genes que codificam proteínas desta família multigênica (BUSCAGLIA et al., 2006). Brevemente, as mucinas são glicoproteínas que apresentam uma densa matriz de oligossacarídeos em sua composição, as quais são essenciais para os mecanismos de infecção do

parasito, pois estão envolvidas nos processos de reconhecimento, adesão e internalização celular do parasito na célula hospedeira (MUCCI et al., 2002; DE SOUZA; DE CARVALHO; BARRIAS, 2010).

Em conjunto, esses dois grupos de elementos genômicos desempenham papeis essenciais na biologia de tripanosomatídeos, pois são responsáveis por influenciar a sobrevivência, a diversidade, a infectividade e a patogenicidade destes parasitos. Desta forma, a correta predição e anotação gênica das famílias multigênicas no genoma de *T. rangeli* é fundamental.

1.7 PREDIÇÃO E ANOTAÇÃO GÊNICA

Em sequência à etapa de montagem do genoma, é necessário contextualizar biologicamente o novo conjunto de dados obtido. Esta etapa de identificação do conteúdo gênico de um genoma recém montado chama-se predição gênica, a qual consiste em buscar e caracterizar janelas abertas de leitura (do inglês, *Open Reading Frames* – ORF), CDS, genes que codificam RNA e outras regiões funcionais como, por exemplo, regiões promotoras e terminadoras (STEIN, 2001; KORF, 2004).

Assim que a etapa de predição gênica é finalizada, é necessário atribuir função aos genes e seus produtos gênicos, quando possível. O processo pelo qual novos dados de sequências biológicas oriundas da predição gênica de uma espécie são analisados, com a finalidade de se atribuir sua função e seu contexto biológico, chama-se anotação gênica (STEIN, 2001). Essa contextualização de dados moleculares *in silico* é uma das etapas mais importantes e um dos maiores desafios para as análises de sequências biológicas. Geralmente, a identificação da possível função desta sequência gênica predita é inicialmente realizada através da busca automatizada por similaridade com sequências previamente anotadas, que estão depositadas em bancos de dados (MORIYA et al., 2007). Esta análise de similaridade entre sequências biológicas é comumente desempenhada através da utilização de algoritmos de comparação e alinhamento de sequências, sendo o algoritmo Smith-Waterman, usado pelos programas do pacote BLAST+, o mais utilizado (CAMACHO et al., 2009). Desta forma, realiza-se a transferência de anotação da sequência depositada em banco de dados para a sequência gênica predita, respeitando-se critérios de similaridade entre as sequências biológicas estudadas. É importante destacar que não existem critérios fixos para classificar sequências como similares ou

não, de forma que tais critérios precisam ser adequados ao desenho experimental, às sequências estudas e ao conjunto de dados utilizados para a respectiva comparação (PEARSON, 2013).

Além da anotação gênica automática, a revisão manual das anotações realizadas pelos programas computacionais faz-se necessária para a redução de redundâncias, para a correta identificação das características gênicas e funções biológicas dos genomas, sendo uma tarefa lenta, contínua e laboriosa (OUZOUNIS; KARP, 2002). As bases de dados que passaram por este processo são denominadas "curadas" e usualmente constituem bancos especializados, com constante atualização. A qualidade da anotação gênica é, portanto, diretamente ligada à qualidade dos bancos de dados utilizados neste processo e, quando possível, deve ainda ser validada de forma experimental. Para a etapa de anotação gênica deste trabalho o banco de dados especializado mais relevante é o *Kinetoplastid Informatics Resources*, conhecido como TriTrypDB (disponível em: https://tritrypdb.org/tritrypdb/), que reúne informações acerca de diferentes espécies de organismos tripanosomatídeos, o qual faz parte do VEuPathDB – *Eukaryotic Pathogen, Vector & Host Informatics Resources* (AMOS et al., 2021).

1.8 GENÔMICA COMPARATIVA E HOMOLOGIA DE SEQUÊNCIAS BIOLÓGICAS

Uma vez que os genes tenham sido devidamente preditos e anotados, é possível realizar estudos de genômica comparativa, que é definida como ramo da genética, onde características genômicas como estrutura, sequência e função dos genomas de diferentes espécies são estudados. Os objetos de estudo da genômica comparativa podem ser divididos em três principais tipos de análises, que envolvem: (i) estrutura genômica, incluindo regiões de repetição, rearranjos, sintenia, pontos de quebra e conteúdo G+C; (ii) regiões não-codificantes, que estuda regiões regulatórias; e (iii) regiões codificantes, que inclui o estudo do conteúdo proteico e das relações de homologia entre genes (WEI et al., 2002; XIA, 2013). Nos últimos anos, estudos de genômica comparativa se tornaram mais amplos e precisos graças ao advento das tecnologias de sequenciamento em larga escala, os quais provêm maior qualidade e confiabilidade das bases sequenciadas, sendo possível utilizar uma grande variedade

de marcadores nas abordagens e análises *in silico*, genéticas e de biologia celular para estudar os genes de um determinado genoma (PEVSNER, 2015).

Ainda, é importante considerar a história evolutiva das espécies para a realização da comparação genotípica, onde apenas é possível investigar as diferenças e as similaridades entre os genes, os transcritos e as proteínas de diferentes organismos a partir das relações evolutivas entre as espécies estudadas. Um exemplo claro da relação evolutiva entre sequências gênicas de diferentes organismos é justamente a homologia entre genes, os quais podem ser classificados como genes ortólogos ou genes parálogos (ALTENHOFF; DESSIMOZ, 2012). Genes ortólogos são aqueles que possuem origem evolutiva comum e divergiram por um processo de especiação, normalmente mantendo a função biológica original observada no organismo ancestral, enquanto genes parálogos originam-se através de eventos de duplicação gênica, podendo manter ou não as suas funções biológicas originais (ALTENHOFF; DESSIMOZ, 2012).

Desta forma, a genômica comparativa é uma importante ferramenta para também se entender aspectos evolutivos das espécies estudadas. Destaca-se também a possibilidade de identificação de genes conservados – em outras palavras, preservados em diferentes organismos ao longo da história evolutiva da espécie – que é um passo importante para a compreensão do próprio genoma. Esta abordagem comparativa ajuda a estabelecer quais genes desempenham papéis importantes nos organismos, o que por sua vez pode se traduzir em novas abordagens para o tratamento de doenças e melhoria da condição de saúde humana (NIH, 2020).

2 RELEVÂNCIA

Dados dos Centros de Controle e Prevenção de Doenças dos Estados Unidos (do inglês, *Centers for Disease Control and Prevention* – CDC) estimam que, em 2017, ao menos sete milhões de pessoas estavam infectadas com o *T. cruzi* nas Américas. Acredita-se que cerca de 30% das pessoas infectadas desenvolvem algum sintoma da Doença de Chagas, que vão desde: febre, surgimento de ínguas e inflamação das meninges; ou patologias associadas, como aumento do volume do baço e do fígado, acidente vascular cerebral e infarto agudo do miocárdio. Mundialmente, essa doença causa aproximadamente 10 mil mortes por ano (CDC, 2017).

Estudos que avaliam o padrão de dispersão dessa doença pelo mundo demonstram que países europeus que recebem imigrantes das Américas estão observando um aumento de casos de cardiomiopatias associadas à pacientes chagásicos (BASILE et al., 2011; ANTINORI et al., 2017). Ainda, estima-se que os gastos globais em cuidados de saúde associados ao tratamento da Doença de Chagas sejam em torno de 630 milhões de dólares anualmente, que é um montante considerável quando se leva em consideração o contexto de pobreza associado a pacientes chagásicos na América Latina, em especial no Brasil (LEE et al., 2013).

Segundo dados do boletim epidemiológico do Ministério da Saúde do Brasil, para o ano de 2020, a região Norte do país representa 95% dos novos casos de Doença de Chagas. O perfil demográfico deste levantamento indica que 63,70% dos casos acometem jovens adultos do sexo masculino, de raça autodeclarada parda (BRASIL, 2021). Este cenário é agravado pelos registros de infecções mistas pelo *T. cruzi* e o *T. rangeli* em regiões endêmicas, tanto nos vetores triatomíneos quanto nos hospedeiros humanos, de forma que o diagnóstico desta doença costuma ser inespecífico justamente pela soro-reatividade cruzada que ocorre entre os antígenos destas espécies (PORCEL et al., 1996; GRISARD et al., 1999; MORAES et al., 2008; MAIA DA SILVA et al., 2009). A aposentadoria precoce, estipulada pela legislação brasileira para indivíduos com sorologia positiva para *T. cruzi*, representa outra faceta dos desafios socioeconômicos potencialmente decorrentes de infecções mistas, causada por resultados falso-positivos de pessoas infectadas pelo *T. rangeli*. Ainda, levando em consideração um contexto global, a modificação dos habitats de insetos vetores devido a fatores como desmatamento, expansão urbana e mudanças climáticas pode ampliar as áreas propensas a infecções por esses organismos,

contribuindo para o aumento de infecções e maior incidência da doença (MEDONE et al., 2015; LIDANI et al., 2019).

Desta forma, evidencia-se a necessidade de se obter uma nova versão do genoma do *T. rangeli*, que é uma espécie não patogênica para o hospedeiro mamífero, combinando dados de sequenciamento de tecnologias distintas, para uma análise comparativa com outras espécies de tripanosomatídeos patogênicos. De posse desta nova versão, é possível focar principalmente no que diz respeito às regiões repetitivas do genoma dos parasitos, visando compreender as diferenças estruturais e funcionais destas regiões entre as espécies. São justamente estas regiões repetitivas no genoma de tripanosomatídeos onde há perda de sintenia entre espécies e a maior concentração de proteínas que participam dos processos de sinalização e comunicação celular na interface parasito-hospedeiro (BERNÁ et al., 2018; MÜLLER et al., 2018). Nas espécies patogênicas, tais processos estão diretamente relacionados ao estabelecimento da infecção e ao desenvolvimento de doenças nos hospedeiros humanos (CARDOSO; REIS-CUNHA; BARTHOLOMEU, 2016). Desta forma, o resultado de uma análise de genômica comparativa focada na análise de regiões repetitivas nos genomas das espécies patogênicas e a não patogênica do gênero *Trypanosoma* representa uma fonte para a exploração e caracterização de novos alvos para o avanço dos métodos de diagnóstico e o desenvolvimento de fármacos, dentre outros.

A apresentação deste documento será feita em dois capítulos, contemplando os experimentos realizados e resultados obtidos durante o período de doutoramento. O **Capítulo I** se refere ao artigo de desenvolvimento e validação do *AnnotaPipeline*, que é uma ferramenta de anotação gênica automática que utiliza uma abordagem proteogenômica para predição e anotação de CDS de organismos eucarióticos. O **Capítulo II** se refere às análises para a obtenção de uma nova montagem genômica da cepa SC58 de *T. rangeli* utilizando novos dados de sequenciamento, assim como os resultados obtidos das análises *in silico* iniciais deste novo conjunto de dados.

3 HIPÓTESE

O refinamento do genoma do *T. rangeli*, por meio da utilização de dados de sequenciamento mais modernos e montados de forma híbrida, permitirá uma comparação mais completa entre o conteúdo genômico de tripanosomatídeos patogênicos e a espécie não patogênica.

4 OBJETIVOS

4.1 OBJETIVO GERAL

Gerar uma nova versão do genoma de *Trypanosoma rangeli* (cepa SC58), combinando dados obtidos através de diferentes tecnologias de sequenciamento, e realizar um estudo comparativo de aspectos estruturais e funcionais com outras espécies de tripanosomatídeos.

4.2 OBJETIVOS ESPECÍFICOS

- Desenvolver uma *pipeline* que utilize dados experimentais de transcriptômica e proteômica para validação da predição e anotação gênica em organismos eucariotos;

- Utilizar uma abordagem proteogenômica para realizar predição e anotação gênica na nova versão do genoma de *T. rangeli* SC58;

- Identificar, analisar e comparar *in silico* proteínas de superfície obtidas na nova versão do genoma de *T. rangeli* SC58 com outros tripanosomatídeos;

- Identificar o genoma central compartilhado entre espécies patogênicas de tripanosomatídeos e a nova versão do genoma de *T. rangeli* SC58;

- Investigar regiões repetitivas e a presença de genes multicópias associados a processos envolvidos na interação parasito-hospedeiro na nova versão do genoma de *T. rangeli* SC58 e outros tripanosomatídeos.

5 CAPÍTULO I: ANNOTAPIPELINE – AN INTEGRATED TOOL TO ANNOTATE EUKARYOTIC PROTEINS USING MULTI-OMICS DATA

5.1 CONTEXTUALIZAÇÃO

Com base na necessidade de se estabelecer um fluxograma que utilizasse simultaneamente dados genômicos, transcriptômicos e proteômicos obtidos de bases de dados públicas ou gerados previamente pelo nosso grupo de pesquisa para realizar a anotação ou a reanotação de CDS de diferentes tripanosomatídeos, buscamos o desenvolvimento de uma ferramenta computacional que permitisse tal análise.

Resumidamente, a *pipeline* desenvolvida utiliza uma abordagem proteogênomica para validar as predições e anotações gênicas, ao mesmo tempo em que permite a personalização dos bancos de dados consultados e de palavras-chaves para classificação das CDS. Assim, esta ferramenta para anotação automática recebeu o nome de “AnnotaPipeline”, cuja descrição é “ferramenta integrada para anotação de proteínas de organismos eucariotos, que utiliza dados multi-ômicos” (MAIA et al., 2022).

Neste estudo, foram realizados testes de validação com dados genômicos, transcriptômicos e proteômicos de organismos modelo, como *Arabidopsis thaliana*, *Caenorhabditis elegans* e *Candida albicans*, obtidos dos bancos de dados do GenBank, BioProject e ProteomeXchange, tendo sido o trabalho publicado na revista “Frontiers In Genetics”, o qual apresentamos a seguir (MAIA et al., 2022).

Durante o seu desenvolvimento, a *pipeline* foi utilizada pelo grupo de pesquisa para realizar anotação genômica automática de diversos organismos como *Eimeria* spp., *Leishmania infantum*, *T. cruzi* e *T. rangeli*.

Frontiers In Genetics. 2022, 13: 1020100. DOI: 10.3389/fgene.2022.1020100

AnnotaPipeline: An integrated tool to annotate eukaryotic proteins using multi-omics data.

Guilherme Augusto Maia[1], Vilmar Benetti Filho[1], Eric Kazuo Kawagoe[1], Tatiany Aparecida Teixeira Soratto[1], Renato Simões Moreira[1,2], Edmundo Carlos Grisard[1,3] and Glauber Wagner[1,3*]

[1] Laboratório de Bioinformática, Universidade Federal de Santa Catarina (UFSC), Campus João David Ferreira Lima, Florianópolis, Brazil. [2] Instituto Federal de Santa Catarina (IFSC), Campus Lages, Lages, Brazil. [3] Laboratório de Protozoologia, Universidade Federal de Santa Catarina (UFSC), Campus João David Ferreira Lima, Florianópolis, Brazil.

* Correspondence: glauber.wagner@ufsc.br

5.2 ABSTRACT

Assignment of gene function has been a crucial, laborious, and time-consuming step in genomics. Due to a variety of sequencing platforms that generates increasing amounts of data, manual annotation is no longer feasible. Thus, the need for an integrated, automated pipeline allowing the use of experimental data towards validation of in silico prediction of gene function is of utmost relevance. Here, we present a computational workflow named AnnotaPipeline that integrates distinct software and data types on a proteogenomic approach to annotate and validate predicted features in genomic sequences. Based on FASTA (i) nucleotide or (ii) protein sequences or (iii) structural annotation files (GFF3), users can input FASTQ RNA-seq data, MS/MS data from mzXML or similar formats, as the pipeline uses both transcriptomic and proteomic information to corroborate annotations and validate gene prediction, providing transcription and expression evidence for functional annotation. Reannotation of the available *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Candida albicans*, *Trypanosoma cruzi*, and *Trypanosoma rangeli* genomes was performed using the AnnotaPipeline, resulting in a higher proportion of annotated proteins and a reduced proportion of hypothetical proteins when compared to the annotations publicly available for these organisms. AnnotaPipeline is a Unix-based pipeline developed using Python and is available at: https://github.com/bioinformatics-ufsc/AnnotaPipeline.

Keywords: workflow, proteogenomics, genome annotation, functional annotation, hypothetical proteins

5.3 INTRODUCTION

Genome annotation involves a detailed description and understanding of the genome structure and assignment of biological functions to the genes (Stein, 2001). Structural annotation thus characterizes the physical structure of coding and non-coding regions on a given genome, resulting in a physical map of the genes' number and positioning. Along determination of the structure and organization of the protein-coding sequences (CDS) located within open reading frames (ORF) of each gene, annotation also includes a description of other genomic elements such as promoters and enhancers (Korf, 2004; Danchin et al., 2018). Several computational tools known as gene predictors, such as AUGUSTUS (Stanke and Waack, 2003) and GeneMark (Brůna, Lomsadze, and Borodovsky, 2020), have been widely used to perform structural annotation (Yandell and Ence, 2012).

Functional annotation consists of assigning biological information to genes, such as their involvement in biological processes, molecular functions, presence of functional protein domains, and subcellular localization, among others (Stein, 2001; Yandell and Ence, 2012). The assignment of biological functions to protein-coding genes is generally performed through similarity analysis with databases containing previously annotated protein sequences using sequence aligners such as BLAST (Camacho et al., 2009) or DIAMOND (Buchfink, Reuter, and Drost, 2021). The biological function of a predicted CDS is therefore assumed to be the same as the protein in the database that demonstrates the most significant similarity, leading to an annotation transfer (Hegyi and Gerstein, 2001). Thus, the accuracy of the annotated database is fundamental for genome annotation, allowing the quality of downstream analyses based on the transferred annotations. Especially with the use of high-throughput sequencing during the past years, several public genomic and proteomic databases from a variety of organisms are nowadays available. However, the exponential growth of datasets impairs the quality of a proper and detailed structural and functional annotation of genomes. For that, the use of curated databases such as SwissProt/UniProtKB (The UniProt Consortium, 2021) and Ensembl (Flicek et al., 2014), or even organism-specific databases, such as those contained in the VEuPathDB (Amos et al., 2022), is highly recommended to ensure high quality to the genome annotation.

Considering the growing datasets of genomic and proteomic databases, and the specific genomic features across taxa, combining different computational tools or pipelines to automatically assess gene structural and functional annotation has been widely used (Danchin et al., 2018). Composed of a set of data processing methods connecting inputs and outputs in series, automated pipelines can perform genome annotation by sequence similarity (Hyatt et al., 2010; Steinbiss et al., 2016) or functional annotation of proteins (Gotz et al., 2008; Vlasova et al., 2021; Törönen and Holm, 2022). Nevertheless, only a few genome annotation pipelines use expression experimental data (RNA-Seq or MS/MS) to validate the *in silico* annotation (Ghali et al., 2014; Sheynkman et al., 2014).

Large-scale genomic and transcriptomic studies based on high-throughput sequencing platforms in the past decade have provided increasing amounts of data (Kumar et al., 2016a), also providing extensive gene expression profiles based on transcribed RNAs (RNA-seq) sequencing. Moreover, extensive proteomic data acquired from sensitive mass spectrometry (MS) technologies are available from several databases (Vaudel et al., 2016), such as PRIDE (Perez-Riverol et al., 2022), MassIVE (Miao et al., 2012), and the ProteomeXchange Consortium (Vizcaíno et al., 2014). Thus, using transcription and expression evidence to annotate newly predicted CDS or reannotate formerly analyzed genomes would reveal novel biological aspects. The proteogenomic approach allows the cross-validation of genomic, transcriptomic, and proteomic data on both intra- and inter-specific analyzes (Nesvizhskii, 2014). However, this approach requires novel computational methods and pipelines. Thus, integrating the classic annotation analysis by sequence similarity with customizable parameters and databases, combined with functional prediction validated with RNA-seq and MS/MS data evidence, would enhance genome annotation as an essential step toward comprehending biological mechanisms.

In this study, we developed AnnotaPipeline, a proteogenomic computational tool for automatic annotation of eukaryotic genomes using support from high-throughput transcriptomic and proteomic data, allowing validation of gene function and expression.

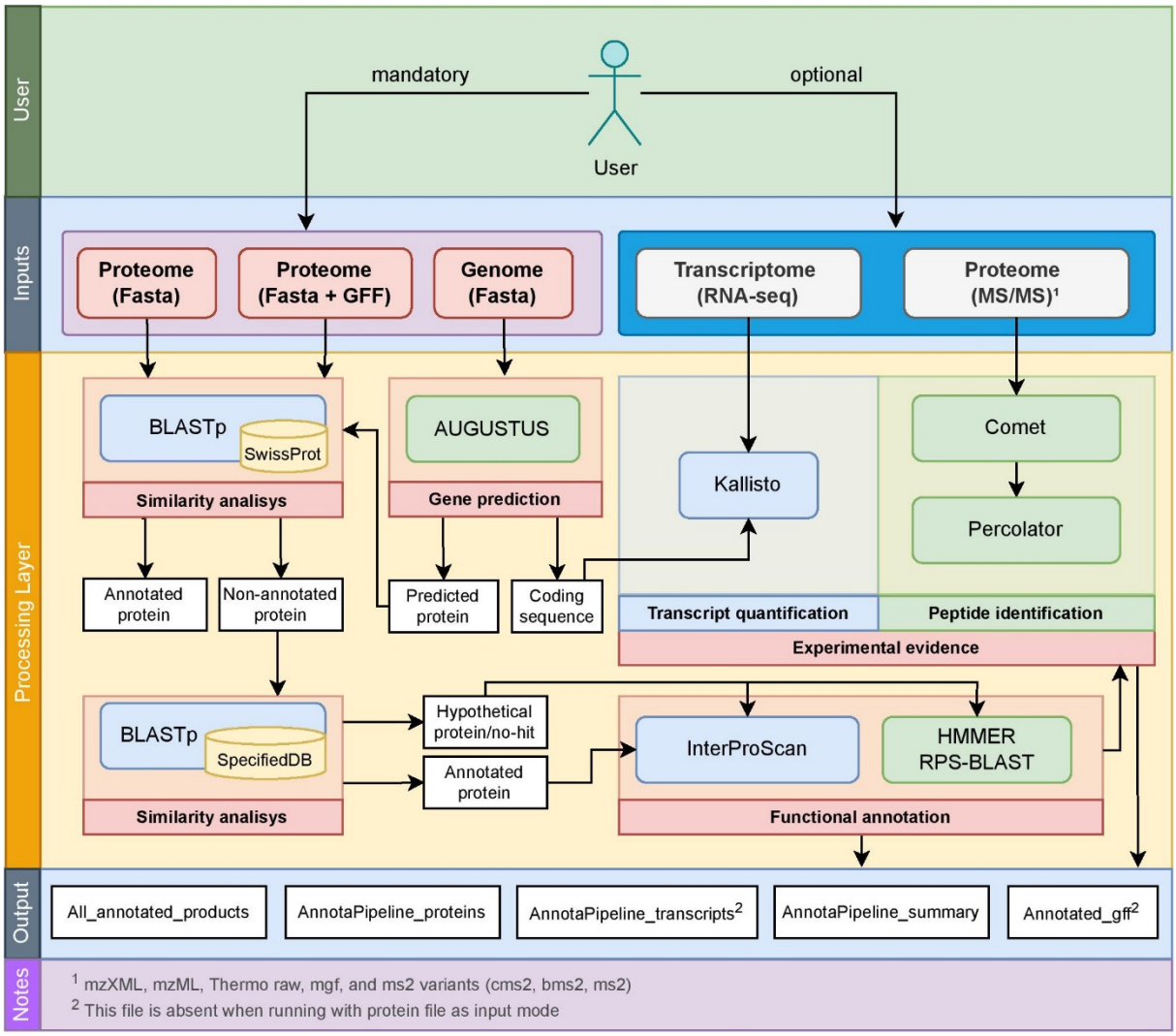
5.4 METHODS

5.4.1 AnnotaPipeline

5.4.1.1 Development and overview

The AnnotaPipeline overall scheme and processes are shown in Figure 1. This pipeline was developed using Python and runs on Unix-based systems, consisting of a series of tolls and in- house scripts for data preparation, processing, and analysis. Documentation related to installation instructions and scripts to run AnnotaPipeline are available at https://github.com/bioinformatics-ufsc/AnnotaPipeline.

Figure 1. Overview of AnnotaPipeline workflow, indicating the optional and the required inputs from the user, the internal processes, and the output layers.



5.4.1.2 Input and configuration files

AnnotaPipeline requires the input of at least one of the following different FASTA files: 1) a nucleotide sequence file, 2) a protein sequence file, 3) a protein sequence file, and structural annotation files in GFF3 format. If the first option is selected, AnnotaPipeline will perform gene prediction on the provided nucleotide sequence. Therefore, it is essential to use a trained AUGUSTUS model for the gene prediction process before executing the pipeline. This execution will produce an annotated GFF3, and CDS sequences will contain a complete header. For the second option, gene prediction will be skipped, and the final output file will contain only a simplified sequence header. The third option is executed equally to the second option, the pipeline will include annotations for each CDS from the provided GFF file. Also, it is recommended that the submitted GFF file is in GFF3 format, preferably from a previous AUGUSTUS gene prediction.

Aside from the molecular data input, it is also required from the user to access the YAML configuration file prior to running the pipeline, where locations of both software and databases required for the personalized analysis must be provided. Similarly, if analyses with experimental data will be carried out, it is also necessary to provide the locations of folders containing RNA-seq and MS/MS data.

Users can define the number of processing threads that will be used during the execution of the pipeline (default is set to 4 threads) and are required to define the cutoff parameters and specific keywords to classify hypothetical proteins during the similarity analysis process. This configuration step is facilitated if the user installs AnnotaPipeline using Conda from the environment file available at https://github.com/bioinformatics-ufsc/AnnotaPipeline.

5.4.1.3 Annotation process

The annotation process starting with a genomic file input is divided into three steps. Initially, gene prediction is performed by AUGUSTUS (Stanke and Waack, 2003). Although AnnotaPipeline is mainly focused on eukaryotic organisms, the pipeline accepts input of further gene prediction training models if absent in the AUGUSTUS standalone version. It is recommended to use the WebAUGUSTUS platform to generate custom training models (Hoff and Stanke, 2013).

Following gene prediction, the annotation process continues into similarity analysis performed by the BLASTp algorithm (Camacho et al., 2009) using (i) the SwissProt database, which contains about 570,000 manually curated protein sequences from a wide variety of organisms (The UniProt Consortium, 2021), and (ii) a user-specified database such as TrEMBL/UniProtKB, VEuPathDB and GenBank NR, or additional databases that must be specified in the AnnotaPipeline.yaml configuration file. Despite the used database, the pipeline contains parsing scripts that automatically will transfer the protein annotation for the predicted CDS on the output file. Proteins are then classified into three groups: annotated proteins (known function), hypothetical proteins, and no-hit proteins. Annotated proteins are those with attributed annotation either by the SwissProt or the user-specified database. In AnnotaPipeline, hypothetical proteins are considered those presenting similarities with proteins with no specific annotation in the databases (unknown function) and that contain filter keywords in their descriptions, such as “fragment”, “hypothetical”, “partial”, “uncharacterized”, “unknown”, and “unspecified”. These are the default keywords used by the pipeline, but users can change these in the AnnotaPipeline.yaml configuration file. Annotations in subject proteins will be disregarded if at least one description contains any of the provided keywords. No-hit proteins are proteins with no available match, and therefore no annotation, in either database used in the similarity analysis step. For downstream analysis steps, the no-hit and the hypothetical proteins are grouped by the pipeline. Furthermore, proteins revealing no matches with databases and presenting no supporting evidence from experimental data are considered true negative proteins.

The third step consists of the functional annotation of proteins, starting with analyzing both annotated and hypothetical protein groups by InterProScan software (Jones et al., 2014). Exclusively for the hypothetical protein/no-hit group, further analysis using the hmmscan algorithm of the HMMER suite (Finn, Clements, and Eddy, 2011) and the RPS-BLAST (Camacho et al., 2009) are performed. The resulting functional annotation is contained in a single output file where all predicted proteins will be annotated and can be used as input for the experimental validation analyses.

5.4.1.4 Experimental validation with proteogenomic data

The AnnotaPipeline accepts the input of RNA-seq and MS/MS data that will allow experimental validation of CDS prediction and annotation. Upon activation of the

experimental analysis module, transcriptomic data will be processed by Kallisto (Bray et al., 2016), which performs a pseudo-alignment of RNA-seq reads to the annotated protein file. The result will be refined based on a quantification of aligned transcripts, which are accounted for transcripts per million (TPM). Users may concatenate their transcriptomic data into a single FASTQ file (for single-end RNA-seq) or two FASTQ files (R1 and R2, for paired-end RNA-seq) to run multiple experiments at once. For experimental validation using proteomic data (MS/MS), users can provide a single folder containing their MS/MS data files to run multiple experiments simultaneously. The search for MS/MS-derived peptides among the annotated proteins will be performed using Comet (Eng, Jahan, and Hoopmann, 2013), following the user-provided search parameters in comet.params configuration file, generating the input for the Percolator software (The et al., 2016). Then, the proteomic data will be searched among the annotated proteins dataset and parsed by the *q*-value threshold of the Percolator software.

5.4.1.5 Output files

The pipeline will create a log file and an output folder in the AnnotaPipeline directory. The log file contains details of script processing, software execution, and outputs of each computational tool. Also, this log may contain any possible warnings or errors relative to the software execution. Within the output folder, the pipeline will create (i) two FASTA files containing the annotated proteins and their respective annotated CDS, (ii) a GFF file including a transcript product field containing the final annotation for each CDS, (iii) a TXT file containing the all CDS product ID and annotated description, and (iv) a TSV file summarizing all annotated CDS and information regarding transcription (RNA-Seq) or expression (MS/MS) evidence. In addition to these main output files, within each of the folders created by AnnotaPipeline, other outputs can help the user manually curate the annotations suggested by the pipeline (Supplementary Table 1).

5.4.1.6 Comparative evaluation of AnnotaPipeline performance

Performance tests were carried out using a computational cluster equipped with 40 threads processor (3.2 GHz), 285 GB RAM memory (DDR4, 2,400 MHz), and 5 TB storage space (2.5 SATA HD, 7,200 RPM). Storage was mainly used for RNA-

seq and MS/MS data of the testing organisms. Despite the availability of computing power, the number of processing threads used for testing was set to 12 in the AnnotaPipeline.yaml configuration file.

Molecular data from three different model organisms were used to test AnnotaPipeline: *Arabidopsis thaliana* (strain TAIR10), an essential model for plant biology and genetics; *Caenorhabditis elegans* (strain WBcel235), an important model for molecular and developmental biology; and *Candida albicans* (strain SC5314), a fungal pathogen model. Genomic data for each of these organisms were retrieved from GenBank under the following accession numbers: GCA_000001735.2, GCA_000002985.3, and GCA_000182965.3, respectively. RNA-seq data for each of these organisms were obtained from BioProject/NCBI under the following accession numbers: PRJNA779571, PRJNA809747, and PRJNA750749 for *A. thaliana*; PRJNA734346, PRJNA658149, and PRJNA755869 for *C. elegans*; PRJNA714869, PRJNA496318, PRJNA752883, and PRJNA744166 for *C. albicans*. MS/MS data for each of these organisms were obtained from ProteomeXchange, under the following accession numbers: PXD012708 and PXD010730 for *A. thaliana*; PXD025128 for *C. elegans*; PXD005364 for *C. albicans*.

For the similarity analysis step, in addition to the SwissProt database, a specific database of protein sequences was used for each model organism: for *A. thaliana*, a subset of 370,680 protein sequences was obtained from the GenBank NR dataset; for *C. albicans*, the FungiDB v56 containing 2,331,868 protein sequences was obtained from VEuPathDB; and for *C. elegans*, a subset of 23,010 protein sequences was obtained from TrEMBL.

AnnotaPipeline was independently run with default parameters for every organism, using the genome FASTA file obtained for each organism as input. AUGUSTUS (version 3.4.0) prediction was performed with the gene model argument set to partial and using the prediction model dataset already provided by the software, as in: arabidopsis, for *A. thaliana*; candida_albicans, for *C. albicans*; and caenorhabditis, for *C. elegans*. Therefore, the gene prediction step was not optimized. BLASTp (version 2.12.0) execution was done assuming an *e*-value of 1e-5, the number of maximum target sequences set to 10. Also, a minimum threshold value of sequence coverage was set to 30, sequence identity 40, and sequence positivity 60 for the annotation transfer. The annotation was chosen based on the highest bit score between the analyzed sequences.

InterProScan (version 5.52–86.0) was run for the functional annotation step, allowing for the lookup of corresponding Gene Ontology annotation (--goterms). HMMscan (version 3.3.2) had the *e*-value of both sequences and domains set to 1e-5, and RPSblast (version 2.12.0) also had the minimum *e*-value of target sequences set to 1e-5. Kallisto (version 0.48.0) pseudo-alignment of RNA-seq dataset was run with 1,000 bootstraps, and the minimum threshold of TPM was selected as the mean. Comet (version 2021.01) was run for each MS/MS dataset with a scan range minimum and a maximum set to 200 and 4,000, respectively. After, Percolator (version 3.5) was run with Comet output files, and the results obtained were filtered by a *q*-value threshold of 0.05. As a complete example, all the output files from the *A. thaliana* dataset are available at https://github.com/bioinformatics-ufsc/AnnotaPipeline/blob/v1.0/Output%20Example/Annota_Athaliana.tar.xz.

The pipeline was further tested using two taxonomically close protozoa species of medical relevance containing over 50% of their CDS annotated as hypothetical proteins: *Trypanosoma cruzi* (strain Sylvio X10/1), the etiological agent of Chagas disease (Talavera-López et al., 2021) and *Trypanosoma rangeli* (strain SC58) an avirulent trypanosomatid of mammals (Stoco et al., 2014). Genomic data was retrieved from the TriTrypDB (version 57) under the following accession numbers: DS_107bdce9bb, and DS_9d0531db8e, respectively. For both organisms, the Augustus prediction model was trained online based on their respective available genome file and annotated transcripts files (tcruzi_sylviox10, for *T. cruzi*; and trypanosoma_rangeli, for *T. rangeli*). For the similarity analysis step, a database of 648,560 protein sequences obtained from the TriTrypDB was used, along with the mandatory SwissProt database. The AnnotaPipeline was run using default parameters for both trypanosomatid species, as previously mentioned.

5.5 RESULTS

5.5.1 AnnotaPipeline workflow

The complete execution of AnnotaPipeline resulted in the expected output files that were named <basename>_AnnotaPipeline_<file>.<format>, allowing users to identify the results and perform multiple experiments in the same directory by swapping the <basename> of the experiments in the AnnotaPipeline.yaml configuration file.

The generated annotation files in FASTA format display for each sequence a header containing the following information separated by a pipe character “|”: sequence identification; source organism; scaffold number; CDS start; CDS end; strand orientation; and sequence description, were functional annotations provided by GO and IPR are included. If no structural annotation GFF file is included in the analysis, information concerning strand orientation and scaffold location will be absent. Also, AnnotaPipeline changes the “transcript product” field of each CDS in the annotated GFF file to the corresponding sequence description present in the header of the FASTA file.

5.5.2 Comparative analysis of AnnotaPipeline results

AnnotaPipeline was comparatively tested using genomic data of different model organisms for which genome annotation is available. The pipeline enabled experimental evidence analyses and no gene prediction optimization. The summary of the obtained annotations, functional annotations, and experimental evidence results for the *A. thaliana*, *C. albicans*, and *C. elegans* datasets are presented in Table 1.

For *A. thaliana*, the pipeline annotated a total of 19,651 protein sequences in 29 h and 07 min; 5,377 protein sequences for *C. albicans* in 10 h and 06 min; and 14,278 protein sequences for *C. elegans* in 20 h and 58 min.

Among the genome analyzed, C. albicans had the highest percentage of annotated proteins with 99.48%, followed by A. thaliana with 98.90%. C. elegans had 22.62% of their protein sequences annotated as hypothetical proteins, and another 3.24% of proteins with no matches available in the analyzed databases. Comparatively to the current data from analyzed genomes available in public databases, AnnotaPipeline provided a higher number of annotated proteins (known function) and fewer hypothetical proteins. Consequently, the number of hypothetical proteins in the A. thaliana dataset went down from 8.75% to 1.10% using the AnnotaPipeline, while for C. elegans and C. albicans datasets, the reduction was from 34.02% to 25.85% and 38.19%–0.52%, respectively.

Table 1. Summary of AnnotaPipeline annotations, functional annotations, and experimental evidence results for different model organisms.

Parameter	*Arabidopsis thaliana* TAIR10		*Candida albicans* SC5314		*Caenorhabditis elegans* WBcel235	
	GenBank	**AnnotaPipeline**	**GenBank**	**AnnotaPipeline**	**GenBank**	**AnnotaPipeline**
Predicted proteins	27,562	19,651	6,043	5,377	19,984	14,278
Annotated proteins	25,151 (91.25%)	19,434 (98.90%)	3,735 (61.81%)	5,349 (99.48%)	13,186 (65.98%)	10,587 (74.15%)
Annotated by SwissProt	-	13,444 (69.18% of annotated)	-	2,914 (54.48% of annotated)	-	5,395 (50.96% of annotated)
Annotated by SpecificDB	-	5,990 (30.82% of annotated)	-	2,435 (45.52% of annotated)	-	5,192 (49.04% of annotated)
Hypothetical proteins	2,411	169	2,308	13	6,798	3,229
No hit proteins (true negative) *	-	48 (45)	-	15 (9)	-	462 (440)
Total hypothetical proteins	2,411 (8.75%)	217 (1.10%)	2,308 (38.19%)	28 (0.52%)	6,798 (34.02%)	3,691 (25.85%)
Proteins with at least 1 IPR term	-	17,974 (91.47%)	-	4,704 (87.48%)	-	11,050 (77.39%)
Proteins with at least 1 GO term	-	13,612 (69.27%)	-	3,705 (68.90%)	-	7,587 (53.14%)
Proteins with transcript evidence	-	3,228 (16.43%)	-	716 (13.32%)	-	1,714 (12.0%)
Proteins with peptide evidence	-	1,546 (7.87%)	-	809 (15.05%)	-	0 (%)

* True negative are proteins with no match on studied databases and no supporting evidence from experimental data, which could possibly be artifacts from gene prediction. Reference genome GenBank accessions number: *Arabidopsis thaliana* (strain TAIR10) = GCA_000001735.2; *Caenorhabditis elegans* (strain WBcel235) = GCA_000002985.3; *Candida albicans* (strain SC5314) = GCA_000182965.3.

Functional annotation of the *A. thaliana*, *C. albicans* and *C. elegans* genomes using the AnnotaPipeline revealed 69.27%, 68.90%, and 53.14% of their CDS associated with at least one GO term associated, respectively. When RNA-Seq and MS/MS data were included for the analysis of experimental evidence of transcription or expression, *A. thaliana*, *C. albicans* and *C. elegans* had 16.43%, 15.05%, and 12.00% of their annotated proteins validated with transcriptomic and proteomic data, respectively. Interestingly, no *C. elegans* annotated CDS were validated by the available MS/MS dataset.

Comparative analysis of the genome annotation for *T. cruzi* and *T. rangeli* retrieved from the TriTrypDB (version 57) and the annotation generated using AnnotaPipeline is shown in Supplementary Table 2. Although not including experimental data for validation (RNA-Seq or MS/MS), the pipeline was able to reduce the number of hypothetical proteins by 60.46% and 42.84% for *T. cruzi* and *T. rangeli*, respectively, while increasing the proportion of annotated CDS having at least one GO term assigned (Supplementary Table 2).

Considering the annotation provided by AnnotaPipeline, it is possible to classify the annotated protein sequences into eight different categories based on three different criteria: 1) available annotation based on sequence similarity with provided databases; 2) transcription evidence by quantifying RNA-seq reads; and 3) translation evidence supported by the identification of peptides matches from MS/MS information. As an example, result of the analysis of the *A. thaliana* dataset is shown in Table 2. From a total of 19,651 annotated CDS, the less represented categories are those who contains CDS having support from either RNA-Seq (12.65%) or MS/MS (4.09%) support, or both (3.78%).

5.6 DISCUSSION

Whole genome annotation is one of the first and most essential steps in any genome study, consisting in a time-consuming and laborious work depending on the genome size, and no longer can be performed manually due to the amount of data generated by high-throughput sequencing (Ouzounis and Karp, 2002). AnnotaPipeline was designed to perform automatic annotation of genomes, having the unique feature to include experimental data derived from transcriptomic (RNA-Seq) or proteomic

(MS/MS) approaches towards experimental validation of an annotated CDS. The pipeline is easy to install, runs on operating systems that support command-line options, such as Unix-based systems, and does not require high computational demands, although the time-consuming tasks can be reduced while using more robust machines. It is also user-friendly and customizable to meet the user needs in terms of analysis stringency.

Table 2. Classification table of annotated proteins by AnnotaPipeline for the Arabidopsis thaliana dataset.

Categories	Hypothetic Annotation	Transcript Evidence	Peptide Evidence	Number of Sequences	Percentage (%)
1	Yes	No	No	203	1.03
2	No	No	No	15,417	78.45
3	Yes	Yes	No	5	0.03
4	Yes	No	Yes	7	0.04
5	No	Yes	No	2,480	12.62
6	No	No	Yes	796	4.05
7	Yes	Yes	Yes	2	0.01
8	No	Yes	Yes	741	3.77

Although distinct genome annotation pipelines are available (Gotz et al., 2008; Hyatt et al., 2010; Ghali et al., 2014), AnnotaPipeline provides the possibility of using RNA-seq and MS/MS data to improve genome annotation simultaneously. Considering that proteomic data have become increasingly accessible (Nesvizhskii, 2014), and new RNA-seq technologies, such as single-cell or single-molecule sequencing, are improving significantly (Wang et al., 2019), the use of this pipeline would increase que quality and accuracy of the annotated genomes from a variety of organisms by providing several possible annotations for each protein sequence. On top of providing a more accurate automated analysis, the pipeline also offers information to support manual curation of the annotation by the user.

Comparison of the results obtained using AnnotaPipeline with the data available in public databases, it was possible to observe a reduction in the number of

hypothetical proteins for *A. thaliana* (91.0%), *C. elegans* (45.70%), and *C. albicans* (98.79%), as shown in Table 1. This reduction can be due to the use of customizable databases and keywords but also to the use of combined proteogenomic data to complement gene annotation, increasing the reliability of gene prediction and automatic annotation.

In addition to these well-annotated genomes, AnnotaPipeline also showed good performance when used to annotate the repetitive genomes from two closely related species of *Trypanosoma* (*T. cruzi* and *T. rangeli*) retrieved from TriTrypDB, both lacking RNA-seq or MS/MS data for experimental validation. It was possible to observe a relative reduction of more than 60% in the number of proteins annotated as hypothetical (Supplementary Table 2).

The use of experimental data to validate CDS annotation raises a critical discussion, especially regarding hypothetical proteins. Categorizing hypothetical proteins according to their evidence of transcription or expression by AnnotaPipeline revealed interesting results. Although presenting experimental support from RNA-Seq, MS/MS or both, as observed for *A. thaliana* proteins belonging to Class 7 (Table 2), they remain annotated as hypothetical proteins in the studied databases. In this context, annotation pipelines using this multi-omics approach can provide fundamental insights into new and uncharacterized proteins and revise those whose functions are already annotated. Knowledge areas associated with medicine would benefit most since previously annotated hypothetical proteins could now be studied and thus allow for the re-evaluation of disease diagnosis or prognostic methods (Kumar et al., 2016b).

Furthermore, AnnotaPipeline can be used to guide the exploration of proteins because it adds functional annotation to protein annotation through the incorporation of GO and IPR terms. Especially for hypothetical or uncharacterized proteins, the classical description of annotations might not be biologically informative, so the lack of functional annotations (such as GO or IPR terms) increases this information gap (Lubec et al., 2005; Gotz et al., 2008). AnnotaPipeline provides descriptive and functional information for these proteins during the automated annotation process, which helps to identify potential prediction artifacts and streamline the process of manually curating the annotations. Lastly, the AnnotaPipeline summary file can provide to users the SUPERFAMILY protein information, adding yet another layer of detail to annotations. This information can provide new insights into the functionality of

uncharacterized proteins, as they represent possibilities of new structures and functions to be explored (Lubec et al., 2005).

5.7 CONCLUSION

By integrating experimental data from RNA-seq and MS/MS analyses to validate prediction and annotations of protein-coding sequences, AnnotaPipeline, an integrated and modular genomic annotation pipeline, promoted the reduction of the number of hypothetical proteins for various organisms. The use of this original proteogenomic approach on reannotation of *A. thaliana*, *C. elegans*, *C. albicans*, *T. cruzi*, and *T. rangeli* datasets, have increased the proportion of annotated proteins, consequently reducing the number of hypothetical proteins if compared to the currently available annotation. AnnotaPipeline was developed as a generalist annotation pipeline, allowing the assessment of genomes from any eukaryotic organism with available molecular data.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

GM: participated in study design and manuscript writing. VF: participated in study design and manuscript writing; EK: participated in study design and manuscript writing. TS participated in study design and manuscript writing. RM: participated in study design and manuscript writing. EG: participated in manuscript writing. GW: participated in coordination, study design and manuscript writing.

FUNDING

This research was funded by CAPES (Coordination for the Improvement of Higher Education Personnel, Brazil) — Finance Code 001, CNPq (National Council for Scientific and Technological Development, Brazil), FAPESC (Santa Catarina Research Foundation) and UFSC (Federal University of Santa Catarina). GM and VF were

recipients of CAPES scholarship. EK was recipient of CNPq scholarship. TS was recipient of FAPESC scholarship. The open access publication fee was funded by CAPES.

ACKNOWLEDGMENTS

The development process and experimental steps were performed in the Laboratory of Bioinformatics, in the Department of Microbiology, Immunology, and Parasitology (MIP), in the Center for Biological Sciences (CCB) of the Federal University of Santa Catarina (UFSC), Brazil. The authors would like to thank Superintendencia de Governanca Eletronica e Tecnologia da Informacao e Comunicacao (SeTIC/ UFSC) for the computational infrastructure support.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR DISCLAIMER

The opinions expressed by authors contributing to this journal do not necessarily reflect the opinions of the Federal University of Santa Catarina, Brazil, or the institutions with which the authors are affiliated. The funders had no role in the study design, data analysis, or the decision to publish.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1020100/full#supplementary-material

5.8 REFERENCES

Amos, B., Aurrecoechea, C., Barba, M., Barreto, A., Basenko, E. Y., Bazant, W., et al. (2022). VEuPathDB: The eukaryotic pathogen, vector and host bioinformatics resource center. Nucleic Acids Res. 50 (D1), D898–D911. doi: 10.1093/nar/gkab929

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34 (5), 525–527. doi: 10.1038/nbt.3519

Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. Nar. Genom. Bioinform. 2 (2), lqaa026. doi: 10.1093/nargab/lqaa026

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods 18 (4), 366–368. doi: 10.1038/s41592-021-01101-x

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. BMC Bioinforma. 10 (1), 421. doi: 10.1186/1471-2105-10-421

Danchin, A., Ouzounis, C., Tokuyasu, T., and Zucker, J. D. (2018). No wisdom in the crowd: Genome annotation in the era of big data - current status and future prospects. Microb. Biotechnol. 11 (4), 588–605. doi: 10.1111/1751-7915.13284

Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. PROTEOMICS 13 (1), 22–24. doi: 10.1002/pmic.201200439

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37. doi: 10.1093/nar/gkr367

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2014). Ensembl 2014. Nucleic Acids Res. 42 (D1), D749–D755. doi: 10.1093/nar/gkt1196

Ghali, F., Krishna, R., Perkins, S., Collins, A., Xia, D., Wastling, J., et al. (2014). ProteoAnnotator - open source proteogenomics annotation software supporting PSI standards. PROTEOMICS 14 (23–24), 2731–2741. doi: 10.1002/pmic.201400265

Götz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36 (10), 3420–3435. doi: 10.1093/nar/gkn176

Hegyi, H., and Gerstein, M. (2001). Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. Genome Res. 11 (10), 1632–1640. doi: 10.1101/gr.183801

Hoff, K. J., and Stanke, M. (2013). WebAUGUSTUS — A web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 41 (W1), W123–W128. doi: 10.1093/nar/gkt418

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinforma. 11 (1), 119. doi: 10.1186/1471-2105-11-119

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-scale protein function classification. Bioinformatics 30 (9), 1236–1240. doi: 10.1093/bioinformatics/btu031

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinforma. 5, 59. doi: 10. 1186/1471-2105-5-59

Kumar, D., Bansal, G., Narang, A., Basak, T., Abbas, T., and Dash, D. (2016b). Integrating transcriptome and proteome profiling: Strategies and applications. PROTEOMICS 16 (19), 2533–2544. doi: 10.1002/pmic.201600140

Kumar, D., Yadav, A. K., Jia, X., Mulvenna, J., and Dash, D. (2016a). Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. Mol. Cell. Proteomics 15 (1), 329–339. doi: 10.1074/mcp.M114.047126

Lubec, G., Afjehi-Sadat, L., Yang, J. W., and John, J. P. P. (2005). Searching for hypothetical proteins: Theory and practice based upon original data and literature. Prog. Neurobiol. 77 (1–2), 90–127. doi: 10.1016/j.pneurobio.2005.10.001

Miao, J. J., Chen, G. Y., Du, K., and Fang, Z. J. (2012). Towards big data to improve availability of massive database. Appl. Mech. Mater. 263–266, 3326–3329. doi: 10.4028/www.scientific.net/AMM.263-266.3326

Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. Nat. Methods 11 (11), 1114–1125. doi: 10.1038/nmeth.3144

Ouzounis, C. A., and Karp, P. D. (2002). The past, present and future of genome-wide re-annotation. Genome Biol. 3 (2), COMMENT2001. doi: 10.1186/gb-2002-3-2-comment2001

Perez-Riverol, Y., Bai, J., Bandla, C., Garcia-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., et al. (2022). The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. 50 (D1), D543–D552. doi: 10.1093/nar/gkab1038

Sheynkman, G. M., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., et al. (2014). Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics 15 (1), 703. doi: 10.1186/1471-2164-15-703

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19, ii215–ii225. doi: 10.1093/bioinformatics/btg1080

Stein, L. (2001). Genome annotation: From sequence to biology. Nat. Rev. Genet. 2 (7), 493–503. doi: 10.1038/35080529

Steinbiss, S., Silva-Franco, F., Brunk, B., Foth, B., Hertz-Fowler, C., Berriman, M., et al. (2016). Companion: A web server for annotation and analysis of parasite genomes. Nucleic Acids Res. 44 (W1), W29–W34. doi: 10.1093/nar/gkw292

Stoco, P. H., Wagner, G., Talavera-Lopez, C., Gerber, A., Zaha, A., Thompson, C. E., et al. (2014). ‘Genome of the avirulent human-infective trypanosome — *Trypanosoma rangeli*’, PLoS neglected tropical diseases. PLoS Negl. Trop. Dis. 8 (9), e3176. doi: 10.1371/journal.pntd.0003176

Talavera-López, C., Messenger, L. A., Lewis, M. D., Yeo, M., Reis-Cunha, J. L., Matos, G. M., et al. (2021). Repeat-driven generation of antigenic diversity in a major human pathogen, *Trypanosoma cruzi*. Front. Cell. Infect. Microbiol. 11, 614665. doi: 10.3389/fcimb.2021.614665

The, M., MacCoss, M. J., Noble, W. S., and Kall, L. (2016). Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. J. Am. Soc. Mass Spectrom. 27 (11), 1719–1727. doi: 10.1007/s13361-016-1460-7

The UniProt Consortium (2021). UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res. 49 (D1), D480–D489. doi: 10.1093/nar/gkaa1100

Törönen, P., and Holm, L. (2022). Pannzer — a practical tool for protein function prediction. Protein Sci. 31 (1), 118–128. doi: 10.1002/pro.4193

Vaudel, M., Verheggen, K., Csordas, A., Raeder, H., Berven, F. S., Martens, L., et al. (2016). Exploring the potential of public proteomics data. PROTEOMICS 16 (2), 214–225. doi: 10.1002/pmic.201500295

Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. 32 (3), 223–226. doi: 10.1038/nbt.2839

Vlasova, A., Hermoso Pulido, T., Camara, F., Ponomarenko, J., and Guigo, R. (2021). FA-Nf: A functional annotation pipeline for proteins from non-model organisms implemented in nextflow. Genes 12 (10), 1645. doi:10.3390/genes12101645

Wang, B., Kumar, V., Olson, A., and Ware, D. (2019). Reviving the transcriptome studies: An insight into the emergence of single-molecule transcriptome sequencing. Front. Genet. 10, 384. doi: 10.3389/fgene.2019.00384

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13 (5), 329–342. doi: 10.1038/nrg3174

5.9 SUPPLEMENTARY MATERIAL

Supplementary Table 1. Directories and output files generated by the AnnotaPipeline in each of the respective analysis performed during the annotation process.

AnnotaPipeline Directory		
File	**Format**	**Description**
All_Annotated_Products	TXT	Contains all unique sequence identifiers and their respective annotations
AnnotaPipeline_<basename>_proteins	FASTA	Contains all protein sequences and their annotations
AnnotaPipeline_<basename>_transcripts	FASTA	Contains all transcript sequences and their annotations
AnnotaPipeline_<basename>_Summary	TSV	Summarizes hits for each protein in similarity, functional, transcriptomics (if used) and proteomics analysis (if used)
<basename>_Annotated_GFF	GFF3	Contains all sequences and their annotations in GFF3 format
1_GenePrediction		
File	**Format**	**Description**
AUGUSTUS_<basename>	GFF	Predicted genes in gff format
AUGUSTUS_<basename>	AA	Predicted amino acid sequences in fasta format
AUGUSTUS_<basename>	CDSEXONS	Predicted exon sequences in fasta format
AUGUSTUS_<basename>	CODINGSEQ	Predicted coding sequences in fasta format
AUGUSTUS_<basename>	MRNA	Predicted mRNA sequences in fasta format
Clear_AUGUSTUS_<basename>	AA	Amino acid sequences in fasta format after minimum size filter (minsize-seq in config.yaml file)

2_SimilarityAnalysis		
File	**Format**	**Description**
SimilarityAnalysis	LOG	Log from blastp_parser.py script (include command lines for debugging)
<basename>_BLASTp_AAvsSwissProt	OUTFMT6	Raw output file from blast similarity analysis with SwissProtDB
<basename>_SwissProt_annotations	TXT	List of proteins containing all valid annotations found during similarity analysis with SwissProtDB. This output shows all hits that passes cutoffs parameters (coverage, identity and positivity) and semantic analysis (keywords scape values)
<basename>_BLASTp_AA_SwissProted	FASTA	Amino acid sequences of proteins that were not annotated with swissprot (input for SpecifiedDB similarity analysis)
<basename>_BLASTp_AAvsSpecifiedDB	OUTFMT6	Raw output file from blast similarity analysis with SpecifiedDB
<basename>_SpecifiedDB_annotations	TXT	List of proteins containing all valid annotations found during similarity analysis with SpecifiedDB. This output shows all hits that passes cutoffs parameters (coverage, identity and positivity) and semantic analysis (keywords scape values)
<basename>_annotated_products	TXT	List of hypothetical proteins (considering only similarity)
<basename>_hypothetical_products	TXT	List of hypothetical proteins (considering only similarity)
<basename>_no_hit_products	TXT	List of proteins with no hit in SwissprotDB and SpecifiedDB
3_FunctionalAnnotation		
File	**Format**	**Description**
FunctionalAnnotation	LOG	Log from Funcannotation.py parser (include command lines for debugging)
Annotated_Products	FASTA	Fasta file containing proteins annotated

Hypothetical_Products	FASTA	Fasta file containing hypothetical proteins
<basename>_interproscan_annotated_ output	GFF3	Raw output from interproscan when running with annotated proteins
<basename>_interproscan_hypothetical_ output	GFF3	Raw output from interproscan when running with hypothetical proteins
<basename>_hmmscan_output	TXT	Raw output from hmmscan running only with hypothetical proteins
<basename>_rpsblast_output	OUTFMT6	Raw output from rpsblast running only with hypothetical proteins
InterProScan_Out_<basename>	TSV	Parsed file with Interproscan annotations (for hypothetical and annotated proteins)
Hmmscan_Out_<basename>	TSV	Parsed file with hmmscan annotations
RPSblast_Out_<basename>	TSV	Parsed file with RPSblast annotations
<basename>_Grouped_Hypothetical_ Information	TSV	File that groups information of each hypothetical protein (InterproScan + Hmmscan + RPSblast)
4_TranscriptQuantification		
File	**Format**	**Description**
<basename>_kallisto_index	IDX	Index file to run Kallisto
<basename>_Transcript_Quantification	TSV	Parsed file with tpm found in each protein
<basename>_kallisto_output	FOLDER*	Directory containing abundance quantification outputs from kallisto
5_PeptideIdentification		
File	**Format**	**Description**
<basename>_Total_Proteomics_ Quantification	TSV	Count of unique/total peptide and spectrum for each protein (using files after parsering)

COMET_Output	FOLDER*	Original output files from comet
PERCOLATOR_Raw	FOLDER*	Raw output files from percolator, divided by sample
PERCOLATOR_Parsed	FOLDER*	Output files from percolator, divided by sample, after filtering by q-value cutoff

* Directory containing original output file from software

Supplementary Table 2. Comparative analysis of the TriTrypDB and AnnotaPipeline annotations of *Trypanosoma cruzi* and *Trypanosoma rangeli* genomes.

Parameter	*Trypanosoma cruzi* Sylvio X10/1		*Trypanosoma rangeli* SC58	
	TriTrypDB	**AnnotaPipeline**	**TriTrypDB**	**AnnotaPipeline**
Predicted proteins	20,619	9,127	7,475	5,649
Annotated proteins	5,075 (24.61%)	7,759 (85.01%)	2,400 (32.11%)	4,234 (74.95%)
Annotated by SwissProt	-	2,569 (33.11% of annotated)	-	1,252 (29.57% of annotated)
Annotated by SpecificDB	-	5,190 (66.89% of annotated)	-	2,982 (70.43% of annotated)
Hypothetical Proteins	15,544	1,348	5,075	1,411
No hit proteins	-	20	-	4
Total hypothetical proteins	15,544 (75.39%)	1368 (14.99%)	5,075 (67.89%)	1,415 (25.05%)
Proteins with at least 1 IPR term	-	5,757 (63.08%)	-	4,052 (71.73%)
Proteins with at least 1 GO term	-	3,809 (41.73%)	-	2,990 (52.93%)

Reference genome TriTrypDB (version 57) accession number: *Trypanosoma cruzi* (strain Sylvio X10/1) = DS_107bdce9bb; *Trypanosoma rangeli* (strain SC58) = DS_9d0531db8e.

6 CAPÍTULO II: GENOMA DA CEPA SC58 DE *TRYPANOSOMA RANGELI*

6.1 METODOLOGIA

6.1.1 Parasitos e sequenciamento de DNA

Anteriormente à realização deste trabalho, a cepa SC58 de *T. rangeli* foi isolada pelo grupo de pesquisa do Laboratório de Protozoologia da UFSC, a partir de um roedor (*Phyllomys dasythrix*) na Ilha de Santa Catarina (STEINDEL, 1993; GRISARD et al., 1999). Foi hipotetizado que o triatomíneo *R. domesticus* é o vetor natural para o *T. rangeli* nesta região, pois é a única espécie de triatomíneo descrita para a região de Florianópolis (GRISARD et al., 1999). Considerando-se diferentes marcadores moleculares utilizados em filogenia de tripanosomatídeos, a cepa SC58 de *T. rangeli* pertence ao clado TrD (GRISARD et al., 1999; ESPINOSA-ÁLVAREZ et al., 2018), sendo pertencente à linhagem KP1(-) (STOCO et al., 2014), quando se consideram aspectos genômicos baseados no arranjo de sequências de kDNA minicírculo.

Brevemente, para geração de dados genômicos de alta cobertura do tipo *short reads* e *long reads*, o DNA total de *T. rangeli* foi extraído a partir de formas epimastigotas de cultura, crescidas em meio LIT, e utilizado para construção de bibliotecas do tipo *paired-end* e *single molecule real time sequencing* (SMRT), para sequenciamento nas plataformas Illumina (modelo HiSeq 2500) e PacBio (modelo RSII), respectivamente, de acordo com os protocolos padrões estabelecidos por cada fabricante.

Toda a etapa de sequenciamento do material biológico foi realizada no *Science for Life Laboratory* do Karolinska Institutet, em Estocolmo, na Suécia, sob a supervisão do Prof. Dr. Björn Andersson. Todos os dados genômicos de *T. rangeli* previamente gerados estão acessíveis nos bancos de dados do Laboratório de Protozoologia e no Karolinska Institutet. O genoma anotado da cepa SC58 de *T. rangeli* gerado no presente estudo será disponibilizado publicamente no GenBank e no TriTrypDB depois de sua publicação.

Para as análises comparativas, foram utilizados os genomas das espécies *T. brucei* (cepa TREU927), *T. cruzi* (cepas CL Brener Esmeraldo-like e Sylvio X10/1) e *T. rangeli* (cepa SC58) disponíveis no banco de dados do TriTrypDB (versão 57), o genoma da espécie *T. rangeli* (cepa AM80) obtido do banco de dados do GenBank

(código de acesso: GCF_003719475.1), bem como o genoma de *T. conorhini* (cepa 025E), gerados pelo nosso grupo de pesquisa em colaboração com o Laboratório de Biologia Molecular de *Trypanosoma cruzi*, do Prof. Dr. José Franco da Silveira (Universidade Federal de São Paulo – UNIFESP), que estão em fase de redação de artigos e submissão dos dados.

6.1.2 Controle de qualidade e cobertura genômica

Os dados brutos obtidos do sequenciamento foram submetidos à etapa de controle de qualidade, visando a remoção de bases nucleotídicas com qualidade *phred* abaixo de um limiar mínimo estipulado. O limiar de qualidade mínima das bases nucleotídicas geradas foi determinado a partir de valores observados através do programa FastQC v0.11.8 (ANDREWS, 2010), sendo que todas as bases com *Q-score* abaixo deste valor foram removidas utilizando os programas Trimmomatic v0.39 (BOLGER; LOHSE; USADEL, 2014) para os dados de Illumina, e Filtlong v0.2.0 (disponível em: https://github.com/rrwick/Filtlong) para os dados da plataforma PacBio. O programa Trimmomatic também faz a remoção de sequências adaptadoras inerentes da técnica de sequenciamento em plataformas Illumina. Desta forma, as bases nucleotídicas com *Q-score* abaixo do limiar são substituídas por bases *N* e *reads* que apresentam inúmeras dessas substituições são removidos do conjunto de dados.

O método de remoção das bases nucleotídicas de baixa qualidade geradas no sequenciamento pela plataforma Illumina foi realizado através do método de janelas corridas (do inglês, *sliding window*), para garantir a boa qualidade dos *reads* que foram utilizados na montagem. Os parâmetros de remoção de bases nucleotídicas gerais (SLIDINGWINDOW:4:30), iniciais (LEADING:30) e finais (TRAILING:30) foram alterados para requerer um *Q-score* mínimo de 30, assim como foi adotado um tamanho mínimo para os *reads* restantes (MINLEN:75). Para a remoção das bases nucleotídicas oriundas da plataforma PacBio, foi estipulado apenas um tamanho mínimo dos *reads* restantes (--min_length 1000) e garantindo a disponibilidade de pelo menos 90% dos reads (--keep_percent 90).

A cobertura dos *reads* de Illumina no genoma depende do tamanho estimado do genoma ou de um genoma de referência, que neste caso é a cepa SC58 de *T.*

rangeli (STOCO et al., 2014). A equação que estima a cobertura teórica do genoma é tal qual:

$$cobertura = \frac{total\ de\ reads * tamanho\ médio\ dos\ reads}{tamanho\ do\ genoma\ de\ referência}$$

É importante destacar que, considerando *short reads* gerados por bibliotecas de sequenciamento do tipo *paired-end*, é necessário multiplicar o fator *total de reads* por dois, para considerar os *reads* gerados a partir de sequências senso e antissenso.

6.1.3 Montagem genômica e validação experimental da montagem

A montagem da nova versão do genoma de *T. rangeli* foi realizada pelo método *de novo*, sendo que todos os parâmetros descritos nesta seção foram avaliados manualmente após testes empíricos, que visavam a obtenção da melhor montagem dada a respectiva etapa metodológica. Primeiro, foi utilizado o programa Canu v2.2 (KOREN, 2017), para montagens com dados de *long reads* gerados pela plataforma PacBio, cujo único parâmetro fornecido durante sua execução foi uma estimativa do tamanho do genoma haploide de 13 Mb (genomeSize=13000000).

Em seguida, foram incorporados os *shorts reads* do tipo *paired-end* gerados pela plataforma Illumina para a realização da etapa de junção de *contigs* para formação de *scaffolds*. Para tal, primeiro foi utilizado o algoritmo CollectInsertSizeMetrics do programa Picard Tools v2.23.3 (disponível em: https://github.com/broadinstitute/picard) para estimar o tamanho de inserção dos *reads paired-end*. Em seguida, foi utilizado o programa SSPACE v3.0 (BOETZER et al., 2010) para formação de *scaffolds*, o qual foi executado em sete iterações variando o tamanho de cobertura necessário para correção dos *scaffolds* (-k 100, 50, 30, 20, 10, 5 e 5) e aceitando um desvio padrão relativo da estimativa do cálculo de inserção de 0,2%. Esta mesma estimativa de tamanho de inserção dos *reads paired-end* também foi utilizada pelo programa GapFiller v1-10 (BOETZER; PIROVANO, 2012) para preenchimento de espaços entre os *scaffolds* formados. A execução do programa GapFiller foi realizada em 10 iterações (-i 10), considerando-se parâmetros como: (i) interpolações mínimas de 20 nucleotídeos para junção de *scaffolds* adjacentes (-n 20); (ii) remoção de 10 nucleotídeos em cada extremidade de *scaffolds*

que apresentam baixa cobertura (-t 10); e (iii) diferença máxima de 50 nucleotídeos entre espaços fechados e nucleotídeos adicionados (-d 50).

Os *scaffolds* obtidos foram analisados pelo programa Pilon v1.23 (WALKER et al., 2014), para que fosse realizada a substituição de nucleotídeos erroneamente classificados ou não identificados, utilizando apenas informação de sequência de *reads* com *Q-score* acima de 34 (--defaultqual 34) e considerando a natureza diplóide do organismo (--diploid).

Para que fossem respeitados os aspectos biológicos do genoma de *T. rangeli*, todas as etapas da montagem do genoma foram avaliadas pelo programa QUAST v5.0.2 (GUREVICH et al., 2013), para obtenção das principais métricas de montagem. O programa BUSCO v5.5.0 (MANNI et al., 2021) foi utilizado para estimar a completude da montagem genômica, por meio de uma avaliação comparativa universal de genes ortólogos de cópia única (do inglês, *Benchmarking Universal Single-Copy Orthologues*), baseada na composição genética esperada de organismos do Filo Euglenozoa (--lineage_dataset euglenozoa_odb10).

Como validação experimental da montagem, foi realizado um mapeamento *in silico* do genoma do *T. rangeli* utilizando-se sequencias de sondas cromossômicas direcionadas a genes de cópia única do genoma de *T. cruzi* (cepa CL Brener), as quais foram utilizadas em ensaios de hibridização genômica comparativa baseada em arranjo de DNA (do inglês, *array Comparative Genomic Hybridization* – aCGH) com *T. cruzi* e o genoma de referência de *T. rangeli* (STOCO et al., 2014). Estas sondas foram desenvolvidas pelo grupo de pesquisa do Prof. Dr. José Franco da Silveira Filho da UNIFESP, sendo utilizadas pela Drª. Rafaela Andrade do Carmo nos experimentos que decorreram durante seu doutoramento (Apêndice A). O mapeamento foi feito por análise de similaridade entre as sequências das sondas e os *scaffolds* obtidos na montagem, utilizando os algoritmos BLASTn e tBLASTx, do programa BLAST+ v2.9.0 (CAMACHO et al., 2009), admitindo um *E-value* de $1e^{-10}$ como valor mínimo de corte e considerando apenas os resultados que apresentaram cobertura acima de 60% entre o gene marcador e a região do *scaffold* correspondente.

Outra etapa de validação da montagem foi realizada *in silico* através de duas análises: (i) um mapeamento entre a montagem realizada neste trabalho e o genoma de referência de *T. rangeli* (STOCO et al., 2014), utilizando o algoritmo dnadiff do programa MUMmer4 v1.4 (MARÇAIS et al., 2018); e (ii) uma busca por genes de cópia única utilizados em análises laboratoriais pela técnica de tipagem de sequência

multilocus (do inglês, *Multilocus Sequence Typing* – MLST), que inicialmente foram descritos para os dados de *T. cruzi* (cepas CL Brener e Sylvio X10/1), sendo confirmada a correspondência também para os dados do genoma de referência de *T. rangeli* (STOCO et al., 2014). A partir das sequências que obtiveram similaridade no genoma de referência de *T. rangeli*, os algoritmos BLASTn e BLASTp foram utilizados para realizar uma busca por similaridade na montagem do genoma de *T. rangeli* realizada neste trabalho.

Por fim, a visualização da cobertura observada dos *reads*, para cada um dos *scaffolds* obtidos na montagem, foi feita pelos pacotes ggplot2 e tidyverse do programa R v4.3.1 (R CORE TEAM, 2023). Os arquivos de entrada utilizados pelo programa R foram obtidos através de mapeamentos feitos pelos programas Bowtie2 v2.5.1 (LANGMEAD; SALZBERG, 2012), para dados de *short reads*, e Minimap2 v2.26 (LI, 2018), para os dados de *long reads*, sendo preparados pelo programa Samtools v1.18 (DANECEK et al., 2021), todos executados sem nenhum parâmetro opcional.

A partir deste ponto em diante, todas as menções ao "genoma de referência" no texto indicam os dados de Stoco e colaboradores (2014), cuja versão do genoma foi posteriormente analisada pelo TriTrypDB e incorporado naquela base de dados. A nova montagem realizada neste estudo será referida como "versão 2", uma vez que ambos os estudos são baseados na cepa SC58 de *T. rangeli*.

6.1.4 Predição de RNA

A predição de RNA ribossomais (rRNA) e RNA transportadores (tRNA) na versão 2 foi realizada pelo algoritmo cmscan do programa Infernal v1.1.2 (NAWROCKI; EDDY, 2013), que se baseia na busca por similaridade com sequências de RNA não-codificantes depositados no banco de dados Rfam v14.2 (KALVARI et al., 2018). Os parâmetros utilizados pelo algoritmo cmscan foram: (i) utilização de modelos de covariância juntamente com modelos ocultos de Markov (--nohmmonly); (ii) eliminação de resultados truncados em terminações de sequências (--notrunc); (iii) valores de *bitscore* pré-determinados pelo banco de dados curado, ao se considerar uma possível homologia entre sequências (--cut_ga); e (iv) utilização de dados do Rfam como informações de agrupamentos (--clanin).

Os RNA preditos foram validados *in silico* com base em percentuais de cobertura e identidade obtidos através do algoritmo BLASTn do programa BLAST+ v2.9.0. Sequências preditas de rRNA e tRNA que apresentaram valores de cobertura e identidade acima de 90% quando comparadas com sequências de tripanosomatídeos obtidas do banco de dados RNAcentral v20 (RNACENTRAL CONSORTIUM et al., 2018) foram consideradas validadas. No total, foram utilizadas 8.527 sequências do RNAcentral, obtidas no dia 02 de junho de 2022, para validação de RNA preditos na versão 2 do genoma de *T. rangeli*.

6.1.5 Predição e anotação gênica

A etapa de predição e anotação gênica da versão 2 do genoma de *T. rangeli* foi realizada pelo programa AnnotaPipeline (MAIA et al., 2022), utilizando-se *E-value* de $1e^{-5}$ para as análises de similaridade e *Q-value* de 0,05 para identificação de peptídeos.

Primeiro, o programa AUGUSTUS v3.3.3 (STANKE; WAACK, 2003) foi utilizado para realizar predição de genes por meio de um modelo preditivo fornecido pela plataforma online WebAugustus (HOFF; STANKE, 2013), a partir de um conjunto de treino composto por 2.942 etiquetas de sequências expressas (do inglês, *expressed sequence tags* – EST) e de 7.802 sequências aminoacídicas preditas do genoma de referência de *T. rangeli*. Em seguida, a etapa de anotação foi realizada através de análise de similaridade utilizando-se o algoritmo BLASTp, do programa BLAST+ v2.9.0, entre as proteínas preditas e sequências proteicas depositada nos bancos de dados SwissProt (THE UNIPROT CONSORTIUM, 2020), obtido em janeiro de 2022, e o TriTrypDB v57. Neste trabalho, foram consideradas como anotadas as CDS preditas na versão 2 que possuíam uma anotação correspondente a elas nos bancos de dados utilizados. As CDS preditas foram anotadas como hipotéticas quando as anotações atribuídas a elas por similaridade com os bancos de dados utilizados continham qualquer uma das palavras-chave (*fragment*, *hypothetical*, *partial*, *uncharacterized*, *unknown* e *unspecified*) em sua descrição. Ao mesmo tempo que, caso os parâmetros mínimos de 30% de cobertura, 40% de identidade e 60% de positividade não fossem contemplados para a transferência de anotação, as CDS preditas foram automaticamente reanotadas como hipotéticas, desconsiderando a anotação do banco de dados. Por fim, foram automaticamente anotadas como

hipotéticas quaisquer CDS preditas que não encontraram correspondência de anotação nos bancos de dados estudados, agregando-se ao número total de CDS preditas e anotadas como hipotéticas da versão 2.

Em seguida, a etapa de anotação funcional foi realizada utilizando-se os programas InterProScan v5.61-93.0 (JONES et al., 2014), o algoritmo hmmscan do programa HMMER v3.1b2 (FINN; CLEMENTS; EDDY, 2011) e o algoritmo RPS-BLAST, que faz parte do programa BLAST+ v2.9.0.

Com a disponibilidade de dados transcriptômicos (RNA-Seq) e proteômicos (espectros de massas) de *T. rangeli*, gerados previamente pelo grupo de pesquisa do Laboratório de Protozoologia da UFSC (WAGNER, 2006; GRISARD et al., 2010; WAGNER et al., 2013; LÜCKEMEYER, 2014), estes dados experimentais foram utilizados para verificar a evidência de transcrição e tradução das CDS preditas, utilizando o AnnotaPipeline. Desta forma, a etapa de evidência de transcrição foi realizada pelo programa Kallisto v0.48.0 (BRAY et al., 2016), enquanto a etapa de evidência de tradução foi feita pelo programa Comet v2021.01 (ENG; JAHAN; HOOPMANN, 2012) em combinação com o programa Percolator v3.05.0 (THE et al., 2016).

6.1.6 Caracterização *in silico* de CDS preditas e anotadas

Do conjunto total de CDS preditas e anotadas na versão 2 do genoma de *T. rangeli*, foi escolhido um subconjunto contendo todas as proteínas anotadas como hipotéticas, as quais foram analisadas *in silico* quanto a presença de elementos que indicassem um perfil associado a proteínas de superfície. Esta caracterização foi realizada pelo programa FastProtein v1.0 (disponível em: https://github.com/bioinformatics-ufsc/FastProtein), que provêm uma série de inferências a respeito de cada CDS, como massa molecular, ponto isoelétrico, localização celular, presença de domínios transmembranas, domínios de N-glicosilações e outros. A execução do programa foi realizada de forma online, utilizando parâmetros padrão para as análises computacionais realizadas.

Desta forma, foram consideradas como proteínas de superfície as CDS que apresentavam: presença de peptídeo sinal; ancoramento na membrana celular por âncora de GPI; e evidência de localização extracelular.

6.1.7 Análise de homologia

Todas as proteínas obtidas dos processos de predição e anotação gênica da versão 2 do genoma de *T. rangeli* foram submetidas a análise de homologia pelo programa OrthoFinder v2.3.11 (EMMS; KELLY, 2019), com alteração de *E-value* do algoritmo BLASTp para $1e^{-5}$. Conforme descrito anteriormente, para fins comparativos foram utilizadas as proteínas anotadas de *T. brucei* TREU927 e *T. cruzi* CL Brener Esmeraldo-like obtidas a partir do banco de dados TriTrypDB v57, assim como as proteínas anotadas de *T. conorhini* 025E geradas pelo nosso grupo de pesquisa. Foram consideradas como sequências ortólogas aquelas que pertencem a um mesmo agrupamento ortólogo que continha sequências de diferentes espécies, assim como foram consideradas como sequências parálogas aquelas pertencentes a um agrupamento que continha duas ou mais sequências apenas de uma mesma espécie. Ainda, os *singletons* foram considerados como grupos que continham apenas uma proteína de apenas uma espécie.

As principais métricas e estatísticas gerais da análise de homologia, assim como a identificação das CDS pertencentes aos agrupamentos (ortólogos ou parálogos), foram obtidas dos relatórios gerados pelo próprio programa OrthoFinder. Por fim, a representação dos resultados de homologia gerados neste trabalho foi construída através da plataforma *Bioinformatics & Evolutionary Genomics* (disponível em: https://bioinformatics.psb.ugent.be/webtools/Venn/).

6.1.8 Investigação de elementos repetitivos

A análise de elementos genômicos repetitivos na versão 2 do genoma de *T. rangeli* e em genomas tripanosomatídeos foi realizada em duas partes, considerando os conceitos expostos previamente no texto: (i) elementos repetitivos estruturais; e (ii) elementos repetitivos funcionais. Para fins comparativos, quando não tratada em números absolutos de pares de bases, foi realizada uma normalização do conteúdo genômico de cada uma das espécies analisadas, de tal forma que o respectivo tamanho do genoma foi considerado como 100% e as demais comparações foram feitas em relação a este total.

A análise de elementos repetitivos estruturais foi realizada pelo programa RepeatMasker v4.1.5 (SMIT; HUBLEY; GREEN, 2015), utilizando uma metodologia de busca mais lenta, porém mais sensível (-s), desconsiderando elementos de

inserção bacterianos (-no_is) e uma biblioteca personalizada (-lib), com sequências específicas de tripanosomatídeos. A biblioteca personalizada foi montada a partir dos dados disponíveis no msRepDB (LIAO et al., 2022), buscando-se pelo termo “Trypanosoma” e concatenando os conjuntos de dados obtidos desta forma em apenas um arquivo.

A fim de investigar a presença de regiões teloméricas na versão 2, foram utilizados os algoritmos telo do programa Seqtk v1.4 (disponível em: https://github.com/lh3/seqtk) e o programa Tidk v0.2.31 (disponível em: https://github.com/tolkit/telomeric-identifier). Para ambos os programas a repetição telomérica canônica representada pelo monômero (TTAGGG)n foi considerada padrão de busca nas extremidades dos *scaffolds*, conforme previamente descrito para organismos tripanosomatídeos (HORN; SPENCE; INGRAM, 2000; RAMIREZ, 2020a), mais precisamente nos primeiros 10 mil pares de bases e nos últimos 10 mil pares de base de cada *scaffold* da versão 2. Os resultados obtidos foram avaliados manualmente para que fossem desconsiderados quaisquer resultados que indicassem possíveis regiões teloméricas na porção central dos *scaffolds*.

A análise de elementos repetitivos funcionais foi baseada no número de nucleotídeos totais presentes nas seguintes famílias multigênicas: amastinas, DGF-1, GP63, KMP-11, MASP, mucinas, RHS, sialidases, tuzinas e VSG, que foram identificadas durante o processo de anotação automática das CDS preditas. Foi utilizado o algoritmo grep do programa SeqKit v2.5.1 (SHEN et al., 2016) para a realização de uma busca por palavras-chave na anotação do conjunto de CDS preditas e anotadas, tais como: *amastin*, *dgf*, *dispersed*, *dispersed gene family*, *gp63*, *leishmanolysin*, *kmp*, *kinetoplastid*, *kinetoplastid membrane protein*, *masp*, *mucin-associated*, *mucin-associated surface protein*, *muc*, *mucin*, *rhs*, *retrotransposon*, *retrotransposon hot spot*, *sialidase*, *vsg*, *surface glycoprotein*, *variant surface glycoprotein* e *tuzin*. Pseudogenes relacionados aos termos de busca foram desconsiderados da análise. CDS cujas palavras-chave são redundantes quanto a sua atividade biológica, por exemplo *masp* e *mucin-associated surface protein*, foram consideradas como pertencentes a uma mesma família multigênica e sua redundância removida. A anotação de cada CDS foi avaliada manualmente para remoção de falsos positivos relacionados aos termos encontrados e o número de nucleotídeos totais foi calculado pelo algoritmo stats do programa SeqKit v2.5.1.

6.1.9 Análise de sintenia entre genomas de tripanosomatídeos

Foi realizada uma análise de sintenia entre os *scaffolds* da versão 2 do genoma de *T. rangeli* e os cromossomos da cepa Sylvio X10/1 de *T. cruzi*, utilizando-se o programa Synchro (DRILLON; CARBONE; FISCHER, 2014), que foi executado com parâmetros padrões. O preparo da biblioteca necessária para a análise foi feito utilizando o algoritmo convertfasta do programa Synchro, cujos arquivos de entrada foram arquivos de CDS preditas e anotadas das respectivas espécies, com cabeçalhos das sequências modificados para conter apenas: o identificar da CDS; o cromossomo que contém aquela CDS; a posição de início da CDS; a posição de término da CDS; e o sentido do cromossomo.

6.2 RESULTADOS E DISCUSSÃO

6.2.1 Montagem e validação experimental da nova versão do genoma

De maneira geral, a versão 2 do genoma da cepa SC58 de *T. rangeli* apresentada neste trabalho utilizou dados de sequenciamento gerados pelas plataformas Illumina e PacBio e permitiu refinar o tamanho do genoma, melhorar sua contiguidade e reduzir a quantidade de bases não identificadas, quando comparado com o genoma de referência depositado em bancos de dados públicos, o qual foi realizado exclusivamente com dados de sequenciamento da plataforma Roche-454 (STOCO et al., 2014). O panorama geral das métricas de montagem avaliadas está disposto na Tabela 1.

Tabela 1. Comparação das principais métricas de montagem da versão 2 do genoma de *Trypanosoma rangeli* SC58 e do genoma de referência desta cepa.

	Versão 2	**Referência**
Tamanho do genoma (Mpb)	30,27	13,97
Número de *scaffolds*	465	7.321
Conteúdo G+C genômico (%)	53,87	53,27
***Scaffolds* com mais de 50 mil pb**	71	0
Tamanho do maior *scaffold* (Mpb)	1,37	0,25
N50	380.491	2.208
L50	22	2.085
Bases *N* a cada 100 mil pb	0,08	21,37

Legenda: Mpb = milhões de pares de bases. pb = pares de bases.

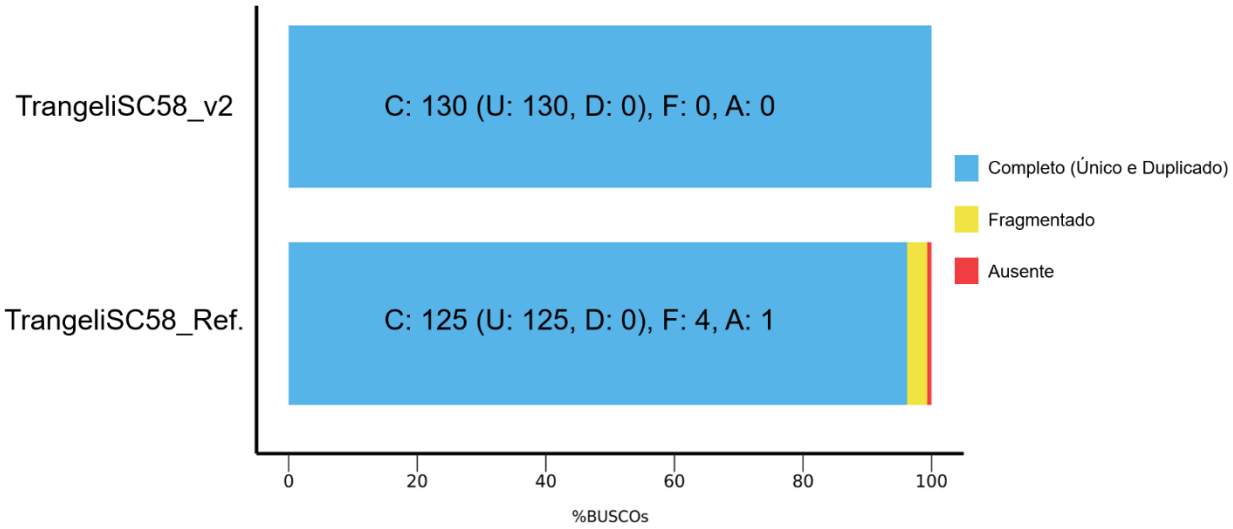
O tamanho observado na versão 2 – de aproximadamente 30 milhões de pares de bases – está de acordo com o esperado de um genoma haplóide para um organismo tripanosomatídeo (EL-SAYED et al., 2005b; STOCO et al., 2014; REIS-CUNHA; BARTHOLOMEU, 2019). Destaca-se também que foi possível reduzir em mais de 90% o número de *scaffolds* no genoma desta espécie na versão 2. Mesmo quando comparado com o genoma de referência de *T. rangeli*, foi observado que a versão 2 apresenta correspondência com 99,95% da informação contida no genoma

de referência, apontando sua devida cobertura, ainda que essas montagens tenham sido geradas por metodologias distintas de sequenciamento.

A baixa quantidade de bases não identificadas na versão 2 se deve, entre outros fatores, ao uso de dados de sequenciamento do tipo *long reads*, uma vez que a utilização de *short reads* para montagem de regiões repetitivas no genoma acarreta problemas, como ambiguidade na chamada de bases, vieses de cobertura e colapso de regiões (LOGSDON; VOLLGER; EICHLER, 2020; WANG et al., 2021). Levando em consideração que os genomas de tripanosomatídeos possuem uma abundância de sequências repetitivas e genes dispostos de maneira cistrônica, é justamente nessas regiões de grande repetição, ou de elevado conteúdo G+C, que a utilização de *long reads* pode melhorar significativamente as montagens genômicas (ENGLISH et al., 2012; BERNÁ et al., 2018). Ainda, a baixa quantidade de bases não identificadas na versão 2 também garante que as etapas seguintes de predição e anotação gênica sejam mais confiáveis.

Como complemento às métricas de montagem, a estimativa da completude da versão 2 do genoma de *T. rangeli* sugere para uma versão completa dos elementos gênicos desta espécie. Do conjunto de dados de genes ortólogos de cópia única analisados, todos os 130 marcadores foram encontrados na versão 2; sem genes duplicados, sem sequências fragmentadas, tampouco marcadores ausentes (Figura 4).

Figura 4. Estimativa e comparação da completude entre as montagens genômicas da versão 2 e do genoma de referência de *Trypanosoma rangeli* SC58.



Legenda: Resultado obtido considerando-se a presença de um conjunto predefinido e esperado de genes marcadores de cópia única (n: 130), presentes em organismos do Filo Euglenozoa. “Ref.” indica o genoma de referência de *T. rangeli* SC58 (STOCO et al., 2014), obtido do banco de dados do TriTrypDB (v57). “v2” indica a versão 2 do genoma de *T. rangeli* SC58.

Com relação à validação experimental da versão 2, foi possível mapear nove das 13 sondas de genes de cópia única de *T. cruzi* em nove *scaffolds* únicos de *T. rangeli*, duas sondas cada uma em dois *scaffolds* e duas outras sondas cada uma em múltiplos *scaffolds*, resultado que indica que o genoma montado neste trabalho é de alta qualidade. Estes casos em que as sondas foram mapeadas em múltiplos *scaffolds* indicam que ainda há possibilidade de aprimorar a montagem do genoma desta espécie, ou que o genoma de *T. rangeli* apresente múltiplas cópias destes genes, e não cópias únicas como descrito no genoma do *T. cruzi*.

Os resultados do mapeamento das sondas na versão 2, assim como um mapeamento comparativo realizado com outras espécies de tripanosomatídeos, está apresentado na Tabela 2.

Tabela 2. Tabela comparativa do mapeamento de sondas cromossômicas de genes de cópia única de *Trypanosoma cruzi* em diferentes espécies e na versão 2 do genoma de *T. rangeli* SC58.

Gene Marcador *	**Anotação Gênica**	**Mapeamento *T. cruzi* CL Brener**	**Mapeamento *T. conorhini* 025E**	**Mapeamento *T. rangeli* Choachí**	**Mapeamento *T. rangeli* SC58**	
					Referência	**Versão 2**
TcCLB.503793.20	phosphatidylinositol (3,5) kinase, putative (fragment)	banda 16	-	-	-	scaffold16
TcCLB.504109.170	hypothetical protein, conserved	banda 16	banda 10	banda 8	banda 14	scaffold16
TcCLB.506127.90	Hereditary spastic paraplegia protein strumpellin, putative	banda 16	banda 18	banda 15	banda 14	scaffold2
TcCLB.507009.90	iron-sulfur cluster assembly protein, putative	banda 16	banda 17	banda 15	banda 14	scaffold10
TcCLB.507609.40	delta-4 fatty acid desaturase, putative	banda 10	banda 17	banda 15	banda 14	scaffold18
TcCLB.507611.380	hypothetical protein	banda 10	banda 18	banda 15	banda 14	scaffold7
TcCLB.507641.120	ATP-dependent DEAD/H RNA helicase, putative	banda 10	bandas 18 e 17	banda 15	banda 14	scaffold6
TcCLB.507681.160	40S ribosomal protein S24E, putative	banda 10	banda 16	banda 6	bandas 6 e 7	scaffold30
TcCLB.507713.30	heat shock protein 85, putative	banda 10	banda 18	banda 15	banda 14	*Múltiplos*
TcCLB.508465.120	syntaxin, putative	banda 16	banda 17	banda 15	banda 14	scaffold10
TcCLB.509715.120	phosphatidylinositol-4-phosphate 5-kinase, putative	banda 10	-	bandas 6 e 15	banda 6	scaffold30 e scaffold186
TcCLB.510129.20	Surface membrane protein	banda 10	banda 17	banda 15	-	scaffold14 e scaffold5
TcCLB.511041.40	hexose transporter, putative	banda 10	banda 18	banda 15	banda 14	*Múltiplos*

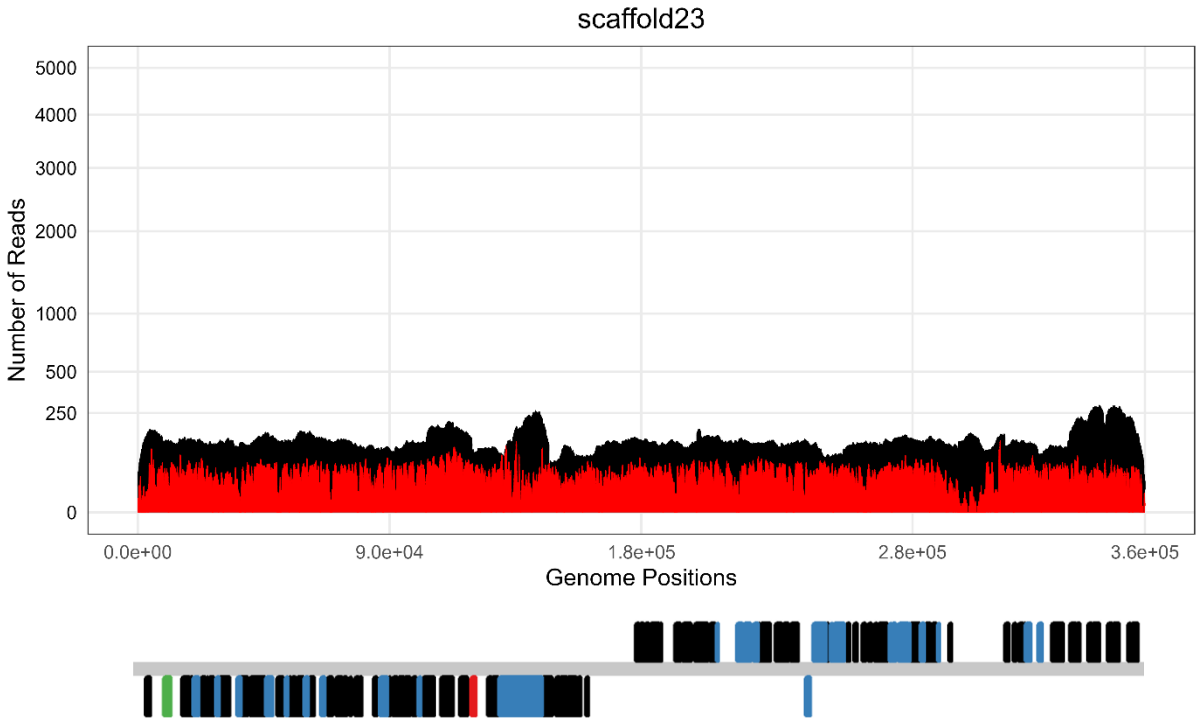
Legenda: * Código de acesso do respectivo gene marcador no banco de dados do TriTrypDB. “Referência” sinaliza o genoma de referência de *T. rangeli* SC58 (STOCO et al., 2014). “Versão 2” indica a montagem do genoma da cepa SC58 de *T. rangeli* realizada neste trabalho. O termo “Múltiplos” se refere ao mapeamento do gene marcador em cinco ou mais *scaffolds*. Fonte: adaptado de DO CARMO, 2022.

Dados de sequenciamento de *T. rangeli* SC58 da plataforma Illumina geraram 9.264.332 *reads* limpos, a medida em que 361.938 *reads* limpos foram obtidos da plataforma PacBio. Isto corresponde a aproximadamente 4,8 GB e 6,9 GB de dados sequenciados, respectivamente.

De acordo com a equação que estima a cobertura teórica de um genoma, apresentada anteriormente (secção 6.1.2), a cobertura esperada em relação ao genoma de referência de *T. rangeli* era de 143,27x para os dados de *long reads* e de 71,13x para os dados de *short reads*. No entanto, os resultados obtidos apontam uma cobertura média de 19,88x para os dados de *short reads* e de 192,32x para os dados de *long reads*. A diferença entre a cobertura esperada e a cobertura real para os *short reads* é compreensível, porque é improvável que todos os *reads* sejam utilizados durante a montagem genômica, especialmente quando a montagem é realizada combinando dados de diferentes plataformas de sequenciamento. Por outro lado, a diferença entre a cobertura esperada e a cobertura observada para os dados de *long reads* é intrigante. Uma possibilidade é que a equação utilizada para estimar a cobertura teórica pode não ser adequada para dados do tipo *long reads*. Outra possibilidade é que a montagem do genoma de referência tenha limitações, como o colapso de regiões repetitivas, que podem diminuir o tamanho do genoma, que é um dos fatores considerados na equação.

Como exemplo ilustrativo, na Figura 5 estão dispostas as coberturas observadas para o *scaffold*23 da versão 2.

Figura 5. Figura da profundidade de cobertura observada de *reads* no *scaffold*23 da versão 2 do genoma de *Trypanosoma rangeli* SC58.

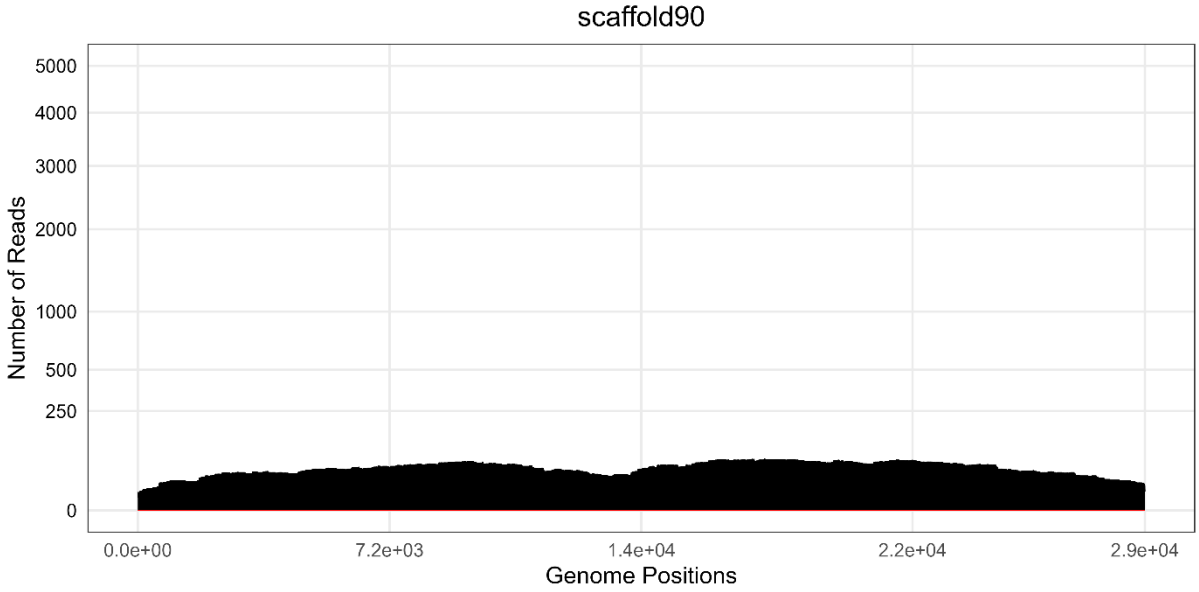


Legenda: Os *short reads* oriundos pela plataforma Illumina estão representados em vermelho. Os *long reads* obtidos da plataforma PacBio estão representados em preto. Eixo X representa as posições genômicas no *scaffold*. Eixo Y representa o número de *reads*, em ordem crescente de grandeza. Abaixo do gráfico, estão dispostas verticalmente as CDS (preto) ao longo da extensão do *scaffold* (cinza). Estão destacadas as CDS anotadas como hipotéticas (azul), GP63 (vermelho) e sialidases (verde).

Este tipo de modelo de visualização é particularmente útil, pois permitiu que casos como o do *scaffold*90 fossem identificados e investigados, uma vez que o mesmo não apresenta nenhuma cobertura de *short reads* e uma cobertura média de *long reads* de 43,12x (Figura 6). Por meio de uma análise de similaridade entre a sequência completa do *scaffold*90 e o banco de dados do NR do GenBank foi revelada uma similaridade nucleotídica de aproximadamente 74% com uma sequência anotada como kDNA maxicírculo de *Trypanosoma rangeli* (código de acesso do GenBank: KJ803830.1). Este resultado parece sugerir a existência de uma preferência pela geração de fragmentos longos do kDNA desses parasitos, em detrimento de fragmentos curtos. É possível que a falta de fragmentos curtos de kDNA seja devida à um viés de sequenciamento em regiões pobres em conteúdo G+C. Westenberger e colaboradores (2006) apontam que o genoma mitocondrial de tripanosomatídeos apresenta uma das menores porcentagens de conteúdo G+C de todos os

protozoários, os quais apresentam a menor porcentagem de conteúdo G+C dentre os genomas mitocondriais de metazoários, exceto os de insetos. Ainda, Browne e colaboradores (2020) destacam um viés da metodologia de sequenciamento em função da amplificação por PCR de fragmentos curtos com baixa porcentagem G+C, que fica mais evidente em regiões fora da faixa de 45–65% de G+C, levando a uma cobertura falsamente baixa em sequências pobres em conteúdo G+C. De fato, a porcentagem G+C do *scaffold*90 é de 30,13%, estando abaixo da média de porcentagem G+C da versão 2 do genoma, o que torna possível validar essa hipótese. Além disso, também não se descarta a possibilidade de que este seja o resultado de um viés metodológico, decorrente dos procedimentos de preparo de biblioteca para sequenciamento. Por fim, é interessante notar que o *scaffold*90 possui um tamanho de 28.984 bases nucleotídicas, enquanto a sequência de kDNA maxicírculo de *T. rangeli* depositada no GenBank tem 25.288 bases nucleotídicas, o que pode ser um indicativo de que a sequência de kDNA maxicírculo da versão 2 do genoma de *T. rangeli* seja mais completa.

Figura 6. *Scaffold*90 não apresenta mapeamento de *short reads*, gerados pela plataforma Illumina, na versão 2 do genoma de *Trypanosoma rangeli* SC58.



Legenda: Os *long reads* obtidos da plataforma PacBio estão representados em preto. Eixo X representa as posições genômicas no *scaffold*. Eixo Y representa o número de *reads*, em ordem crescente de grandeza. A análise de similaridade feita entre a sequência deste *scaffold* e sequências do banco de dados do GenBank evidencia sua correspondência com sequências de kDNA maxicírculo de tripanosomatídeos.

Este resultado é importante do ponto de vista metodológico, pois demonstra o poder de resolução das estruturas genômicas pelos *long reads*, o que é de suma importância quando se considera o contexto genômico altamente repetitivo de organismos tripanosomatídeos. Ao utilizarem dados de *long reads* em sua montagem, Wang e colaboradores (2021) conseguiram refinar a estrutura do genoma das cepas Brazil e Y do *T. cruzi*, que são genomas maiores e mais repetitivos do que o genoma do *T. rangeli*. Neste mesmo estudo, os autores destacam a possibilidade do emprego deste tipo de dado de sequenciamento para obter uma montagem à nível de pseudo-cromossomos e uma maior resolução do conteúdo gênico da espécie, permitindo inclusive que fossem feitas análises de variação genética entre as cepas estudas (cepas Brazil e Y de *T. cruzi*). São poucos os genomas de tripanosomatídeos que foram obtidos ou revisitados por dados de sequenciamento de terceira geração (PacBio), sendo que a maioria ainda são conjuntos de dados obtidos de técnicas de sequenciamento de primeira (Sanger) ou segunda geração (454 e Illumina) (HERREROS-CABELLO et al., 2020). Outro fator a ser levado em consideração é o montador genômico utilizado neste trabalho: o programa Canu é conhecido por sua capacidade em lidar com regiões de repetição e de gerar montagens com menos erros, dando preferência pela devida geração de sequências consenso do que apenas pela contiguidade (KOREN et al., 2017; JAYAKUMAR; SAKAKIBARA, 2019).

Com relação às investigações de regiões teloméricas na versão 2 do genoma, foi possível identificar 13 regiões que contém repetições canônicas (TTAGGG)n, em diferentes *scaffolds*, as quais são consideradas possíveis regiões teloméricas (Tabela 3). Este número representa um acréscimo significativo às três regiões teloméricas previamente descritas no genoma de referência de *T. rangeli*. Entretanto, é importante destacar que não foram identificadas regiões teloméricas em ambas as extremidades de qualquer um dos *scaffolds* da versão 2, o que significa que não foi possível avaliar regiões centroméricas nos resultados gerados neste trabalho.

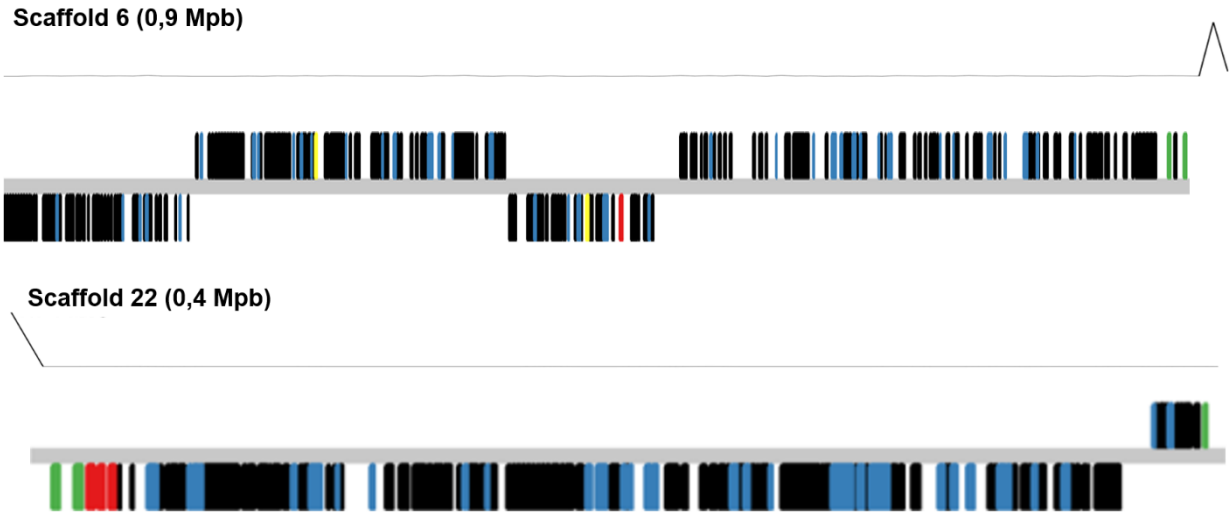
Tabela 3. Regiões teloméricas identificadas na versão 2 do genoma de *Trypanosoma rangeli* SC58.

Scaffold	**Início da repetição (pb)**	**Término da repetição (pb)**	**Tamanho do *scaffold* (pb)**
*scaffold*6	848.419	850.150	850.150
*scaffold*15	474.975	477.707	477.707
*scaffold*22	0	5.529	380.901
*scaffold*26	299.410	308.182	308.182
*scaffold*29	0	7.111	293.810
*scaffold*32	0	3.344	279.601
*scaffold*34	229.787	238.211	238.211
*scaffold*57	0	4.261	98.993
*scaffold*66	63.600	68.845	68.845
*scaffold*91	0	9.699	28.961
*scaffold*137	19.747	22.319	22.319
*scaffold*187	0	6.974	18.233
*scaffold*324	0	9.385	12.696

Legenda: pb = pares de bases.

Na Figura 7 são apresentados dois *scaffolds* representativos da análise de regiões teloméricas, evidenciando a concentração de repetições (TTAGGG)n na porção final do *scaffold6* e na porção inicial do *scaffold22*.

Figura 7. Identificação de regiões teloméricas de *Trypanosoma rangeli* SC58, ilustradas em dois *scaffolds* obtidos na versão 2 do genoma.



Legenda: Regiões genômicas estão representadas por um gráfico de linha. Pico na porção final do *scaffold*6 e na porção inicial do *scaffold*22 apontam para regiões de alta concentração da repetição telomérica canônica (TTAGGG)n. Tamanho do respectivo *scaffold* está indicado dentro dos parênteses. Mpb = milhões de pares de bases. Abaixo do gráfico, estão dispostas verticalmente as CDS (preto) ao longo da extensão dos *scaffolds* (cinza). Estão destacadas as CDS anotadas como hipotéticas (azul), GP63 (vermelho), sialidases (verde) e MASP (amarelo).

A identificação de regiões teloméricas permite o melhor entendimento da biologia de *T. rangeli*, uma vez que regiões subteloméricas representam sítios importantes de expressão gênica de genes multicópias em outros tripanosomatídeos. Em *T. brucei*, a região subtelomérica representa uma porção especializada do genoma, responsável pela expressão de diversas cópias de VSG (RAMIREZ, 2020a; LI, 2021). Em *T. cruzi*, esta região adjacente aos telômeros apresenta uma grande quantidade de cópias de trans-sialidases, especificamente do tipo II (também chamadas de gp85), DGF-1 e RHS (RAMIREZ, 2020a). Para o genoma de referência de *T. rangeli*, é descrito que a região subtelomérica é curta, em comparação ao *T. cruzi*, evidenciando também a perda de sintenia nestas regiões, justamente pela ausência de genes multicópias das famílias das (trans-) sialidases, DGF-1 e outros (STOCO et al., 2014). O menor número ou completa ausência de genes multicópias nestas regiões repetitivas do genoma de *T. rangeli* é mais uma evidência da diminuição do seu genoma, quando comparado ao *T. cruzi*. Desta forma, a resolução e a devida identificação destas 13 regiões teloméricas na versão 2 do genoma são um

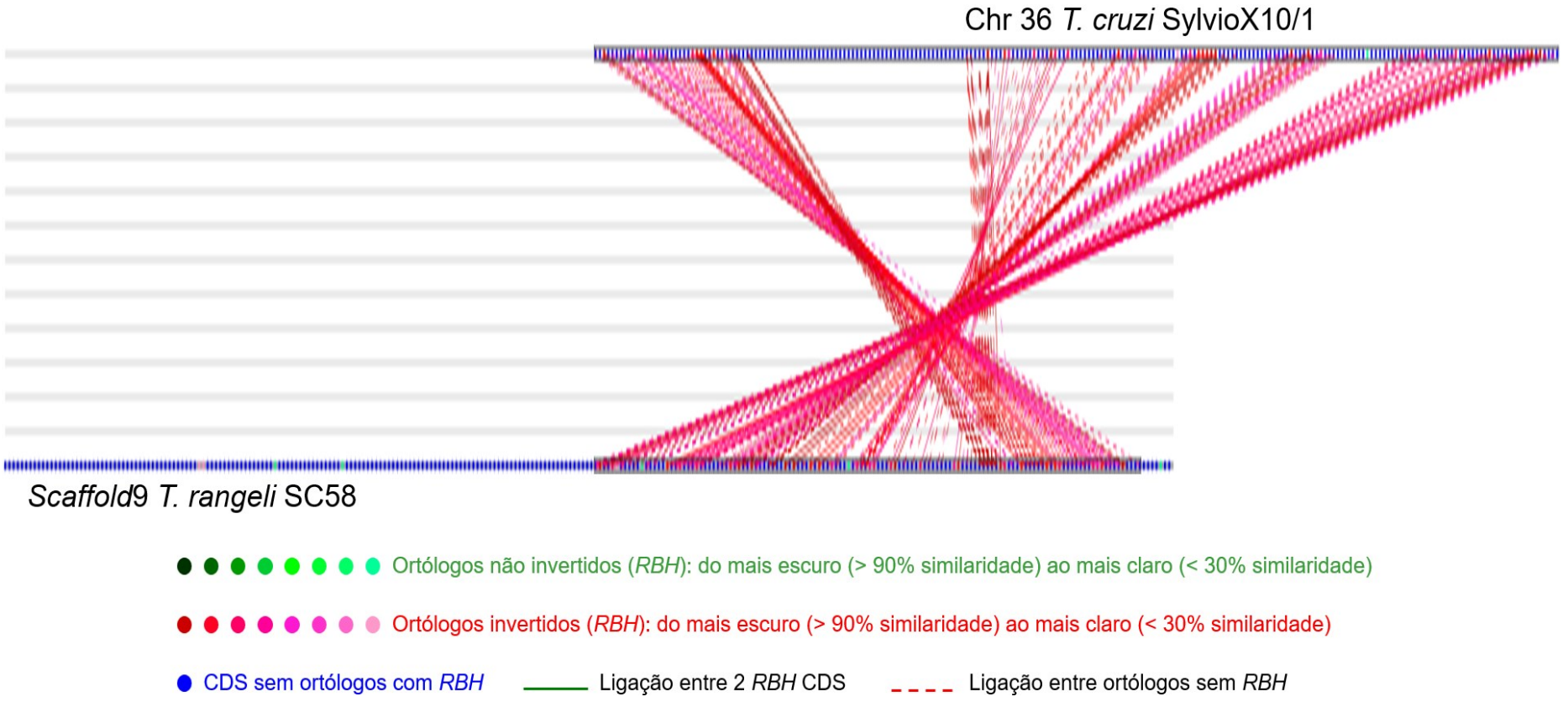
passo para possibilitar que este componente genômico possa ser estudado em *T. rangeli*.

O resultado da análise de sintenia entre os *scaffolds* obtidos na versão 2 do genoma de *T. rangeli* e os cromossomos de *T. cruzi* Sylvio X10/1 estão dispostos no Apêndice B. Como exemplo, na Figura 8, apresenta-se a sintenia entre o *scaffold*9 da versão 2 e o cromossomo 36 de *T. cruzi*, que é mútua e exclusiva, onde a porção final do *scaffold* cobre praticamente todo o cromossomo, apesar da região sintênica se apresentar através de uma inversão de sentido.

Praticamente todas as CDS presentes nesse cromossomo de *T. cruzi* apresentam correspondência com CDS localizadas na porção final do *scaffold* de *T. rangeli*, mas o contrário não é verdadeiro: uma vez que há uma perda de sintenia na porção inicial do *scaffold*. Essa diferença observada sugere que a sintenia com o cromossomo do *T. cruzi* foi estabelecida a partir da porção final do *scaffold*, uma vez que a sintenia foi determinada por meio de ortologia baseada em melhor correspondência mútua (do inglês, *Reciprocal Best Hit* – RBH). Desta forma, a exclusividade e a sintenia em apenas uma porção do *scaffold* indicam a possibilidade de que o *scaffold*9 represente uma região de duplicação genômica, porém não foram feitas nenhuma análise para testar esta hipótese.

Também foram observados casos em que dois ou mais *scaffolds* da versão 2 apresentam sintenia com todo um cromossomo de *T. cruzi*, inclusive no mesmo sentido. Em um desses casos, foi possível observar uma sintenia entre os *scaffolds* 61 e 10, com o cromossomo 19, que aponta uma sobreposição na porção final do *scaffold*61 e na porção inicial do *scaffold*10, o que parece sugerir uma possível ligação entre ambos os *scaffolds*. Nestes casos, fica evidente como a análise de sintenia pode ser utilizada como ferramenta para ajudar no fechamento de uma nova montagem genômica, a exemplo da estratégia aplicada inicialmente para o fechamento da montagem da cepa CL Brener do *T. cruzi* (RAMIREZ, 2020b).

Figura 8. Região sintênica identificada entre o cromossomo 36 de *Trypanosoma cruzi* Sylvio X10/1 e o *scaffold*9 da versão 2 do genoma de *T. rangeli* SC58.



Legenda: RBH = Melhor correspondência mútua.

6.2.2 Predição e anotação de características genômicas

Após a etapa de montagem genômica, a primeira predição realizada para o conjunto de dados de *T. rangeli* focou na predição de genes não traduzidos, especificamente rRNA e tRNA. No total, foram preditos 389 rRNA e 108 tRNA, dos quais 358 rRNA e 80 tRNA de tripanosomatídeos foram validados *in silico*. O conjunto de rRNA descrito na versão 2 contém todas as unidades típicas descritas em tripanosomatídeos, como 5.8S, 18S e 28S, assim como o conjunto de tRNA incluem todos aqueles necessários para carreamento dos 20 aminoácidos (Tabela 4). Apesar de nos genomas de referência de outros tripanosomatídeos esses genes também terem sido identificados (BERRIMAN et al., 2005; EL-SAYED et al., 2005a; STOCO et al., 2014), o número de RNA identificados na versão 2 foi superior a estes genomas, possivelmente em razão de uma melhor contiguidade no genoma montado.

Em seguida, foi realizada a predição de CDS da versão 2 do genoma de *T. rangeli*. No total, foram identificadas 9.102 CDS na versão 2, que é um número acima da quantidade de CDS preditas no genoma de referência. É importante ressaltar que o conjunto de treino utilizado na etapa de predição continha pouco mais de 7.800 sequências aminoacídicas descritas no genoma de referência. Neste sentido, é possível afirmar que o incremento no total de CDS preditas na montagem realizada neste trabalho pode ser devido a fatores como: (i) maior contiguidade da montagem; (ii) ao tamanho do genoma montado; e (iii) ao menor número de bases não identificadas e de espaçamentos na montagem. As informações gerais sobre as predições e anotações gênicas feitas neste trabalho estão presentes na Tabela 4.

No que diz respeito a anotação das CDS, considerando o estudo das diferentes famílias multigênicas, estão descritas na versão 2: quatro amastinas; quatro mucinas; 11 MASP; 286 GP63; e cinco KMP-11. Também foram anotadas CDS como: 28 DGF-1; 34 RHS; uma tuzina; e nenhuma VSG. Interessantemente, na versão 2 estão descritas 239 CDS anotadas como sialidases, das quais 196 foram previamente anotadas automaticamente como trans-sialidase e depois manualmente reanotadas, para se adequar ao fato de que a literatura aponta a ausência de atividade *trans*- em sialidases de *T. rangeli* (STOCO et al., 2014; CHIURILLO et al., 2016). Desta forma, o número de CDS anotadas como sialidases na versão 2 do genoma do *T. rangeli* é quase o dobro do que havia sido descrito no genoma de referência.

Tabela 4. Informações quantitativas sobre as predições e anotações de características genômicas na versão 2 do genoma de *Trypanosoma rangeli* SC58.

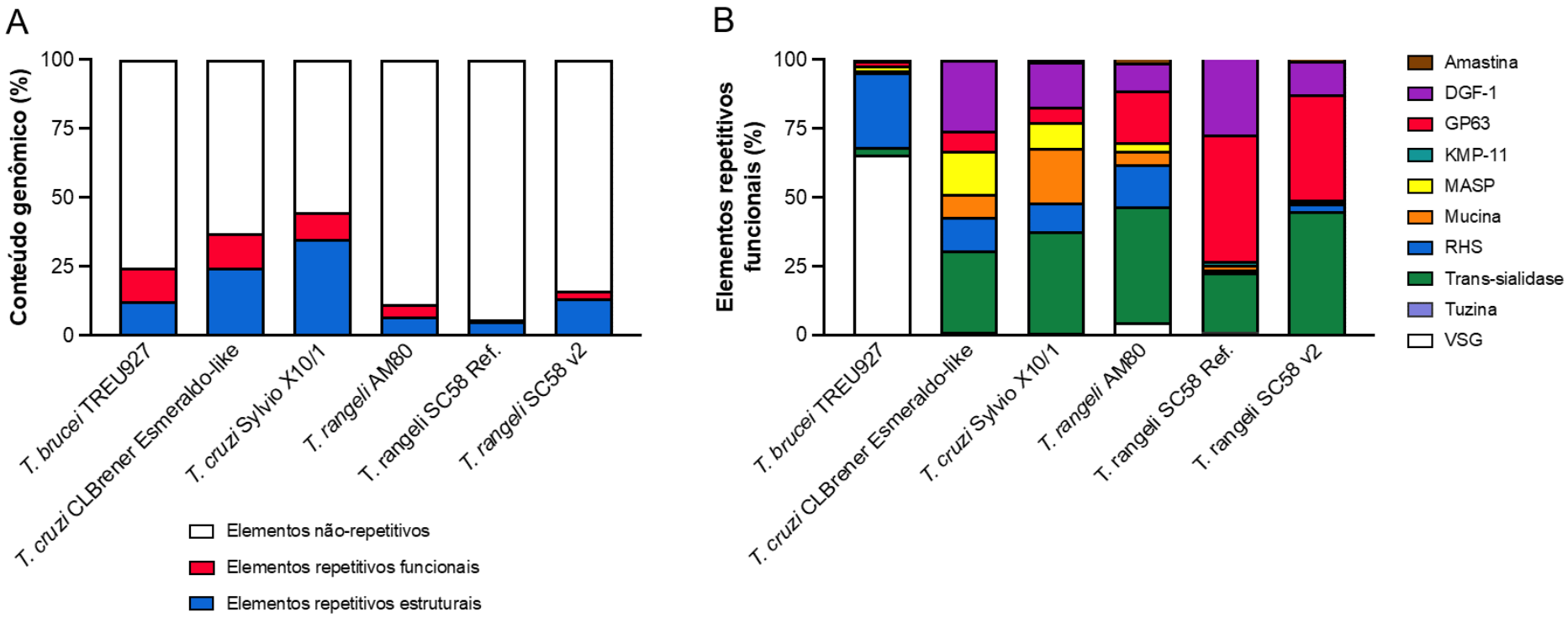
Número rRNA	358
Número tRNA	80
Número de CDS	9.102
Número de CDS anotadas	6.655
Número de CDS hipotéticas	2.447
Tamanho médio das CDS (aa)	468
CDS menores que 100 aa	360

Legenda: O “número de CDS hipotéticas” representa sequências que foram anotadas como hipotéticas durante o processo de anotação automática. aa = aminoácidos.

Uma das possíveis explicações para essa discrepância no número de CDS que codificam para proteínas de superfície, é baseada no fato que estes genes podem estar contidos em regiões repetitivas do genoma que não foram resolvidas na publicação do genoma de referência. Enquanto o genoma de referência é um bom retrato do genoma central da cepa SC58, graças a sua montagem feita com dados de *short reads* e suporte experimental, a estrutura repetitiva do genoma não foi resolvida. Proteínas como as sialidases, dentre outras que participam dos processos de sinalização e comunicação celular na interface parasito-hospedeiro, estão contidas nessas regiões repetitivas (BERNÁ et al., 2018; PITA et al., 2019), que foram elucidadas na versão 2 do genoma de *T. rangeli*. Além disso, quando se considera o conjunto de CDS descritas no genoma de referência de *T. rangeli*, é importante lembrar que a anotação gênica feita na época estava sujeita às limitações dos bancos de dados então utilizados. Além da contribuição de anotações contidas em bancos de dados modernos, também é possível destacar a contribuição do aumento da contiguidade do genoma como fator que possibilitou que, de modo geral, mais CDS fossem identificadas.

Considerando o aspecto de elementos repetitivos funcionais, na Figura 9 está apresentada uma comparação relativa da proporção de famílias multigênicas identificadas no genoma de tripanosomatídeos e na atual versão 2 do genoma de *T. rangeli*.

Figura 9. Gráficos de comparações normalizadas entre elementos repetitivos funcionais identificados nos genomas de tripanosomatídeos e na versão 2 do genoma de *Trypanosoma rangeli* SC58.



Legenda: (A) Quantidade normalizada de elementos repetitivos nos genomas de tripanosomatídeos e da versão 2, desconsiderando regiões teloméricas. (B) Porções relativas dos elementos repetitivos funcionais nos diferentes genomas e na versão 2, considerando as diferentes famílias multigênicas. “Ref.” indica o genoma de referência de *T. rangeli* SC58 (STOCO et al., 2014). “v2” indica a versão 2 do genoma de *T. rangeli* SC58. Para a versão 2 todas as CDS anotadas automaticamente como “trans-sialidase” foram manualmente reanotadas como “sialidases”, vide texto.

Dos elementos repetitivos funcionais da versão 2, destaca-se a grande representatividade de genes multicópias das famílias das sialidases e GP63, correspondendo a aproximadamente 45% e 38% do total, respectivamente. Este resultado corrobora a participação significativa já bem documentada destas famílias multigênicas no conteúdo genômico desta espécie (STOCO et al., 2014; BRADWELL et al., 2018). Añez-Rojas e colaboradores (2005) apontam que uma possível explicação para a quantidade considerável de sialidases no genoma de *T. rangeli* seja o seu papel importante durante os processos de adesão e invasão celular, no qual estas enzimas apenas estariam em seus estados ativos no ambiente do inseto vetor, mas não no do hospedeiro mamífero. Entretanto, mais recentemente, Wagner e colaboradores (2013), por meio de experimentos utilizando cromatografia líquida acoplada à espectrometria de massas, apontam que sialidases foram as proteínas mais abundante identificadas tanto nas formas epi- quanto nas formas tripomastigotas do *T. rangeli*. Uma análise de similaridade realizada entre as sialdiases identificadas na versão 2 aponta correspondências no conjunto de proteínas identificadas como sialidases no trabalho de Wagner e colaboradores (2013), com similaridade média de 59,93% e cobertura média de 87,64%, reforçando os achados resultantes de espectrometria de massas e das predições e anotações de CDS realizadas neste trabalho.

Deste resultado, chama a atenção nos dados de *T. rangeli* AM80 é a presença de cinco CDS anotadas como VSG, que não estão presentes nos dados da versão 2 tampouco no genoma de referência de *T. rangeli* SC58. As VSG são um grupo de moléculas de superfície bem descritas em tripanosomatídeos africanos (como *T. brucei*, *T. congolense* e *T. vivax*) e até então não foram descritas na superfície do *T. cruzi* ou do *T. rangeli* (EL-SAYED et al., 2005b; BERRIMAN et al., 2005; BRADWELL et al., 2018). Desta forma, é bastante provável que este seja um erro de anotação genômica no conjunto de dados da cepa AM80 de *T. rangeli*.

Das 9.102 CDS preditas, 2.447 foram anotadas como proteínas hipotéticas na versão 2, as quais representam 26,88% do total de CDS preditas, que é um número inferior às 66,25% CDS hipotéticas descritas no genoma de referência. Para fins comparativo, foi aplicada a mesma metodologia utilizada na versão 2 para fazer a reanotação das 7.475 CDS do genoma de referência. O resultado aponta uma parcela de apenas 33,78% de CDS anotadas como hipotéticas, que é uma diminuição considerável, mas ainda assim é superior ao total de CDS preditas e anotadas como

hipotéticas na versão 2 do genoma. Ainda, comparativamente, a quantidade de CDS hipotéticas na versão 2 do genoma de *T. rangeli* é menor com relação ao conjunto de CDS em genomas de outros tripanosomatídeos como *T. brucei* TREU927 e *T. cruzi* CL Brener Esmeraldo-like, com 35,82% e 47,15% das CDS totais anotadas como hipotéticas, respectivamente. A combinação de bancos de dados mais modernos e um genoma mais contíguo possibilitou uma redução na quantidade de CDS anotadas como hipotéticas na versão 2 do genoma de *T. rangeli*. Essa redução é importante, pois um dos desafios atuais em genômica é a grande quantidade de sequências pouco informativas ou artefatos de ferramentas preditivas que inundam os bancos de dados moleculares (IJAQ et al., 2015; GOUDEY et al., 2022). Por outro lado, a detecção de CDS anotadas como hipotéticas pode ampliar a quantidade de sequências potencialmente relevantes, que é um dos benefícios de estudos genômicos, uma vez que tais sequências podem ser testadas como novos marcadores moleculares ou alvos farmacológicos para a descoberta, triagem e desenvolvimento de fármacos (MOHAN, 2012; SHAHBAAZ; IMTAIYAZHASSAN; AHMAD, 2013). Ijaq e colaboradores (2015) defendem que apesar da existência de diversos métodos *in silico* para caracterizar proteínas, esses devem ser realizados em conjunto com experimentos de bancada para confirmar se as proteínas identificadas são reais ou artefatos, para garantir que os estudos genômicos sejam mais informativos.

Como ilustração, na Figura 10 estão representadas as disposições das CDS preditas e anotadas nos *scaffolds* com mais de 100 mil pares de bases da versão 2, enquanto a representação das CDS nos demais *scaffolds* pode ser encontrada no Apêndice C.

Figura 10. Distribuição das CDS preditas e anotadas nos *scaffolds* com mais de 100 mil pares de bases obtidos na versão 2 do genoma de *Trypanosoma rangeli* SC58, com destaque para famílias multigênicas.

Legenda: As CDS estão dispostas verticalmente (preto), ao longo da extensão de cada *scaffold* (cinza). Estão destacadas as CDS anotadas como hipotéticas (azul), GP63 (vermelho), sialidases (verde), DGF-1 (roxo), MASP (amarelo), amastinas (marrom) e mucinas (laranja).

Além de apresentar uma visão geral das CDS que compõem a versão 2 do genoma de *T. rangeli*, a Figura 10 também destaca uma das características marcante do genoma de tripanosomatídeos, que é sua organização estruturada em DGCs. Díaz-Viraqué e colaboradores (2023) apontam que são nestas regiões de alteração da fita de leitura das CDS onde os sensos de transcrição convergem ou divergem. Berná e colaboradores (2018) discutem como é possível compartimentalizar o genoma de *T. cruzi* em duas porções, considerando a sintenia de DGCs: em (i) compartimento central, que são regiões sintênicas compostas por CDS com funções conhecidas ou por CDS anotadas como hipotéticas e conservadas entre diferentes espécies de tripanosomatídeos; ou (ii) compartimento disruptivo, que é composto principalmente por famílias multigênicas, geralmente localizadas em regiões subteloméricas, onde há uma perda considerável de sintenia entre as espécies. Em um trabalho que utiliza dados de *long reads*, Müller e colaboradores (2018) foram capazes de refinar a estrutura genômica da cepa Lister 427 do *T. brucei*, apontando que a porção disruptiva do genoma desta espécie também está localizada em regiões subteloméricas, dando destaque para o grande número de genes multicópias que codificam para VSG nestas regiões.

Fazendo uma ligação entre as predições e anotações de CDS e as regiões subteloméricas de *T. rangeli*, Stoco e colaboradores (2014) descreveram apenas uma CDS anotada como “mercaptopyruvate sulfurtransferase”, a qual estava localizada na região subtelomérica do *scaffold* AUPL01006408; enquanto cinco CDS foram identificadas como “mercaptopyruvate sulfurtransferase” na versão 2. Uma dessas CDS (g3726) está localizada na porção central do *scaffold*14, à medida em que as outras quatro CDS (g4110, g4111, g4112 e g4113) estão todas dispostas em tandem na porção final do *scaffold*16, mais precisamente dentro da região dos 10 mil pares de bases finais do *scaffold*. Apesar disso, a análise de regiões teloméricas não identificou repetições (TTAGGG)n no *scaffold*16, tampouco as quatro CDS anotadas como “mercaptopyruvate sulfurtransferase” são as últimas CDS preditas na porção final deste *scaffold*. Embora os fragmentos longos tenham permitido identificar e anotar mais CDS, ainda não foi possível resolver completamente certas regiões no genoma de *T. rangeli*, como as repetições canônicas de regiões teloméricas. De certa forma, este achado enaltece montagens genômicas que combinam o uso de dados de *long reads* associados a técnicas como a captura da conformação de cromatina (Hi-C) (BELTON et al., 2012). Tal metodologia de montagem foi capaz de resolver

regiões repetitivas até mesmo em genomas maiores e mais repetitivos de outras espécies de tripanosomatídeos do que o de *T. rangeli* (MÜLLER et al., 2018; WANG et al., 2021; DÍAZ-VIRAQUÉ et al., 2023).

A partir da anotação das CDS da versão 2, foi possível avaliar a presença da maquinaria molecular de RNA de interferência (iRNA). Dos cinco principais componentes de iRNA descritos em *T. brucei* (TbDCL1, TbDCL2, TbAGO1, TbRIF4 e TbRIF5) (NGÔ et al., 1998), apenas três deles puderam ser identificados através dos resultados de anotação automática: DCL1, AGO1 e RIF4. Ainda, foi feita uma análise de similaridade entre as proteínas de *T. brucei* e os *scaffolds* da versão 2, para avaliar a presença ou ausência dos demais componentes. Os resultados desta análise confirmam: (i) a presença de DCL1 na versão 2, com 98% de cobertura a partir da sequência do *T. brucei*, assim como uma CDS anotada como hipotética de tamanho similar à DCL1, na mesma região do respectivo *scaffold*; (ii) a identificação de fragmentos pseudogenizados de DCL2 na versão 2, apresentando múltiplos *stop* códons em suas sequências; (iii) a presença de AGO1 na versão 2, porém em forma de pseudogene com relação à sequência do *T. brucei*, sendo identificando uma CDS distinta anotada automaticamente como “argonaute-like protein”; (iv) a presença de RIF4 na versão 2, apesar de apresentar a menor cobertura (56%) na análise de similaridade dentre as sequências utilizadas; e (v) a identificação de fragmentos de RIF5 na versão 2, todos com cobertura considerável (86%) com a sequência correspondente de *T. brucei*, os quais também possuem stop códons em suas sequências.

Posteriormente, as anotações das CDS da versão 2 foram refinadas com base em dados de transcriptômica e proteômica de *T. rangeli*, por meio do AnnotaPipeline (MAIA et al., 2022). Considerando os resultados gerados pela *pipeline* foi possível classificar as CDS da versão 2 em oito categorias diferentes, baseando-se em três critérios: (1) anotação disponível nos bancos de dados utilizados ou anotadas como hipotética; (2) evidência de transcrição, obtida através da quantificação de *reads* de RNA-seq; e (3) evidência de tradução, sustentada pela identificação de peptídeos correspondentes. O número total de classes e o quão representativo é cada uma delas, considerando o conjunto total de 9.102 CDS, está apresentado na Tabela 5.

Tabela 5. Tabela de classificação das CDS preditas e anotadas na versão 2 do genoma de *Trypanosoma rangeli* SC58.

Classificação	Anotação Hipotética	Evidência de Transcrição	Evidência de Tradução	Número de Sequências	%
Classe 1	Sim	Não	Não	1.571	17,26
Classe 2	Não	Não	Não	3.157	34,68
Classe 3	Sim	Sim	Não	323	3,55
Classe 4	Sim	Não	Sim	464	5,10
Classe 5	Não	Sim	Não	523	5,75
Classe 6	Não	Não	Sim	2.145	23,57
Classe 7	Sim	Sim	Sim	89	0,98
Classe 8	Não	Sim	Sim	830	9,12

As CDS classificadas na classe 1 podem ser consideradas como verdadeiras hipotéticas, uma vez que são anotadas como hipotéticas e não apresentam qualquer evidência experimental. Por outro lado, uma das classificações que mais chamam a atenção são as CDS classificadas na classe 7, que são anotadas como hipotéticas, porém demonstram evidência experimental de transcrição e tradução, por experimentos independentes. Esta classe é composta por CDS de grande interesse biológico, pois não possuem função conhecida, mas são expressas. Por causa disso, essas seriam os primeiros alvos de interesse para serem estudadas e caracterizadas.

Por fim, as informações acerca da caracterização *in silico* realizada para todas as CDS preditas e anotadas na versão 2 do genoma de *T. rangeli* estão dispostas no Apêndice D.

6.2.3 Elementos repetitivos

Da análise de elementos repetitivos, desconsiderando regiões teloméricas apresentadas anteriormente, a versão 2 do genoma de *T. rangeli* apresenta 13,53% de elementos repetitivos estruturais, que correspondem à aproximadamente 4,1 milhões de pares de bases. Destes, a maior parte é representada por repetições

intercaladas, com 65,10%, seguidas por regiões de repetições simples, com 21,42%. Por outro lado, a fração de elementos repetitivos estruturais no genoma de *T. cruzi* cepa Sylvio X10/1 representa 35,08% do tamanho total do genoma, ou aproximadamente 15 milhões de pares de bases. Da mesma forma, a maior parcela destas repetições genômica é representada por repetições intercaladas, com 81,83%, porém, diferente de *T. rangeli*, a segunda maior parcela é representada por repetições em regiões de macro- e microssatélites, juntas correspondendo a 9,13%.

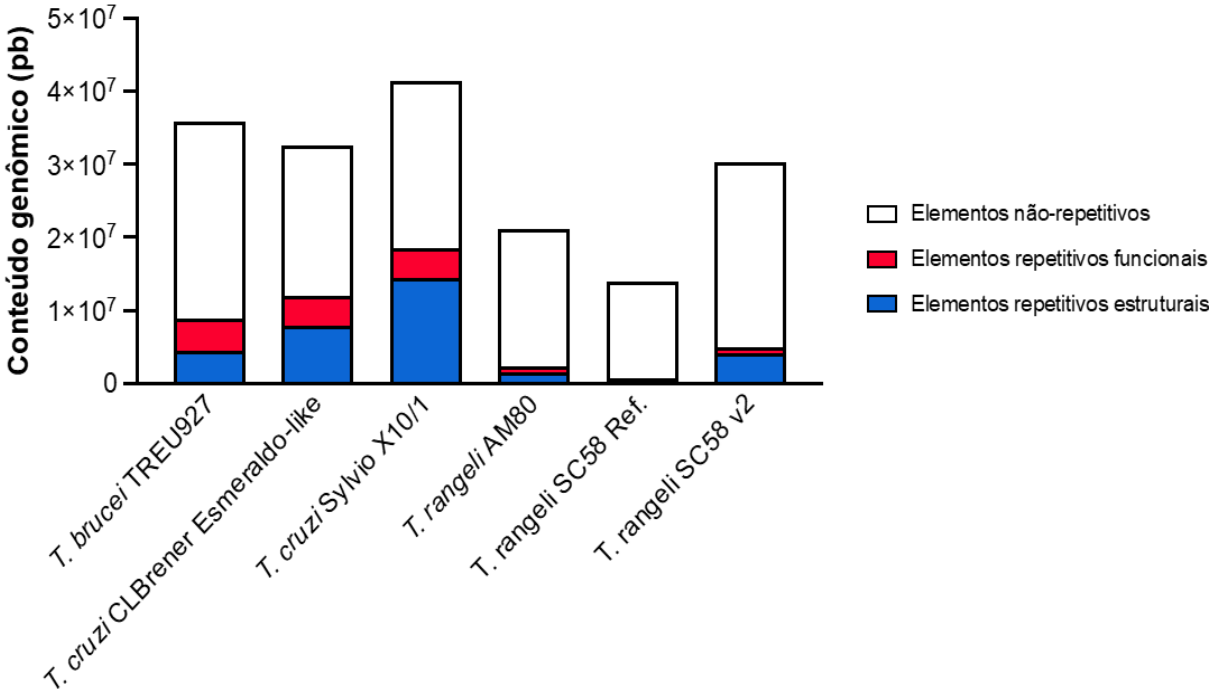
Conceitualmente, microssatélites são pequenos agrupamentos de repetições (geralmente menores do que 200 pares de bases), cujas unidades geralmente têm menos do que 5 pares de base, enquanto macrossatélites são grandes regiões de repetição (mais de centenas de milhares de pares de bases), com unidades repetitivas acima de 25 pares de bases (WICKSTEAD; ERSFELD; GULL, 2003). De fato, estes resultados corroboram os achados de El-Sayed e colaboradores (2005a), que já haviam descrito a participação significativa de regiões satélites na porção repetitiva do genoma de *T. cruzi*.

As repetições intercaladas presentes na versão 2 do genoma de *T. rangeli*, que representam 8,81% do genoma, são compostas principalmente por elementos de repetição terminal longa (do inglês, *Long Terminal Repeats* – LTR) e LINEs, dentre outros. Apesar de ambos serem retroelementos, os mecanismos de origem desses elementos transponíveis no genoma são distintos: os elementos LTR são transcritos reversos de um intermediário de RNA, que se duplica e então é transposto como DNA dupla-fita, possivelmente em outras regiões do genoma; enquanto LINEs são transpostos diretamente no sítio de sua integração a partir da transcrição reversa de seu próprio mRNA, por meio de uma RNA polimerase contida em sua sequência (CHARLESWORTH; SNIEGOWSKI; STEPHAN, 1994; WICKSTEAD; ERSFELD; GULL, 2003). O outro tipo de elemento repetitivo estrutural mais prevalente na versão 2 são as repetições simples, que representam 2,90% do conteúdo desta montagem. Este tipo de elemento repetitivo é representado por motivos de DNA repetitivo compostos por di-, tri-, tetra- ou poli-nucleotídeos (CHARLESWORTH; SNIEGOWSKI; STEPHAN, 1994). Na versão 2, os três tipos mais prevalentes de repetições simples são as que contém os motivos (AT)n, (CA)n e (TG)n, com tamanhos variados. Talavera-López e colaboradores (2021) apontam que na versão mais recente do genoma da cepa Sylvio X10/1 de *T. cruzi* as repetições simples foram mais comumente identificadas ao redor de regiões que concentram CDS de famílias

multigênicas. Neste contexto, a resolução e identificação destas regiões no genoma do *T. rangeli* pode permitir que sítios de recombinação de fatores de virulência sejam melhor estudados.

A comparação do número absoluto de pares de bases correspondentes às regiões repetitivas identificadas nos genomas de diferentes espécies de tripanosomatídeos está representado na Figura 11.

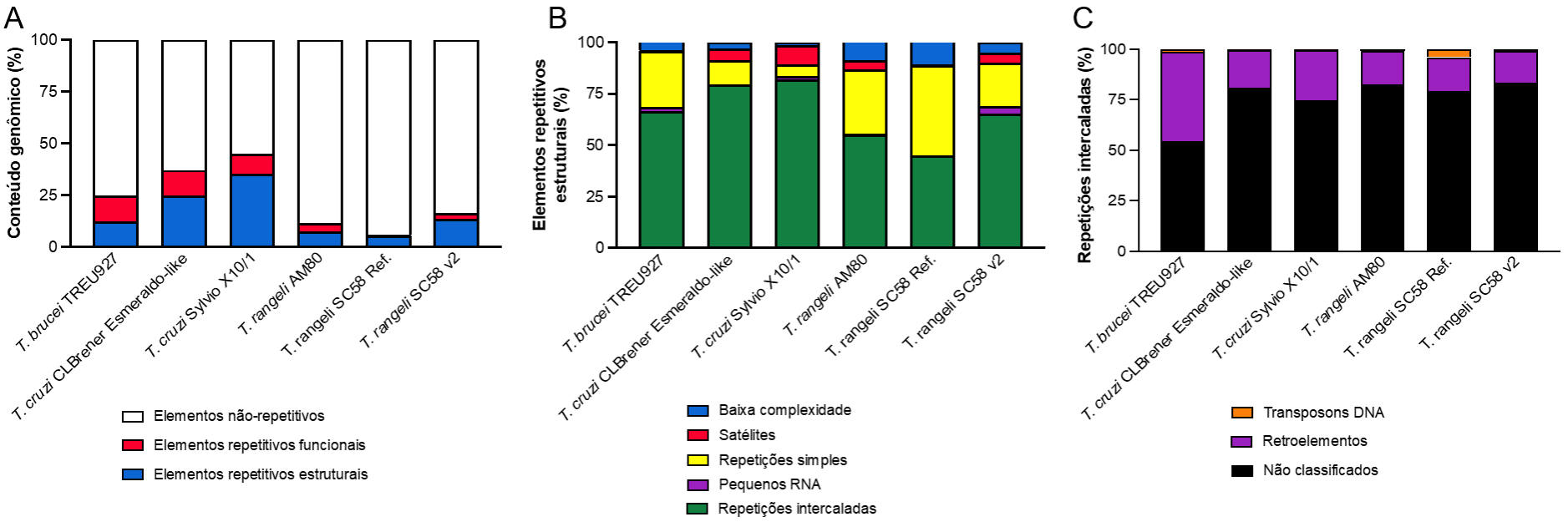
Figura 11. Comparação do número de pares de bases totais e porções repetitivas nos genomas de tripanosomatídeos e na versão 2 do genoma de *Trypanosoma rangeli* SC58.



Legenda: pb = pares de bases. “Ref.” indica o genoma de referência de *T. rangeli* SC58 (STOCO et al., 2014). “v2” indica a versão 2 do genoma de *T. rangeli* SC58.

Dentre os conjuntos de dados da espécie *T. rangeli* analisados, a versão 2 é o que apresenta o maior número de pares de bases. Além dos fatores dispostos anteriormente no texto – o refinamento do tamanho do genoma montado, a melhoraria de sua contiguidade e a redução da quantidade de bases não identificadas do genoma – é possível também que a presença de sequências de kDNA, montadas graças a utilização de dados de *long reads*, influenciem a quantidade total de pares de bases observadas. Um panorama comparativo da análise de elementos repetitivos estruturais da versão 2 com outros genomas de espécies de tripanosomatídeos está exposto na Figura 12.

Figura 12. Gráficos de comparações normalizadas entre elementos repetitivos estruturais identificados nos genomas de tripanosomatídeos e na versão 2 do genoma de *Trypanosoma rangeli* SC58.



Legenda: (A) Quantidade de elementos repetitivos nos genomas de tripanosomatídeos e da versão 2, desconsiderando regiões teloméricas. (B) Explanação dos elementos repetitivos estruturais nos diferentes genomas e na versão 2. (C) Detalhamento da classificação de sequências repetitivas intercaladas. “Unclassified” são todos os elementos repetitivos que encontram correspondência com sequências não identificadas no banco de dados utilizado. “Ref.” indica o genoma de referência de *T. rangeli* SC58 (STOCO et al., 2014). “v2” indica a versão 2 do genoma de *T. rangeli* SC58.

Considerando tanto aspectos repetitivos estruturais quanto aspectos funcionais, a montagem realizada neste trabalho possui 16,26% de conteúdo genômico repetitivo, que correspondem à aproximadamente 5 milhões de pares de base. Comparativamente, a porção repetitiva no conjunto de dados de *T. cruzi* Sylvio X10/1 corresponde à 44,86%, ou aproximadamente 18,5 milhões de pares de base, que é uma proporção bastante próxima do esperado, de acordo com dados da literatura (EL-SAYED et al., 2005b; PITA et al., 2019).

A disparidade entre a expectativa da fração repetitiva do genoma de *T. cruzi*, estimada em aproximadamente 50% para os dados da cepa CL Brener (EL-SAYED et al., 2005a), e a observação de 44,86% de repetição nos dados de *T. cruzi* Sylvio X10/1 pode ser atribuída principalmente a variações intraespecíficas, principalmente à nível genômico (ZINGALES, 2018), mas também não podem ser desconsideradas as diferenças metodológicas empregadas neste trabalho.

No que diz respeito à cepa CL-Brener de *T. cruzi*, um de seus principais aspectos biológicos é sua característica híbrida (EL-SAYED et al., 2005a). Zingales (2018) aponta o questionamento sobre a prevalência do modelo evolutivo clonal nas populações de *T. cruzi*, pois evidências sugerem que eventos de hibridização, como observado na cepa CL Brener, contribuem para a estrutura populacional e evolução de grupos distintos em *T. cruzi*. A contribuição de dois haplótipos de grupos diferentes (TcII e TcIII) para o conteúdo gênico da cepa CL Brener (TcVI) destaca sua heterozigose e o acúmulo de porções repetitivas (FRANZÉN et al., 2011). Isso contrasta com o genoma menor, menos repetitivo e não híbrido da cepa Sylvio X10/1 (TcI) do *T. cruzi* (FRANZÉN et al., 2011; TALAVERA-LÓPEZ et al., 2021), o que pode explicar as diferenças nos resultados obtidos aqui. Além disso, para ilustrar as diferenças metodológicas, na análise de elementos repetitivos na versão 2 do genoma de *T. rangeli*, foram empregados bancos de dados mais atualizados, que continham apenas repetições previamente identificadas em tripanosomatídeos (LIAO et al., 2022), em comparação às informações contidas nos bancos de dados utilizados por El-Sayed e colaboradores (2005). No entanto, é importante apontar que a anotação desses bancos de dados modernos também representa um desafio, pois a classificação automatizada de dados moleculares inevitavelmente incorpora os vieses inerentes do banco de dados consultado. Neste contexto, destaca-se novamente o fenômeno bem conhecido do acúmulo de sequências pouco informativas em bancos de dados, exemplificado aqui pela considerável proporção de sequências repetitivas

categorizadas como "Não classificados" na porção C da Figura 12, juntamente com duplicatas de sequências e anotações funcionais incorretas ou inadequadamente redigidas que permeiam os repositórios de informações moleculares (KARP, 1998; GOUDEY et al., 2022).

6.2.4 Homologia de CDS preditas

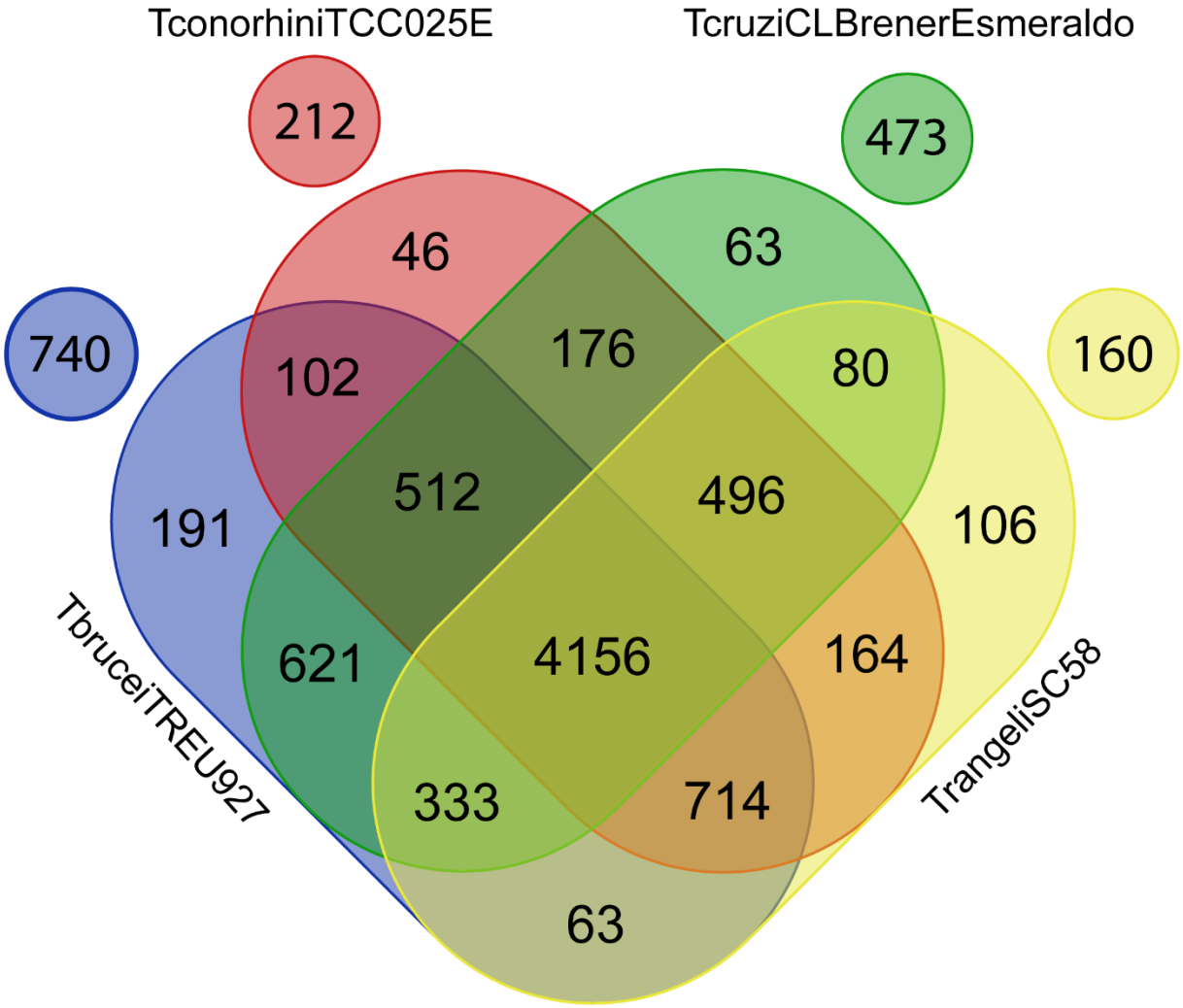
A análise de homologia realizada entre as quatro espécies de tripanosomatídeos comparadas neste trabalho demonstra que 95,51% de todas as CDS das espécies são ortólogas. Este resultado está de acordo com dados da literatura, que apontam que organismos tripanosomatídeos compartilham cerca de 90% dos seus genes (MOREL et al., 2005; STOCO et al., 2014; REIS-CUNHA; BARTHOLOMEU, 2019). Ainda, em média, os agrupamentos formados contêm 4,3 CDS cada, em sua grande maioria apresentando ao menos uma CDS de cada espécie. As métricas gerais da análise de homologia das espécies de tripanosomatídeos estão apresentadas na Tabela 6.

Tabela 6. Métricas gerais da análise de homologia realizada entre *Trypanosoma brucei* (cepa TREU927), *T. conorhini* (cepa 025E), *T. cruzi* (cepa CL Brener Esmeraldo-like) e *T. rangeli* (cepa SC58, versão 2).

Número de CDS analisadas	35.298
Número de agrupamentos formados	7.823
Número de CDS em agrupamentos	33.713 (95,51%)
Agrupamentos com CDS de todas espécies	4.156 (53,13%)

A distribuição dos 7.823 agrupamentos formados entre as quatro espécies de tripanosomatídeos estudados, assim como os seus respectivos *singletons*, está representada na Figura 13.

Figura 13. Diagrama de Venn representando a distribuição dos grupos formados entre as CDS de *Trypanosoma brucei* TREU927, *T. conorhini* 025E, *T. cruzi* CL Brener Esmeraldo-like e da versão 2 do genoma de *T. rangeli* SC58, assim como seus respectivos *singletons*.



Legenda: Resultado da análise de homologia entre as CDS de quatro espécies de tripanosomatídeos. Estão representados: *T. brucei*, em azul; *T. conorhini*, em vermelho; *T. cruzi*, em verde; e *T. rangeli*, em amarelo. Os números dentro do Diagrama de Venn representam agrupamentos ortólogos e parálogos, os quais contém as CDS analisadas. Os *singletons* estão representados como círculos fora do Diagrama de Venn.

A análise de homologia destaca a presença de um genoma central compartilhado entre esses organismos, que é representado por um agrupamento de 4.156 grupos de CDS ortólogas. Comparativamente, Stoco e colaboradores (2014) identificaram 403 agrupamentos ortólogos compartilhados entre o *T. cruzi* e o *T. rangeli*, à medida que neste trabalho, apenas 80 grupos ortólogos são compartilhados entre as respectivas espécies. Entretanto, se desconsiderarmos a contribuição dos

dados de *T. conorhini* na organização dos agrupamentos ortólogos, encontramos neste estudo 576 agrupamentos ortólogos entre o *T. cruzi* e o *T. rangeli* (80+496). Isto representa um aumento de aproximadamente 30% na identificação de sequências ortólogas entre as espécies. As diferenças entre os resultados deste estudo e os de Stoco e colaboradores (2014) podem ser atribuídas a quatro fatores: (i) diferentes conjuntos de dados, pois as CDS de ambas espécies fazem parte de diferentes montagens genômicas; (ii) diferentes metodologias, uma vez que a análise de ortologia de Stoco e colaboradores (2014) foi manualmente curada a partir dos resultados gerados pelo programa OrthoMCL, enquanto que no presente estudo os resultados foram obtidos de forma automática pelo programa OrthoFinder; (iii) diferentes espécies analisadas, de forma que a inclusão dos dados da espécie *T. conorhini*, que é mais próxima do clado do *T. cruzi* do que a espécie *Leishmania major*, alterou a formação dos agrupamentos ortólogos; e (iv) características da montagem da versão 2 do genoma de *T. rangeli*, pois como exposto anteriormente, esta montagem representa um conjunto de dados mais completo da espécie, permitindo a identificação de mais sequências ortólogas compartilhadas com outras espécies de tripanosomatídeos.

Com relação aos *singletons* que fizeram parte da análise de homologia das espécies analisadas, 740 CDS de *T. brucei* foram classificadas como *singletons*, enquanto 473 *singletons* foram identificados em *T. cruzi*. Os dados de *T. conorhini* demonstram que 212 CDS são *singletons*, das quais 52 contém anotação em banco de dados e 160 são anotadas como hipotéticas (13 dessas CDS hipotéticas contém anotação funcional pela IPR ou GO). Da versão 2 do genoma de *T. rangeli*, 160 CDS são consideradas *singletons*, das quais 95 contém anotação em banco de dados e 65 são anotadas como hipotéticas (oito dessas CDS hipotéticas dispõe de anotação funcional pela IPR ou GO). Existe a possibilidade de que essas CDS sejam artefatos de predição, principalmente as que não possuem anotação gênica ou funcional, uma vez que preditores gênicos raramente conseguem ultrapassar uma acurácia de 80% em suas predições (YANDELL; ENCE, 2012). Porém, também existe a possibilidade de que essas CDS representem sequências exclusivas de cada espécie, ou seja, sejam *singletons* verdadeiros da cepa.

Do conjunto de 160 *singletons* identificados em *T. rangeli*, 92 desses encontram correspondência com CDS descritas no genoma de referência e, portanto, 68 podem ser considerados exclusivos da versão 2. Esse conjunto de 68 CDS tiveram

suas possíveis funções genéticas analisadas pelas ferramentas de ortologia do banco de dados do KEGG, para que fossem identificadas suas respectivas participações em vias metabólicas (KANEHISA; SATO; MORISHIMA, 2016). Foi possível atribuir tal classificação de função para 11 das 68 CDS analisadas. Quatro CDS foram categorizados em diferentes vias metabólicas, como: metabolismo de carboidratos (2, KO9101), representado por uma coenzima NAD+ (do inglês, *Nicotinamide Adenine Dinucleotide*) e uma malato desidrogenase; metabolismo de amino ácidos (1, KO09105), no qual está descrito uma IGP desidratase; e metabolismo de lipídeos (1, KO09103), por uma Acil-CoA desidrogenase, que é uma enzima que participa do metabolismo de beta-oxidação de ácidos graxos de cadeia longa. As demais sete CDS não foram mapeadas em vias metabólicas, mas foram classificadas em diversos processos biológicos, como por exemplo: biogênese mitocondrial, tráfego de vesículas membranares e proteínas associadas ao complexo de gama-tubulina.

O complexo de gama-tubulina é um arranjo de componentes proteicos que auxilia no evento de formação de microtúbulos, chamado de nucleação de microtúbulos (KOLLMAN, et al., 2010). Tipicamente, os microtúbulos são polimerizados espontaneamente a partir de dímeros de alfa- e beta-tubulina, mas este processo é limitado pela disponibilidade de energia – na forma de GTP – o que faz com que as taxas de polimerização sejam controladas (KOLLMAN et al., 2011). Em alguns eucariotos, esta limitação é atenuada graças a participação de proteínas do complexo gama-tubulina, que formam uma estrutura em forma de anel que age como suporte para a nucleação de alfa- e beta-tubulina (KOLLMAN et al., 2011). No genoma de referência de *T. rangeli* há uma CDS (TRSC58_04341) identificada como “gamma tubulin”, portanto a identificação de outras duas CDS nesta mesma família proteica é uma adição importante que agrega ao conteúdo genômico da espécie. Biologicamente, McKean e colaboradores (2003) avaliaram a função da gama-tubulina na morfogênese de flagelo do *T. brucei*, através da indução por iRNA, sendo capazes de verificar que essas proteínas são importantes para a nucleação do par de microtúbulos centrais do axonema. Brevemente, esse mesmo estudo observou que o fenótipo de parasitos transfectados nas primeiras 48 horas não apresentavam diferenças morfológicas, mas sim um grande número de células imóveis, ao passo que o fenótipo populacional observado após 72 horas de transfecção era de células com um grande núcleo, indicativo da falha na formação de fuso intranuclear (MCKEAN et al., 2003). É possível que o número de genes associados a esse complexo proteico

no genoma do *T. rangeli* seja mantido por uma pressão seletiva para a conservação da estrutura dos microtúbulos centrais do flagelo, sendo necessários mais estudos para verificar este e outros aspectos. Por fim, o principal fator que contribuiu para que a maioria dessas 68 CDS exclusivas da versão 2 não tivessem sua classificação efetuada pela análise do KEGG é a grande quantidade de repetições em suas sequências, o que indica que muito provavelmente essas CDS são, na verdade, artefatos de predição gênica.

Também foi realizada uma análise do conjunto de CDS anotadas como hipotéticas na versão 2 do genoma de *T. rangeli* que foram classificadas como *singletons* pela análise de ortologia. Este conjunto é representado por 65 sequências, que podem ser considerados um conjunto de genes órfãos. Por definição, genes órfãos são aqueles que não compartilham similaridade com nenhum outro gene anotado em genomas disponíveis em bancos de dados, sendo que tais genes podem compreender entre 10% e 20% das CDS preditas em um novo genoma montado (SATOSHI, 2004; TAUTZ; DOMAZET-LOŠO, 2011). No entanto, a quantidade de potenciais genes órfãos na versão 2 é de 0,71%, que é uma porcentagem muito abaixo da estimada pela literatura. Deste conjunto, destaca-se um grupo de oito CDS que apresentam anotação funcional pela IPR ou pela GO, apesar de serem anotadas como hipotéticas. Como exemplo, destas oito, apenas uma CDS (g7649.t1) apresenta evidência de expressão, com a respectiva anotação funcional do Pfam e InterProScan indicando que se trata de uma proteína de função desconhecida, cujo domínio "P-loop containing nucleoside triphosphate hydrolase" já havia sido descrito anteriormente e classificado em uma superfamília gênica (código de acesso do Superfamily: SSF52540). Este domínio é o mais prevalente entre os vários domínios distintos de proteínas que se ligam a nucleotídeos. A reação catalisada mais comum por enzimas desta família é a hidrólise da ligação beta-gama de um nucleosídeo tri fosfatado (NTP), que geralmente é utilizada para induzir mudanças conformacionais em outras moléculas (LEIPE; KOONIN; ARAVIND, 2004). Deste mesmo conjunto, outra CDS (g6813.t1) chama a atenção por possuir: evidência de sinal peptídeo e localização extracelular; anotação funcional proteolítica pela GO; e é anotada como um pertencente da família multigênica GP63, pelos bancos de dados do InterProScan e Pfam.

6.3 CONCLUSÕES

Neste trabalho foram utilizados dados de sequenciamento de segunda e terceira geração para a obtenção de uma nova versão do genoma da cepa SC58 de *Trypanosoma rangeli*. Os resultados obtidos demonstram que esta nova versão do genoma é mais contígua (possui *scaffolds* maiores), menos fragmentada (possui um menor número de *scaffolds*) e contém praticamente toda a informação existente no genoma de referência desta cepa. Logo, é possível concluir que a utilização de dados de *long reads* fornece uma maior resolução estrutural do genoma, apesar de não ter sido possível atingir uma resolução em nível cromossômico (telômero-a-telômero). Ainda, fica evidente que o genoma de referência da cepa é representativo e informativo, a despeito das tecnologias de sequenciamento utilizadas para sua montagem. Interessantemente, este é o maior genoma descrito para uma cepa de *T. rangeli* até então, sendo constituído por aproximadamente 30 milhões de pares de base. De toda forma, os dados obtidos acerca de elementos repetitivos estruturais e funcionais na versão 2 reforçam as evidências de que o genoma de *T. rangeli* é menor e menos repetitivo comparado ao do *T. cruzi*.

A utilização de dados de sequenciamento do tipo *long reads* permitiu não somente a identificação de sequências de kDNA, como também auxiliou na resolução de regiões repetitivas do genoma. As repetições intercaladas e repetições simples representam quase que toda a porção de elementos repetitivos estruturais identificados na versão 2 do genoma de *T. rangeli*. Esta identificação possibilita um melhor entendimento sobre como os elementos repetitivos surgem e se transpõe pelo genoma desta espécie, influenciando sua adaptação e evolução. Ainda, destaca-se que são justamente em regiões repetitivas onde é possível encontrar a maior concentração de CDS classificadas como fatores de virulência em tripanosomatídeos patogênicos. Desta forma, é possível que os resultados obtidos neste trabalho ajudem a esclarecer um pouco mais sobre a complexa biologia da espécie não patogênica, quando utilizada como modelo comparativo em estudos com outras espécies de tripanosomatídeos. Considerando as regiões de repetições simples identificadas na versão 2, é possível que essas ajudem a elucidar como ocorrem os mecanismos de recombinação para geração e manutenção da variabilidade antigênica do *T. rangeli*.

Na versão 2 do genoma, também foi possível identificar todas as unidades de rRNA e tRNA descritas em outros tripanosomatídeos, assim como um maior

número de CDS do que havia sido descrito no genoma de referência desta cepa. Por meio da utilização de bancos de dados modernos, a porção de CDS que foram anotadas automaticamente também é maior, quando comparado ao conjunto de CDS descritas em outras espécies como *T. brucei*, *T. cruzi* e *T. rangeli*. Mesmo para as CDS que foram anotadas como hipotéticas, foi possível detectar um subconjunto dessas que apresentam evidências de sua expressão e tradução, por meio de análises que utilizaram dados transcriptômicos e proteômicos da espécie, gerados previamente pelo grupo de pesquisa do Laboratório de Protozoologia. Este achado retoma um debate importante sobre a classificação de CDS como “hipotéticas” ou como “de função desconhecida”, uma vez que tanto experimentos de bancada quanto ferramentas computacionais apontam para a existência de um conjunto de proteínas que ainda não foram caracterizadas em tripanosomatídeos. Em alguns casos, CDS anotadas como hipotéticas podem ser desconsideradas durante etapas de análise, então é possível que a mudança de terminologia auxilie na melhor exploração do conjunto de proteínas destes organismos.

Ainda, mesmo utilizando esta nova versão do genoma, foi possível confirmar a existência de um genoma central compartilhado entre tripanosomatídeos. Dentre este resultado, destaca-se um número baixo de *singletons* identificados no genoma de *T. rangeli*, cuja espécie apresenta fragmentos, pseudogenes ou CDS que fazem parte da maquinaria de iRNA. Comparativamente, foram identificados mais *singletons* no conjunto de dados de espécies que apresentam genomas maiores e mais repetitivos, como o *T. cruzi*. Por fim, é importante destacar que foram identificadas CDS anotadas como hipotéticas em *T. rangeli*, que fazem parte de grupos ortólogos com os outros tripanosomatídeos analisados. Estas CDS podem ser alvos promissores para estudos futuros, principalmente os de cunho comparativo.

REFERÊNCIAS

ACOSTA-SERRANO, A.; ALMEIDA, I.C.; FREITAS-JUNIOR, L. H.; YOSHIDA, N.; SCHENKMAN, S. The mucin-like glycoprotein super-family of *Trypanosoma cruzi*: structure and biological roles. **Molecular And Biochemical Parasitology**, v. 114, n. 2, p. 143-150, mai. 2001. Elsevier BV. http://dx.doi.org/10.1016/s0166-6851(01)00245-6.

AFCHAIN, D.; RAY, D.L.; FRUIT, J.; CAPRON, A. Antigenic Make-Up of *Trypanosoma cruzi* Culture Forms: identification of a specific component. **The Journal Of Parasitology**, v. 65, n. 4, p. 507-515, ago. 1979. JSTOR. http://dx.doi.org/10.2307/3280312.

ALTENHOFF, A.M.; DESSIMOZ, C. Infering orthology and paralogy. In: ANISIMOVA, M. (Ed.). **Evolutionary Genomics**, v. 855, p. 259-279, 2012. Totowa, NJ: Humana Press. http://dx.doi.org/10.1007/978-1-61779-582-4_9.

ALVAR, J.; VÉLEZ, I.D.; BERN, C.; HERRERO, M.; DESJEUX, P.; CANO, J.; JANNIN, J.; BOER, M.D. Leishmaniasis Worldwide and Global Estimates of Its Incidence. **Plos One**, v. 7, n. 5, 31 mai. 2012. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0035671.

AMOS, B.; AURRECOECHEA, C.; BARBA, M.; BARRETO, A.; BASENKO, E.Y.; BAŜANT, W.; BELNAP, R.; BLEVINS, A.; BÖHME, U.; BRESTELLI, J. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. **Nucleic Acids Research**, v. 50, n. 1, p. 898-911, 28 out. 2021. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkab929.

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

AÑEZ-ROJAS, N.; PERALTA, A.; CRISANTE, G.; ROJAS, A.; AÑEZ, N.; RAMÍREZ, J.L.; CHIURILLO, M.A. *Trypanosoma rangeli* expresses a gene of the group II trans-sialidase superfamily. **Molecular And Biochemical Parasitology**, v. 142, n. 1, p. 133-136, jul. 2005. Elsevier BV. http://dx.doi.org/10.1016/j.molbiopara.2005.03.012.

ANTINORI, S.; GALIMBERTI, L.; BIANCO, R.; GRANDE, R.; GALLI, M.; CORBELLINO, M. Chagas disease in Europe: a review for the internist in the globalized world. **European Journal Of Internal Medicine**, v. 43, p. 6-15, set. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.ejim.2017.05.001.

ARAÚJO, P.R; TEIXEIRA, S.M. Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in *Trypanosoma cruzi*: a review. **Memórias do Instituto Oswaldo Cruz**, v. 106, n. 3, p. 257-266, mai. 2011. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0074-02762011000300002.

BAPTISTA, R.P.; KISSINGER, J.C. Is reliance on an inaccurate genome sequence sabotaging your experiments? **Plos Pathogens**, v. 15, n. 9, 12 set. 2019. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.ppat.1007901.

BARRIAS, E.S.; REIGNAULT, L.C.; SOUZA, W.; CARVALHO, T.M.U. *Trypanosoma cruzi* uses macropinocytosis as an additional entry pathway into mammalian host cell. **Microbes And Infection**, v. 14, n. 14, p. 1340-1351, nov. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.micinf.2012.08.003.

BASILE, L; JANSÁ, J.M.; CARLIER, Y.; SALAMANCA, D.D.; ANGHEBEN, A.; BARTOLONI, A.; SEIXAS, J.; VAN GOOL, T.; CAÑAVATE, C.; FLORES-CHÁVEZ, M. Chagas disease in European countries: the challenge of a surveillance system. **Eurosurveillance**, v. 16, n. 37, 15 set. 2011. European Centre for Disease Control and Prevention (ECDC). http://dx.doi.org/10.2807/ese.16.37.19968-en.

BELTON, J.M.; MCCORD, R.P.; GIBCUS, J.H.; NAUMOVA, N.; ZHAN, Y.; DEKKER, J. Hi–C: a comprehensive technique to capture the conformation of genomes. **Methods**, v. 58, n. 3, p. 268-276, nov. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.ymeth.2012.05.001.

BERNÁ, L.; RODRIGUEZ, M.; CHIRIBAO, M.L.; PARODI-TALICE, A.; PITA, S.; RIJO, G.; ALVAREZ-VALIN, F.; ROBELLO, C. Expanding an expanded genome: long-read sequencing of *Trypanosoma cruzi*. **Microbial Genomics**, v. 4, n. 5, 1 mai. 2018. Microbiology Society. http://dx.doi.org/10.1099/mgen.0.000177.

BERRIMAN, M.; GHEDIN, E.; HERTZ-FOWLER, C.; BLANDIN, G.; RENAULD, H.; BARTHOLOMEU, D.C.; LENNARD, N.J.; CALER, E.; HAMLIN, N.E.; HAAS, B. The Genome of the African Trypanosome *Trypanosoma brucei*. **Science**, v. 309, n. 5733, p. 416-422, 15 jul. 2005. American Association for the Advancement of Science (AAAS). http://dx.doi.org/10.1126/science.1112642.

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

BOETZER, M.; PIROVANO, W. Toward almost closed genomes with GapFiller. **Genome Biology**, v. 13, n. 6, p. 56, 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/gb-2012-13-6-r56.

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

BRADWELL, K.R; KOPARDE, V.N.; MATVEYEV, A.V.; SERRANO, M.G.; ALVES, J.M.P.; PARIKH, H.; HUANG, B.; LEE, V.; ESPINOSA-ALVAREZ, O.; ORTIZ, A.P., et al. Genomic comparison of *Trypanosoma conorhini* and *Trypanosoma rangeli* to *Trypanosoma cruzi* strains of high and low virulence. **Bmc Genomics**, v. 19, n. 1, p. 1-2, 24 out. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s12864-018-5112-0.

BRASIL; Ministério da Saúde. Boletim epidemiológico – Doença de Chagas, Ano 2, Número Especial. **Secretaria de Vigilância em Saúde**, p. 1-38, abr. 2021. Disponível em: <https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/boletins-epidemiologicos/especiais/2021/boletim_especial_chagas_14abr21_b.pdf>.

BRAY, N.L.; PIMENTEL, H.; MELSTED, P.; PACHTER, L. Near-optimal probabilistic RNA-seq quantification. **Nature Biotechnology**, v. 34, n. 5, p. 525-527, 4 abr. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nbt.3519.

BROWNE, P.D.; NIELSEN, T.K.; KOT, W.; AGGERHOLM, A.; GILBERT, M.T.P.; PUETZ, L.; RASMUSSEN, M.; ZERVAS, A.; HANSEN, L.H. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. **Gigascience**, v. 9, n. 2, 1 fev. 2020. Oxford University Press (OUP). http://dx.doi.org/10.1093/gigascience/giaa008.

BRUCE, D. Classification of the African Trypanosomes pathogenic to man and domestic animals. **Transactions Of The Royal Society Of Tropical Medicine And Hygiene**, v. 8, n. 1, p. 1-22, nov. 1914. Oxford University Press (OUP). http://dx.doi.org/10.1016/s0035-9203(14)90016-5.

BUSCAGLIA, C.A.; CAMPO, V.A.; FRASCH, A.C.C.; NOIA, J.M. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. **Nature Reviews Microbiology**, v. 4, n. 3, p. 229-236, mar. 2006. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrmicro1351.

CALLEJAS-HERNÁNDEZ, F.; GUTIERREZ-NOGUES, Á.; RASTROJO, A.; GIRONÈS, N.; FRESNO, M. Analysis of mRNA processing at whole transcriptome level, transcriptomic profile and genome sequence refinement of *Trypanosoma cruzi*. **Scientific Reports**, v. 9, n. 1, 22 nov. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-019-53924-6.

CAMACHO, C.; COULOURIS, G.; AVAGYAN, V.; MA, N.; PAPADOPOULOS, J.; BEALER, K.; MADDEN, T.L. BLAST+: architecture and applications. **Bmc Bioinformatics**, v. 10, n. 1, dez. 2009. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2105-10-421.

CARDOSO, M.S.; REIS-CUNHA, J.L.; BARTHOLOMEU, D.C. Evasion of the Immune Response by *Trypanosoma cruzi* during Acute Infection. **Frontiers In Immunology**, v. 6, 18 jan. 2016. Frontiers Media SA. http://dx.doi.org/10.3389/fimmu.2015.00659.

CDC; Centers for Disease Control and Prevention. The Global Fight Against Chagas Disease. **U.S. Department of Health and Human Services**, p. 1, 2 out. 2017. Disponível em: <https://www.cdc.gov/globalhealth/infographics/malaria-parasitic-diseases/the-global-fight-against-chagas-disease.htm>.

CHAGAS, C. Neue Trypanosomen: Vorläufige mitteilung. **Archiv für Schiffs- und Tropen-Hygiene**, Leipzig, n.13, p.120-122. 1909a.

CHARLESWORTH, B.; SNIEGOWSKI, P.; STEPHAN, W. The evolutionary dynamics of repetitive DNA in eukaryotes. **Nature**, v. 371, n. 6494, p. 215-220, 15 set. 1994. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/371215a0.

CHIURILLO, M.A.; CORTEZ, D.R.; LIMA, F.M.; CORTEZ, C.; RAMÍREZ, J.L.; MARTINS, A.G.; SERRANO, M.G.; TEIXEIRA, M.M.G.; SILVEIRA, J.F. The diversity and expansion of the trans-sialidase gene family is a common feature in *Trypanosoma cruzi* clade members. **Infection, Genetics And Evolution**, v. 37, p. 266-274, jan. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.meegid.2015.11.024.

CLAESSENS, A.; HAMILTON, W.L.; KEKRE, M.; OTTO, T.D.; FAIZULLABHOY, A.; RAYNER, J.C.; KWIATKOWSKI, D. Generation of Antigenic Diversity in *Plasmodium falciparum* by Structured Rearrangement of Var Genes During Mitosis. **Plos Genetics**, v. 10, n. 12, 18 dez. 2014. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pgen.1004812.

COSTA, R.W.; SILVEIRA, J.F.; BAHIA, D. Interactions between *Trypanosoma cruzi* Secreted Proteins and Host Cell Signaling Pathways. **Frontiers In Microbiology**, v. 7, 31 mar. 2016. Frontiers Media SA. http://dx.doi.org/10.3389/fmicb.2016.00388.

COSTA, R.W.; BATISTA, M.F.; MENEGHELLI, I.; VIDAL, R.O.; NÁJERA, C.A.; MENDES, A.C.; ANDRADE-LIMA, I.A.; SILVEIRA, J.F.; LOPES, L.R.; FERREIRA, L.R.P. Comparative Analysis of the Secretome and Interactome of *Trypanosoma cruzi* and *Trypanosoma rangeli* Reveals Species Specific Immune Response Modulating Proteins. **Frontiers In Immunology**, v. 11, 27 ago. 2020. Frontiers Media SA. http://dx.doi.org/10.3389/fimmu.2020.01774.

CUBA, C.A.C. Revisión de los aspectos biológicos y diagnósticos del *Trypanosoma* (*Herpetosoma*) *rangeli*. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 31, n. 2, p. 207-220, abr. 1998. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0037-86821998000200007.

CUERVO, C.; LÓPEZ, M.C.; PUERTA, C. The *Trypanosoma rangeli* histone H2A gene sequence serves as a differential marker for KP1 strains. **Infection, Genetics And Evolution**, v. 6, n. 5, p. 401-409, set. 2006. Elsevier BV. http://dx.doi.org/10.1016/j.meegid.2006.01.005.

D'ALESSANDRO, A. Biology of *Trypanosoma* (*Herpetosoma*) *rangeli* Tejera, 1920. In: LUMSDEN, W.H.R.; EVANS, D.A. (Ed.). **Biology of the kinetoplastida**. London: London Academic, v. 3, p. 327-403. 1976.

DANECEK, P.; BONFIELD, J.K; LIDDLE, J.; MARSHALL, J.; OHAN, V.; POLLARD, M.O.; WHITWHAM, A.; KEANE, T.; A MCCARTHY, S.; DAVIES, R.M. Twelve years of SAMtools and BCFtools. **Gigascience**, v. 10, n. 2, 29 jan. 2021. Oxford University Press (OUP). http://dx.doi.org/10.1093/gigascience/giab008.

DE SOUZA, W.; DE CARVALHO, T.M.U.; BARRIAS, E.S. Review on *Trypanosoma cruzi*: host cell interaction. **International Journal Of Cell Biology**, v. 2010, p. 1-18, 2010. Hindawi Limited. http://dx.doi.org/10.1155/2010/295394.

DÍAZ-VIRAQUÉ, F.; CHIRIBAO, M.L.; LIBISCH, M.G.; ROBELLO, C. Genome-wide chromatin interaction map for *Trypanosoma cruzi*. **Nature Microbiology**, 12 out. 2023. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41564-023-01483-y.

DIDA, F.; YI, G. Empirical evaluation of methods for *de novo* genome assembly. **Peerj Computer Science**, v. 7, 9 jul. 2021. PeerJ. http://dx.doi.org/10.7717/peerj-cs.636.

DO CARMO, R.A. **Análise genômica comparativa entre tripanossomas do clado *Trypanosoma cruzi* isolados na América do Sul: aplicação da hibridização genômica comparativa em arranjo de DNA (aCGH)**. 2022. Tese de Doutorado (Programa de Pós-Graduação em Imunologia e Microbiologia). Universidade Federal de São Paulo, São Paulo.

DOLLET, M. Plant Diseased Caused by Flagellate Protozoa (*Phytomonas*). **Annual Review of Phytopathology**, v. 22, n. 1, p. 115-132, set. 1984.

DRILLON, G.; CARBONE, A.; FISCHER, G. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. **Plos One**, v. 9, n. 3, 20 mar. 2014. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0092621.

EKBLOM, R.; WOLF, J.B.W. A field guide to whole-genome sequencing, assembly and annotation. **Evolutionary Applications**, v. 7, n. 9, p. 1026-1042, 24 jun. 2014. Wiley. http://dx.doi.org/10.1111/eva.12178.

EL-SAYED, N.M.; MYLER, P.J.; BARTHOLOMEU, D.C.; NILSSON, D.; AGGARWAL, G.; TRAN, A.; GHEDIN, E.; WORTHEY, E.A.; DELCHER, A.L.; BLANDIN, G., et al. The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease. **Science**, v. 309, n. 5733, p. 409-415, 15 jul. 2005a. American Association for the Advancement of Science (AAAS). http://dx.doi.org/10.1126/science.1112631.

EL-SAYED, N.M.; MYLER, P.J.; BLANDIN, G.; BERRIMAN, M.; CRABTREE, J.; AGGARWAL, G.; CALER, E.; RENAULD, H.; WORTHEY, E.A.; HERTZ-FOWLER, C., et al. Comparative Genomics of Trypanosomatid Parasitic Protozoa. **Science**, v. 309, n. 5733, p. 404-409, 15 jul. 2005b. American Association for the Advancement of Science (AAAS). http://dx.doi.org/10.1126/science.1112181.

EMMER, B.T.; NAKAYASU, E.S.; SOUTHER, C.; CHOI, H.; SOBREIRA, T.J. P.; EPTING, C.L.; NESVIZHSKII, A.I.; ALMEIDA, I.C.; ENGMAN, D.M. Global Analysis of Protein Palmitoylation in African Trypanosomes. **Eukaryotic Cell**, v. 10, n. 3, p. 455-463, mar. 2011. American Society for Microbiology. http://dx.doi.org/10.1128/ec.00248-10.

EMMS, D.M.; KELLY, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. **Genome Biology**, v. 20, n. 1, p. 1–14, 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s13059-019-1832-y.

ENG, J.K.; JAHAN, T.A.; HOOPMANN, M.R. Comet: an open-source ms/ms sequence database search tool. **Proteomics**, v. 13, n. 1, p. 22-24, 4 dez. 2012. Wiley. http://dx.doi.org/10.1002/pmic.201200439.

ENGLISH, A.C.; RICHARDS, S.; HAN, Y.; WANG, M.; VEE, V.; QU, J.; QIN, X.; MUZNY, D.M.; REID, J.G.; WORLEY, K.C. Mind the Gap: upgrading genomes with pacific biosciences rs long-read sequencing technology. **Plos One**, v. 7, n. 11, p. 47768, 21 nov. 2012. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0047768.

ESPINOSA-ÁLVAREZ, O.; ORTIZ, P.A.; LIMA, L.; COSTA-MARTINS, A.G.; SERRANO, M.G.; HERDER, S.; BUCK, G.A.; CAMARGO, E.P.; HAMILTON, P.B.; STEVENS, J.R. *Trypanosoma rangeli* is phylogenetically closer to Old World trypanosomes than to *Trypanosoma cruzi*. **International Journal For Parasitology**, v. 48, n. 7, p. 569-584, jun. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.ijpara.2017.12.008.

FINN, R.D.; CLEMENTS, J.; EDDY, S.R. HMMER web server: interactive sequence similarity searching. **Nucleic Acids Research**, v. 39, n. 2, p. 29–37, 2011. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkr367.

FRANZÉN, O.; OCHAYA, S.; SHERWOOD, E.; LEWIS, M.D.; LLEWELLYN, M.S.; MILES, M.A.; ANDERSSON, B. Shotgun Sequencing Analysis of *Trypanosoma cruzi* I Sylvio X10/1 and Comparison

with *T. cruzi* VI CL Brener. **Plos Neglected Tropical Diseases**, v. 5, n. 3, 8 mar. 2011. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pntd.0000984.

FREIRE, E.R.; MALVEZZI, A.M.; VASHISHT, A.A.; ZUBEREK, J.; SAADA, E.A.; LANGOUSIS, G.; NASCIMENTO, J.D.F.; MOURA, D.; DARZYNKIEWICZ, E.; HILL, K. *Trypanosoma brucei* Translation Initiation Factor Homolog EIF4E6 Forms a Tripartite Cytosolic Complex with EIF4G5 and a Capping Enzyme Homolog. **Eukaryotic Cell**, v. 13, n. 7, p. 896-908, jul. 2014. American Society for Microbiology. http://dx.doi.org/10.1128/ec.00071-14.

GOODWIN, S.; MCPHERSON, J.D.; MCCOMBIE, W.R. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333-351, 17 mai. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrg.2016.49.

GOUDEY, B.; GEARD, N.; VERSPOOR, K.; ZOBEL, J. Propagation, detection and correction of errors using the sequence database network. **Briefings In Bioinformatics**, v. 23, n. 6, 20 out. 2022. Oxford University Press (OUP). http://dx.doi.org/10.1093/bib/bbac416.

GRISARD, E.C.; STEINDEL, M.; A GUARNERI, A.; EGER-MANGRICH, I.; A CAMPBELL, D.; ROMANHA, A.J. Characterization of *Trypanosoma rangeli* Strains Isolated in Central and South America: an overview. **Memórias do Instituto Oswaldo Cruz**, v. 94, n. 2, p. 203-209, mar. 1999. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0074-02761999000200015.

GRISARD, E.C.; STOCO, P.H.; WAGNER, G.; SINCERO, T.C.M.; ROTAVA, G.; RODRIGUES, J.B.; SNOEIJER, C.Q.; KOERICH, L.B.; SPERANDIO, M.M.; BAYER-SANTOS, E. Transcriptomic analyses of the avirulent protozoan parasite *Trypanosoma rangeli*. **Molecular And Biochemical Parasitology**, v. 174, n. 1, p. 18-25, nov. 2010. Elsevier BV. http://dx.doi.org/10.1016/j.molbiopara.2010.06.008.

GRISARD, E.C.; ROMANHA, A.J.; STEINDEL, M. *Trypanosoma* (*Herpetosoma*) *rangeli*. In: NEVES, D.P.; MELO A.L.; LINARDI, P.M.; VITOR, R.W.A. (Org.). **Parasitologia Humana**. 13ed. Rio de Janeiro, RJ: Atheneu, 2016, p. 115-120.

GUHL, F.; VALLEJO, G.A. *Trypanosoma* (*Herpetosoma*) *rangeli* Tejera, 1920: an updated review. **Memórias do Instituto Oswaldo Cruz**, v. 98, n. 4, p. 435-442, jun. 2003. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0074-02762003000400001.

GUREVICH, A.; SAVELIEV, V.; VYAHHI, N.; TESLER, G. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072-1075, 19 fev. 2013. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btt086.

HADRYS, H.; BALICK, M.; SCHIERWATER, B. Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. **Molecular Ecology**, v. 1, n. 1, p. 55-63, mai. 1992. Wiley. http://dx.doi.org/10.1111/j.1365-294x.1992.tb00155.x.

HALL, J.P.J.; WANG, H.; BARRY, J.D. Mosaic VSGs and the Scale of *Trypanosoma brucei* Antigenic Variation. **Plos Pathogens**, v. 9, n. 7, 11 jul. 2013. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.ppat.1003502.

HERBIG-SANDREUTER, A. Further studies on *Trypanosoma rangeli* Tejera 1920. **Acta tropica**, v. 14, n. 3, p. 193–207, 1957. http://dx.doi.org/10.5169/seals-310679.

HERREROS-CABELLO, A.; CALLEJAS-HERNÁNDEZ, F.; GIRONÈS, N.; FRESNO, M. *Trypanosoma cruzi* Genome: organization, multi-gene families, transcription, and biological implications. **Genes**, v. 11, n. 10, p. 1196-1221, 14 out. 2020. MDPI AG. http://dx.doi.org/10.3390/genes11101196.

HOARE, C.A.; WALLACE, F.G. Developmental Stages of Trypanosomatid Flagellates: a new terminology. **Nature**, v. 212, n. 5068, p. 1385-1386, dez. 1966. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/2121385a0.

HOFF, K.J.; STANKE, M. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. **Nucleic Acids Research**, v. 41, n. 1, p. 123–128, 2013. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkt418.

HORN, D.; SPENCE, C.; INGRAM, A.K. Telomere maintenance and length regulation in *Trypanosoma brucei*. **The Embo Journal**, v. 19, n. 10, p. 2332-2339, 15 mai. 2000. Wiley. http://dx.doi.org/10.1093/emboj/19.10.2332.

IJAQ, J.; CHANDRASEKHARAN, M.; PODDAR, R.; BETHI, N.; SUNDARARAJAN, V.S. Annotation and curation of uncharacterized proteins- challenges. **Frontiers In Genetics**, v. 6, 31 mar. 2015. Frontiers Media SA. http://dx.doi.org/10.3389/fgene.2015.00119.

ILLUMINA. **HiSeq 2500 System Guide**. 2020. Disponível em: <https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseq2500/hiseq-2500-system-guide-15035786-03.pdf>. Acesso em: 26 mai. 2021.

IVENS, A.C.; PEACOCK, C.S.; WORTHEY, E.A.; MURPHY, L.; AGGARWAL, G.; BERRIMAN, M.; SISK, E.; RAJANDREAM, M.A.; ADLEM, E.; AERT, R. The Genome of the Kinetoplastid Parasite, *Leishmania major*. **Science**, v. 309, n. 5733, p. 436-442, 15 jul. 2005. American Association for the Advancement of Science (AAAS). http://dx.doi.org/10.1126/science.1112680.

JAYAKUMAR, V.; SAKAKIBARA, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. **Briefings In Bioinformatics**, v. 20, n. 3, p. 866-876, 3 mai. 2019. Oxford University Press (OUP). http://dx.doi.org/10.1093/bib/bbx147.

JONES, P.; BINNS, D.; CHANG, H.Y; FRASER, M.; LI, W.; MCANULLA, C.; MCWILLIAM, H.; MASLEN, J.; MITCHEL, A.; NUKA, G.; PESSEAT, S.; QUINN, A.F.; VEGAS-SANGRADOR, A.; SCHEREMETJEW, M.; YONG, S.Y.; LOPEZ, R.; HUNTER, S. InterProScan 5: genome-scale protein function classification. **Bioinformatics**, v. 30, n. 9, p. 1236–1240, 2014. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btu031.

KALVARI, I.; ARGASINSKA, J.; QUINONES-OLVERA, N.; NAWROCKI, E.P.; RIVAS, E.; EDDY, S.R.; BATEMAN, A.; FINN, R.D.; PETROV, A.I. Rfam 13.0: shifting to a genome-centric resource for non-coding rna families. **Nucleic Acids Research**, v. 46, n. 1, p. 335-342, 3 nov. 2017. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkx1038.

KANEHISA, M.; SATO, Y.; MORISHIMA, K. BlastKOALA and GhostKOALA: kegg tools for functional characterization of genome and metagenome sequences. **Journal Of Molecular Biology**, v. 428, n. 4, p. 726-731, fev. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.jmb.2015.11.006.

KARP, P.D. What we do not know about sequence analysis and sequence databases. **Bioinformatics**, v. 14, n. 9, p. 753-754, 1 jan. 1998. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/14.9.753.

KAUFER, A.; ELLIS, J.; STARK, D.; BARRATT, J. The evolution of trypanosomatid taxonomy. **Parasites & Vectors**, v. 10, n. 1, 8 jun. 2017. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s13071-017-2204-7.

KOLLMAN, J.M.; POLKA, J.K.; ZELTER, A.; DAVIS, T.N.; AGARD, D.A. Microtubule nucleating γ-TuSC assembles structures with 13-fold microtubule-like symmetry. **Nature**, v. 466, n. 7308, p. 879-882, 14 jul. 2010. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nature09207.

KOLLMAN, J.M.; MERDES, A.; MOUREY, L.; AGARD, D.A. Microtubule nucleation by γ-tubulin complexes. **Nature Reviews Molecular Cell Biology**, v. 12, n. 11, p. 709-721, 12 out. 2011. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrm3209.

KOREN, S.; WALENZ, B.P.; BERLIN, K.; MILLER, J.R.; BERGMAN, N.H.; PHILLIPPY, A.M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. **Genome Research**, v. 27, n. 5, p. 722-736, 15 mar. 2017. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/gr.215087.116.

KORF, I. Gene finding in novel genomes. **Bmc Bioinformatics**, v. 5, n. 1, p. 59-68, 2004. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2105-5-59.

LANGMEAD, B.; SALZBERG, S.L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357-359, 4 mar. 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nmeth.1923.

LEE, B.; BACON, K.M.; BOTTAZZI, M.E.; HOTEZ, P.J. Global economic burden of Chagas disease: a computational simulation model. **The Lancet Infectious Diseases**, v. 13, n. 4, p. 342-348, abr. 2013. Elsevier BV. http://dx.doi.org/10.1016/s1473-3099(13)70002-1.

LEIPE, D.D.; KOONIN, E.V.; ARAVIND, L. STAND, a Class of P-Loop NTPases Including Animal and Plant Regulators of Programmed Cell Death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. **Journal Of Molecular Biology**, v. 343, n. 1, p. 1-28, out. 2004. Elsevier BV. http://dx.doi.org/10.1016/j.jmb.2004.08.023.

LI, B. Keeping Balance Between Genetic Stability and Plasticity at the Telomere and Subtelomere of *Trypanosoma brucei*. **Frontiers In Cell And Developmental Biology**, v. 9, 5 jul. 2021. Frontiers Media SA. http://dx.doi.org/10.3389/fcell.2021.699639.

LI, H. Minimap2: pairwise alignment for nucleotide sequences. **Bioinformatics**, v. 34, n. 18, p. 3094-3100, 10 mai. 2018. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/bty191.

LIAO, X.; HU, K.; SALHI, A.; ZOU, Y.; WANG, J.; GAO, X. MsRepDB: a comprehensive repetitive sequence database of over 80 000 species. **Nucleic Acids Research**, v. 50, n. 1, p. 236-245, 1 dez. 2021. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkab1089.

LIDANI, K.C.F.; ANDRADE, F.A.; BAVIA, L.; DAMASCENO, F.S.; BELTRAME, M.H.; MESSIAS-REASON, I.J.; SANDRI, T.L. Chagas Disease: from discovery to a worldwide health problem. **Frontiers In Public Health**, v. 7, 2 jul. 2019. Frontiers Media SA. http://dx.doi.org/10.3389/fpubh.2019.00166.

LIMA, F.M.; SOUZA, R.T.; SANTORI, F.R.; SANTOS, M.F.; CORTEZ, D.R.; BARROS, R.M.; CANO, M.I.; VALADARES, H.M.S.; MACEDO, A.M.; MORTARA, R.A. Interclonal Variations in the Molecular Karyotype of *Trypanosoma cruzi*: chromosome rearrangements in a single cell-derived clone of the G strain. **Plos One**, v. 8, n. 5, p. 1-2, 7 mai. 2013. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0063738.

LOGSDON, G.A.; VOLLGER, M.R.; EICHLER, E.E. Long-read human genome sequencing and its applications. **Nature Reviews Genetics**, v. 21, n. 10, p. 597-614, 5 jun. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41576-020-0236-x.

LÜCKEMEYER, D.D. **Avaliação do perfil proteico de *Trypanosoma rangeli* durante o processo de diferenciação celular *in vitro***. 2014. Tese de Doutorado (Programa de Pós-Graduação em Biotecnologia e Biociências). Universidade Federal de Santa Catarina, Florianópolis.

LUKEŠ, J.; SKALICKÝ, T.; TÝČ, J.; VOTÝPKA, J.; YURCHENKO, V. Evolution of parasitism in kinetoplastid flagellates. **Molecular And Biochemical Parasitology**, v. 195, n. 2, p. 115-122, jul. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.molbiopara.2014.05.007.

MAIA DA SILVA, F.; NOYES, H.; CAMPANER, M.; JUNQUEIRA, A.C.V.; COURA, J.R.; AÑEZ, N.; SHAW, J.J.; STEVENS, J.R.; TEIXEIRA, M.M.G. Phylogeny, taxonomy and grouping of *Trypanosoma rangeli* isolates from man, triatomines and sylvatic mammals from widespread geographical origin based on SSU and ITS ribosomal sequences. **Parasitology**, v. 129, n. 5, p. 549-561, 5 out. 2004. Cambridge University Press (CUP). http://dx.doi.org/10.1017/s0031182004005931.

MAIA DA SILVA, F.; JUNQUEIRA, A.C.V.; CAMPANER, M.; RODRIGUES, A.C.; CRISANTE, G.; RAMIREZ, L.E.; CABALLERO, Z.C.E.; MONTEIRO, F.A.; COURA, J.R.; AÑEZ, N. Comparative phylogeography of *Trypanosoma rangeli* and *Rhodnius* (Hemiptera: reduviidae) supports a long coexistence of parasite lineages and their sympatric vectors. **Molecular Ecology**, v. 16, n. 16, p. 3361-3373, ago. 2007. Wiley. http://dx.doi.org/10.1111/j.1365-294x.2007.03371.x.

MAIA DA SILVA, F.; MARCILI, A.; LIMA, L.; CAVAZZANA, M.; ORTIZ, P.A.; CAMPANER, M.; TAKEDA, G.F.; PAIVA, F.; NUNES, V.L.B.; CAMARGO, E.P. *Trypanosoma rangeli* isolates of bats from Central Brazil: genotyping and phylogenetic analysis enable description of a new lineage using spliced-leader gene sequences. **Acta Tropica**, v. 109, n. 3, p. 199-207, mar. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.actatropica.2008.11.005.

MAIA, G.A.; BENETTI FILHO, V.; KAWAGOE, E.K.; SORATTO, T.A.T.; MOREIRA, R.S.; GRISARD, E.C.; WAGNER, G. AnnotaPipeline: an integrated tool to annotate eukaryotic proteins using multi-omics data. **Frontiers In Genetics**, v. 13, 22 nov. 2022. Frontiers Media SA. http://dx.doi.org/10.3389/fgene.2022.1020100.

MANNI, M.; BERKELEY, M.R.; SEPPEY, M.; ZDOBNOV, E.M. BUSCO: assessing genomic data quality and beyond. **Current Protocols**, v. 1, n. 12, dez. 2021. Wiley. http://dx.doi.org/10.1002/cpz1.323.

MARÇAIS, G.; DELCHER, A.L.; PHILLIPPY, A.M.; COSTON, R.; SALZBERG, S.L.; ZIMIN, A. MUMmer4: a fast and versatile genome alignment system. **Plos Computational Biology**, v. 14, n. 1, p. 1-2, 26 jan. 2018. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pcbi.1005944.

MASLOV, D.A.; OPPERDOES, F.R.; KOSTYGOV, A.Y.; HASHIMI, H.; LUKEŁ, J.; YURCHENKO, V. Recent advances in trypanosomatid research: genome organization, expression, metabolism, taxonomy and evolution. **Parasitology**, v. 146, n. 1, p. 1-27, 14 jun. 2018. Cambridge University Press (CUP). http://dx.doi.org/10.1017/s0031182018000951.

MCKEAN, P.G.; BAINES, A.; VAUGHAN, S.; GULL, K. γ-Tubulin Functions in the Nucleation of a Discrete Subset of Microtubules in the Eukaryotic Flagellum. **Current Biology**, v. 13, n. 7, p. 598-602, abr. 2003. Elsevier BV. http://dx.doi.org/10.1016/s0960-9822(03)00174-x.

MEDONE, P.; CECCARELLI, S.; PARHAM, P.E.; FIGUERA, A.; RABINOVICH, J.E. The impact of climate change on the geographical distribution of two vectors of Chagas disease: implications for the force of infection. **Philosophical Transactions Of The Royal Society B**: Biological Sciences, v. 370, n. 1665, 5 abr. 2015. The Royal Society. http://dx.doi.org/10.1098/rstb.2013.0560.

MELLO, C.B.; GARCIA, E.S.; RATCLIFFE, N.A.; AZAMBUJA, P. *Trypanosoma cruzi* and *Trypanosoma rangeli*: interplay with hemolymph components of *Rhodnius prolixus*. **Journal Of Invertebrate Pathology**, v. 65, n. 3, p. 261-268, mai. 1995. Elsevier BV. http://dx.doi.org/10.1006/jipa.1995.1040.

MOHAN, R. Computational structural and functional analysis of hypothetical proteins of *Staphylococcus aureus*. **Bioinformation**, v. 8, n. 15, p. 722-728, 3 ago. 2012. Biomedical Informatics. http://dx.doi.org/10.6026/97320630008722.

MORAES, M.H.; A GUARNERI, A.; GIRARDI, F.P.; RODRIGUES, J.B.; EGER, I.; TYLER, K.M.; STEINDEL, M.; GRISARD, E.C. Different serological cross-reactivity of *Trypanosoma rangeli* forms in *Trypanosoma cruzi*-infected patients sera. **Parasites & Vectors**, v. 1, n. 1, 8 jul. 2008. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1756-3305-1-20.

MOREL, C.M.; ACHARYA, T.; BROUN, D.; DANGI, A.; ELIAS, C.; GANGULY, N.K.; GARDNER, C.A.; GUPTA, R.K.; HAYCOCK, J.; HEHER, A.D., et al. Health Innovation Networks to Help Developing Countries Address Neglected Diseases. **Science**, v. 309, n. 5733, p. 401-404, 15 jul. 2005. American Association for the Advancement of Science (AAAS). http://dx.doi.org/10.1126/science.1115538.

MORIYA, Y.; ITOH, M.; OKUDA, S.; YOSHIZAWA, A.C.; KANEHISA, M. KAAS: an automatic genome annotation and pathway reconstruction server. **Nucleic Acids Research**, v. 35, p. 182-185, 8 mai. 2007. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkm321.

MUCCI, J.; HIDALGO, A.; MOCETTI, E.; ARGIBAY, P.F.; LEGUIZAMÓN, M.S.; CAMPETELLA, O. Thymocyte depletion in *Trypanosoma cruzi* infection is mediated by trans -sialidase-induced apoptosis on nurse cells complex. **Proceedings Of The National Academy Of Sciences**, v. 99, n. 6, p. 3896-3901, 12 mar. 2002. Proceedings of the National Academy of Sciences. http://dx.doi.org/10.1073/pnas.052496399.

MUCCI, J.; LANTOS, A.B.; BUSCAGLIA, C.A.; LEGUIZAMÓN, M.S.; CAMPETELLA, O. The *Trypanosoma cruzi* Surface, a Nanoscale Patchwork Quilt. **Trends In Parasitology**, v. 33, n. 2, p. 102-112, fev. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.pt.2016.10.004.

MÜLLER, L.S. M.; COSENTINO, R.O.; FÖRSTNER, K.U.; GUIZETTI, J.; WEDEL, C.; KAPLAN, N.; JANZEN, C.J.; ARAMPATZI, P.; VOGEL, J.; STEINBISS, S. Genome organization and DNA accessibility control antigenic variation in trypanosomes. **Nature**, v. 563, n. 7729, p. 121-125, 17 out. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41586-018-0619-8.

NAKAYASU, E.S.; YASHUNSKY, D.V.; NOHARA, L.L.; TORRECILHAS, A.C.T.; NIKOLAEV, A.V.; ALMEIDA, I.C. GPIomics: global analysis of glycosylphosphatidylinositol anchored molecules of *Trypanosoma cruzi*. **Molecular Systems Biology**, v. 5, n. 1, jan. 2009. EMBO. http://dx.doi.org/10.1038/msb.2009.13.

NAWROCKI, E. P.; EDDY, S. R. Infernal 1.1: 100-fold faster rna homology searches. **Bioinformatics**, v. 29, n. 22, p. 2933-2935, 4 set. 2013. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btt509.

NGÔ, H.; TSCHUDI, C.; GULL, K.; ULLU, E. Double-stranded RNA induces mRNA degradation in *Trypanosoma brucei*. **Proceedings Of The National Academy Of Sciences**, v. 95, n. 25, p. 14687-14692, 8 dez. 1998. Proceedings of the National Academy of Sciences. http://dx.doi.org/10.1073/pnas.95.25.14687.

NIH; National Institutes of Health. Comparative Genomics Fact Sheet. **National Human Genome Research Institute**, 15 ago. 2020. Disponível em: <https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet>.

OBADO, S.O.; BOT, C.; NILSSON, D.; ANDERSSON, B.; KELLY, J.M. Repetitive DNA is associated with centromeric domains in *Trypanosoma brucei* but not *Trypanosoma cruzi*. **Genome Biology**, v. 8, n. 3, 2007. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/gb-2007-8-3-r37.

OMS; Organização Mundial da Saúde. **Neglected Tropical Diseases - Overview, Impact and Fact Sheets**. 2022. Disponível em: <https://www.who.int/health-topics/neglected-tropical-diseases>. Acesso em: 13 jul. 2022.

OUZOUNIS, C.A.; KARP, P.D. The past, present and future of genome-wide re-annotation. **Genome Biology**, v. 3, n. 2, 2002. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/gb-2002-3-2-comment2001.

PABLOS, L.M.; OSUNA, A. Multigene Families in *Trypanosoma cruzi* and Their Role in Infectivity. **Infection And Immunity**, v. 80, n. 7, p. 2258-2264, jul. 2012. American Society for Microbiology. http://dx.doi.org/10.1128/iai.06225-11.

PACBIO. **Previous system releases: sequel II system**. 2021. Disponível em: <https://www.pacb.com/products-and-services/sequel-system/previous-system-releases/>. Acesso em: 26 mai. 2021.

PEARSON, W.R. An Introduction to Sequence Similarity ("Homology") Searching. **Current Protocols In Bioinformatics**, v. 42, n. 1, jun. 2013. Wiley. http://dx.doi.org/10.1002/0471250953.bi0301s42.

PECH-CANUL, Á.L.C.; MONTEÓN, V.; SOLÍS-OVIEDO, R.L. A Brief View of the Surface Membrane Proteins from *Trypanosoma cruzi*. **Journal Of Parasitology Research**, v. 2017, p. 1-13, 2017. Hindawi Limited. http://dx.doi.org/10.1155/2017/3751403.

PEREIRA-CHIOCCOLA, V.L.; ACOSTA-SERRANO, A.; ALMEIDA, I.C.; FERGUSON, M.A.J.; SOUTO-PADRON, T.; RODRIGUES, M.M.; TRAVASSOS, L.R.; SCHENKMAN, S. Mucin-like molecules form a negatively charged coat that protects *Trypanosoma cruzi* trypomastigotes from killing by human anti-α-galactosyl antibodies. **Journal Of Cell Science**, v. 113, n. 7, p. 1299-1307, 1 abr. 2000. The Company of Biologists. http://dx.doi.org/10.1242/jcs.113.7.1299.

PEVSNER, J. **Bioinformatics and functional genomics**. 2. ed. John Wiley & Sons, 17 aug. 2015.

PITA, S.; DÍAZ-VIRAQUÉ, F.; IRAOLA, G.; ROBELLO, C. The Tritryps Comparative Repeatome: insights on repetitive element evolution in trypanosomatid pathogens. **Genome Biology And Evolution**, v. 11, n. 2, p. 546-551, 1 fev. 2019. Oxford University Press (OUP). http://dx.doi.org/10.1093/gbe/evz017.

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

PORCEL, B.M.; BONTEMPI, E.J.; HENRIKSSON, J.; RYDÅKER, M.; ÅSLUND, L.; SEGURA, E.L.; PETTERSSON, U.; RUIZ, A.M. *Trypanosoma rangeli* and *Trypanosoma cruzi*: molecular characterization of genes encoding putative calcium-binding proteins, highly conserved in

trypanosomatids. **Experimental Parasitology**, v. 84, n. 3, p. 387-399, dez. 1996. Elsevier BV. http://dx.doi.org/10.1006/expr.1996.0127.

PRESTES, E.B.; STOCO, P.H.; MORAES, M.H.; MOURA, H.; GRISARD, E.C. Messenger RNA levels of the Polo-like kinase gene (PLK) correlate with cytokinesis in the *Trypanosoma rangeli* cell cycle. **Experimental Parasitology**, v. 204, set. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.exppara.2019.107727.

R Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. 2023. URL https://www.R-project.org/.

RAMIREZ, J.L. An Evolutionary View of *Trypanosoma cruzi* Telomeres. **Frontiers In Cellular And Infection Microbiology**, v. 9, 10 jan. 2020a. Frontiers Media SA. http://dx.doi.org/10.3389/fcimb.2019.00439.

RAMIREZ, J.L. *Trypanosoma cruzi* Genome 15 Years Later: what has been accomplished?. **Tropical Medicine And Infectious Disease**, v. 5, n. 3, 6 ago. 2020b. MDPI AG. http://dx.doi.org/10.3390/tropicalmed5030129.

REIS-CUNHA, J.L.; BARTHOLOMEU, D.C. *Trypanosoma cruzi* Genome Assemblies: challenges and milestones of assembling a highly repetitive and complex genome. **Methods In Molecular Biology**, p. 1-22, 2019. Springer New York. http://dx.doi.org/10.1007/978-1-4939-9148-8_1.

REIS-CUNHA, J.L.; COQUEIRO-DOS-SANTOS, A.; PIMENTA-CARVALHO, S.A.; MARQUES, L.P.; RODRIGUES-LUIZ, G.F.; BAPTISTA, R.P.; ALMEIDA, L.V.; HONORATO, N.R.M.; LOBO, F.P.; FRAGA, V.G. Accessing the Variability of Multicopy Genes in Complex Genomes using Unassembled Next-Generation Sequencing Reads: the case of *Trypanosoma cruzi* multigene families. **Mbio**, v. 13, n. 6, 20 dez. 2022. American Society for Microbiology. http://dx.doi.org/10.1128/mbio.02319-22.

REQUENA, J.M.; LÓPEZ, M.C.; ALONSO, C. Genomic repetitive DNA elements of *Trypanosoma cruzi*. **Parasitology Today**, v. 12, n. 7, p. 279-283, jul. 1996. Elsevier BV. http://dx.doi.org/10.1016/0169-4758(96)10024-7.

RNACENTRAL CONSORTIUM; SWEENEY, A.B.; PETROV, A.I.; BURKOV, B.; FINN, R.D.; BATEMAN, A.; SZYMANSKI, M.; KARLOWSKI, W.M.; GORODKIN, J.; SEEMANN, S. RNAcentral: a hub of information for non-coding rna sequences. **Nucleic Acids Research**, v. 47, n. 1, p. 221-229, 5 nov. 2018. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gky1034.

ROBERTS, A.J.; FAIRLAMB, A.H. The N-myristoylome of *Trypanosoma cruzi*. **Scientific Reports**, v. 6, n. 1, 5 ago. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/srep31078.

RUIZ-POSTIGO, J.A.; JAIN, S.; MIKHAILOV, A.; MAIA-ELKHOURY, A.N.; VALADAS, S.; WARUSAVITHANA, S.; OSMAN, M.; LIN, Z.; BESHAH, A.; YAJIMA, A.; GASIMOV, E. **Global leishmaniasis surveillance: 2019–2020, a baseline for the 2030 roadmap**. Disponível em: <https://www.who.int/publications/i/item/who-wer9635-401-419>. Acesso em: 13 jul. 2022.

SATOSHI, F. Estimation of the Number of Authentic Orphan Genes in Bacterial Genomes. **Dna Research**, v. 11, n. 4, p. 219-231, 1 jan. 2004. Oxford University Press (OUP). http://dx.doi.org/10.1093/dnares/11.4.219.

SCHRADER, L.; SCHMITZ, J. The impact of transposable elements in adaptive evolution. **Molecular Ecology**, v. 28, n. 6, p. 1537-1549, 4 ago. 2018. Wiley. http://dx.doi.org/10.1111/mec.14794.

SFEIR, A.; SYMINGTON, L.S. Microhomology-Mediated End Joining: a back-up survival mechanism or dedicated pathway?. **Trends In Biochemical Sciences**, v. 40, n. 11, p. 701-714, nov. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.tibs.2015.08.006.

SHAHBAAZ, M.; IMTAIYAZHASSAN, M.; AHMAD, F. Functional Annotation of Conserved Hypothetical Proteins from *Haemophilus influenzae* Rd KW20. **Plos One**, v. 8, n. 12, 31 dez. 2013. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0084263.

SHEN, W.; LE, S.; LI, Y.; HU, F. SeqKit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. **Plos One**, v. 11, n. 10, 5 out. 2016. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0163962.

SILVA, F. M.; NOYES, H.; CAMPANER, M.; JUNQUEIRA, A.C.V.; COURA, J.R.; AÑEZ, N.; SHAW, J.J.; STEVENS, J.R.; TEIXEIRA, M.M.G. Phylogeny, taxonomy and grouping of *Trypanosoma rangeli* isolates from man, triatomines and sylvatic mammals from widespread geographical origin based on SSU and ITS ribosomal sequences. **Parasitology**, v. 129, n. 5, p. 549-561, 5 out. 2004. Cambridge University Press (CUP). http://dx.doi.org/10.1017/s0031182004005931.

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

SOHN, J.; NAM, J. The present and future of *de novo* whole-genome assembly. **Briefings In Bioinformatics**, 14 out. 2016. Oxford University Press (OUP). http://dx.doi.org/10.1093/bib/bbw096.

SOUZA, R.T.; SANTOS, M.R.M.; LIMA, F.M.; EL-SAYED, N.M.; MYLER, P.J.; RUIZ, J.C.; SILVEIRA, J.F. New *Trypanosoma cruzi* Repeated Element That Shows Site Specificity for Insertion. **Eukaryotic**

Cell, v. 6, n. 7, p. 1228-1238, jul. 2007. American Society for Microbiology. http://dx.doi.org/10.1128/ec.00036-07.

STANKE, M.; WAACK, S. Gene prediction with a hidden Markov model and a new intron submodel. **Bioinformatics**, v. 19, n. 2, p. 215-225, 27 set. 2003. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btg1080.

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

STEINDEL, M. **Caracterização de cepas de *Trypanosoma cruzi* e *Trypanosoma rangeli* isoladas de reservatórios e vetores silvestres naturalmente infectados de Santa Catarina**. 1993. Tese de Doutorado (Programa de Pós-Graduação em Parasitologia). Universidade Federal de Minas Gerais, Belo Horizonte.

STEINDEL, M.; DIAS NETO, E.; PINTO, C.J.C.; GRISARD, E.C.; MENEZES, C.L.P.; MURTA, S.M.F.; SIMPSON, A.J.G.; ROMANHA, A.J. Randomly Amplified Polymorphic DNA (RAPD) and Isoenzyme Analysis of *Trypanosoma rangeli* Strains. **The Journal Of Eukaryotic Microbiology**, v. 41, n. 3, p. 261-267, mai. 1994. Wiley. http://dx.doi.org/10.1111/j.1550-7408.1994.tb01506.x.

STEVENS, J.; GIBSON, W. The Evolution of Salivarian Trypanosomes. **Memórias do Instituto Oswaldo Cruz**, v. 94, n. 2, p. 225-228, mar. 1999. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0074-02761999000200019.

STOCO, P.H.; WAGNER, G.; TALAVERA-LOPEZ, C.; GERBER, A.; ZAHA, A.; THOMPSON, C.E.; BARTHOLOMEU, D.C.; LÜCKEMEYER, D.D.; BAHIA, D.; LORETO, E., et al. Genome of the Avirulent Human-Infective Trypanosome—*Trypanosoma rangeli*. **Plos Neglected Tropical Diseases**, v. 8, n. 9, p. 3176, 18 set. 2014. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pntd.0003176.

STOCO, P.H.; MILETTI, L.C.; PICOZZI, K.; STEINDEL, M.; GRISARD, E.C. Other Major Trypanosomiasis. In: MARCONDES, C.B. (eds) **Arthropod Borne Diseases**. Springer, Cham., p. 299-324, 11 nov. 2016. Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-13884-8_19.

TALAVERA-LÓPEZ, C.; MESSENGER, L.A.; LEWIS, M.D.; YEO, M.; REIS-CUNHA, J.L.; MATOS, G.M.; BARTHOLOMEU, D.C.; CALZADA, J.E.; SALDAÑA, A.; RAMÍREZ, J.D. Repeat-Driven Generation of Antigenic Diversity in a Major Human Pathogen, *Trypanosoma cruzi*. **Frontiers In Cellular And Infection Microbiology**, v. 11, 3 mar. 2021. Frontiers Media SA. http://dx.doi.org/10.3389/fcimb.2021.614665.

TAUTZ, D.; DOMAZET-LOŠO, T. The evolutionary origin of orphan genes. **Nature Reviews Genetics**, v. 12, n. 10, p. 692-702, 31 ago. 2011. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrg3053.

THE, M.; MACCOSS, M.J.; NOBLE, W.S.; KÄLL, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. **Journal Of The American Society For Mass Spectrometry**, v. 27, n. 11, p. 1719-1727, 29 ago. 2016. American Chemical Society (ACS). http://dx.doi.org/10.1007/s13361-016-1460-7.

THE UNIPROT CONSORTIUM; BATEMAN, A.; MARTIN, M.; ORCHARD, S.; MAGRANE, M.; AGIVETOVA, R.; AHMAD, S.; ALPI, E.; BOWLER-BARNETT, E.H.; BRITTO, R. UniProt: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, v. 49, n. 1, p. 480-489, 25 nov. 2020. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkaa1100.

VALLEJO, G.A.; GUHL, F.; CARRANZA, J.C; LOZANO, L.e; SÁNCHEZ, J.L; JARAMILLO, J.C; GUALTERO, D.; CASTAÑEDA, N.; SILVA, J.C; STEINDEL, M. KDNA markers define two major *Trypanosoma rangeli* lineages in Latin-America. **Acta Tropica**, v. 81, n. 1, p. 77-82, jan. 2002. Elsevier BV. http://dx.doi.org/10.1016/s0001-706x(01)00186-3.

WAGNER, G. **Geração e análise comparativa de sequências genômicas de *Trypanosoma rangeli***. 2006. Dissertação de Mestrado (Biologia Celular e Molecular). Instituto Oswaldo Cruz, Rio de Janeiro, 2006.

WAGNER, G.; YAMANAKA, L.E.; MOURA, H.; LÜCKEMEYER, D.D.; SCHLINDWEIN, A. D.; STOCO, P.H.; FERREIRA, H.B.; BARR, J.R.; STEINDEL, M.; GRISARD, E.C. The *Trypanosoma rangeli* trypomastigote surfaceome reveals novel proteins and targets for specific diagnosis. **Journal Of Proteomics**, v. 82, p. 52-63, abr. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.jprot.2013.02.011.

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

WALKER, B.J.; ABEEL, T.; SHEA, T.; PRIEST, M.; ABOUELLIEL, A.; SAKTHIKUMAR, S.; CUOMO, C.A.; ZENG, Q.; WORTMAN, J.; YOUNG, S.K. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. **Plos One**, v. 9, n. 11, p. 112963-112976, 19 nov. 2014. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0112963.

WANG, W.; PENG, D.; BAPTISTA, R.P.; LI, Y.; KISSINGER, J.C.; TARLETON, R.L. Strain-specific genome evolution in *Trypanosoma cruzi*, the agent of Chagas disease. **Plos Pathogens**, v. 17, n. 1, 28 jan. 2021. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.ppat.1009254.

WEATHERLY, D.B.; BOEHLKE, C.; TARLETON, R.L. Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. **Bmc Genomics**, v. 10, n. 1, 1 jun. 2009. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2164-10-255.

WEI, L.; LIU, Y.; DUBCHAK, I.; SHON, J.; PARK, J. Comparative genomics approaches to study organism similarities and differences. **Journal Of Biomedical Informatics**, v. 35, n. 2, p. 142-150, abr. 2002. Elsevier BV. http://dx.doi.org/10.1016/s1532-0464(02)00506-3.

WESTENBERGER, S.J.; CERQUEIRA, G.C.; EL-SAYED, N.M.; ZINGALES, B.; CAMPBELL, A.D.; STURM, N.R. *Trypanosoma cruzi* mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. **Bmc Genomics**, v. 7, n. 1, p. 1-2, 22 mar. 2006. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2164-7-60.

WICKSTEAD, B.; ERSFELD, K.; GULL, K. Repetitive Elements in Genomes of Parasitic Protozoa. **Microbiology And Molecular Biology Reviews**, v. 67, n. 3, p. 360-375, set. 2003. American Society for Microbiology. http://dx.doi.org/10.1128/mmbr.67.3.360-375.2003.

XIA, X. Comparative Genomics. **Springerbriefs In Genetics**, 2013. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-37146-2.

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

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

ZINGALES, B. *Trypanosoma cruzi* genetic diversity: something new for something known about chagas disease manifestations, serodiagnosis and drug sensitivity. **Acta Tropica**, v. 184, p. 38-52, ago. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.actatropica.2017.09.017.

APÊNDICE A

Este apêndice é dedicado aos dados moleculares das sondas de genes de cópia única identificados no genoma de *Trypanosoma cruzi* CL Brener.

Tabela 7. Informações dos genes marcadores identificados no genoma de *Trypanosoma cruzi* (cepa CL Brener) utilizados durante a etapa de validação experimental da montagem da versão 2 do genoma de *T. rangeli* SC58.

Código de acesso no TriTrypDB	Anotação do produto gênico	Localização genômica	Iniciador senso	Iniciador antissenso	Tamanho do produto de PCR
TcCLB.503793.20	phosphatidylinositol (3,5) kinase, putative (fragment)	TcChr39-S: 379,142 .. 380,494 (+)	GGTCATGGCAAGCACAGTT	AACAATAACGGAGGGTCGC	510 pb
TcCLB.504109.170	hypothetical protein, conserved	TcChr39-P: 10,454 .. 11,866 (-)	TCTCGAAAGATGGTTCTCAG	ACACATATCGATGTGGAGTG	334 pb
TcCLB.506127.90	Hereditary spastic paraplegia protein strumpellin, putative	TcChr39-S: 612,942 .. 616,493 (+)	AAGCACCATCATCGCTGAG	ATGACACCGTAGAGGTGCA	373 pb
TcCLB.507009.90	iron-sulfur cluster assembly protein, putative	TcChr39-S: 1,325,049 .. 1,325,540 (+)	TCGCTTCACAAGCACTCCA	GGGTTGAAGCCTCTGTTGA	482 pb
TcCLB.507609.40	delta-4 fatty acid desaturase, putative	TcChr37-P: 1,316,342 .. 1,317,616 (+)	TGAAGCTTGCCGCTATTCAT	CCATGGCAAGTGATTGAAGA	366 pb
TcCLB.507611.380	hypothetical protein	TcChr37-S: 100,543 .. 101,184 (+)	GATTTGCCTGAGGAGTCCTA	CGAAGCCTCATGGATAACTC	560 pb
TcCLB.507641.120	ATP-dependent DEAD/H RNA helicase, putative	TcChr37-S: 1,167,741 .. 1,169,567 (-)	GAGGCCTTTAAACTTGTGAG	ATTTCCGCAACGTCCGGC	693 pb
TcCLB.507681.160	40S ribosomal protein S24E, putative	TcChr4-P: 38,962 .. 39,375 (-)	AGAAGAAGAAGGCGGAGGT	TGCATCTTGCCCTTCGCC	380 pb
TcCLB.507713.30	heat shock protein 85, putative	TcChr37-S: 793,891 .. 796,005 (-)	GCGCCGTTTGACATGTTTG	GGCGAGGTCTCCAGCTTCT	479 pb
TcCLB.508465.120	syntaxin, putative	TcChr39-S: 1,536,810 .. 1,537,718 (+)	CTTTGGGTGGAGAAGATGG	GCAAACAACAGCCCAATGAT	763 pb
TcCLB.509715.120	phosphatidylinositol-4-phosphate 5-kinase, putative	TcChr4-P: 107,469 .. 110,231 (-)	GTCATGCTGAATGAGAGGGA	GACGCTCCCTTGAAGCATAA	555 pb
TcCLB.510129.20	Surface membrane protein	TcChr37-P: 870,126 .. 871,505 (-)	CATCATCGCGGGCAGCAT	TGTACAGGCGCACTGTAAC	619 pb
TcCLB.511041.40	hexose transporter, putative	TcChr37-S: 292,428 .. 294,062 (+)	TGACTGGCATCAATGCGGT	TTCGGACGAAACGAGCGC	678 pb

Fonte: adaptado de DO CARMO, 2022.

APÊNDICE B

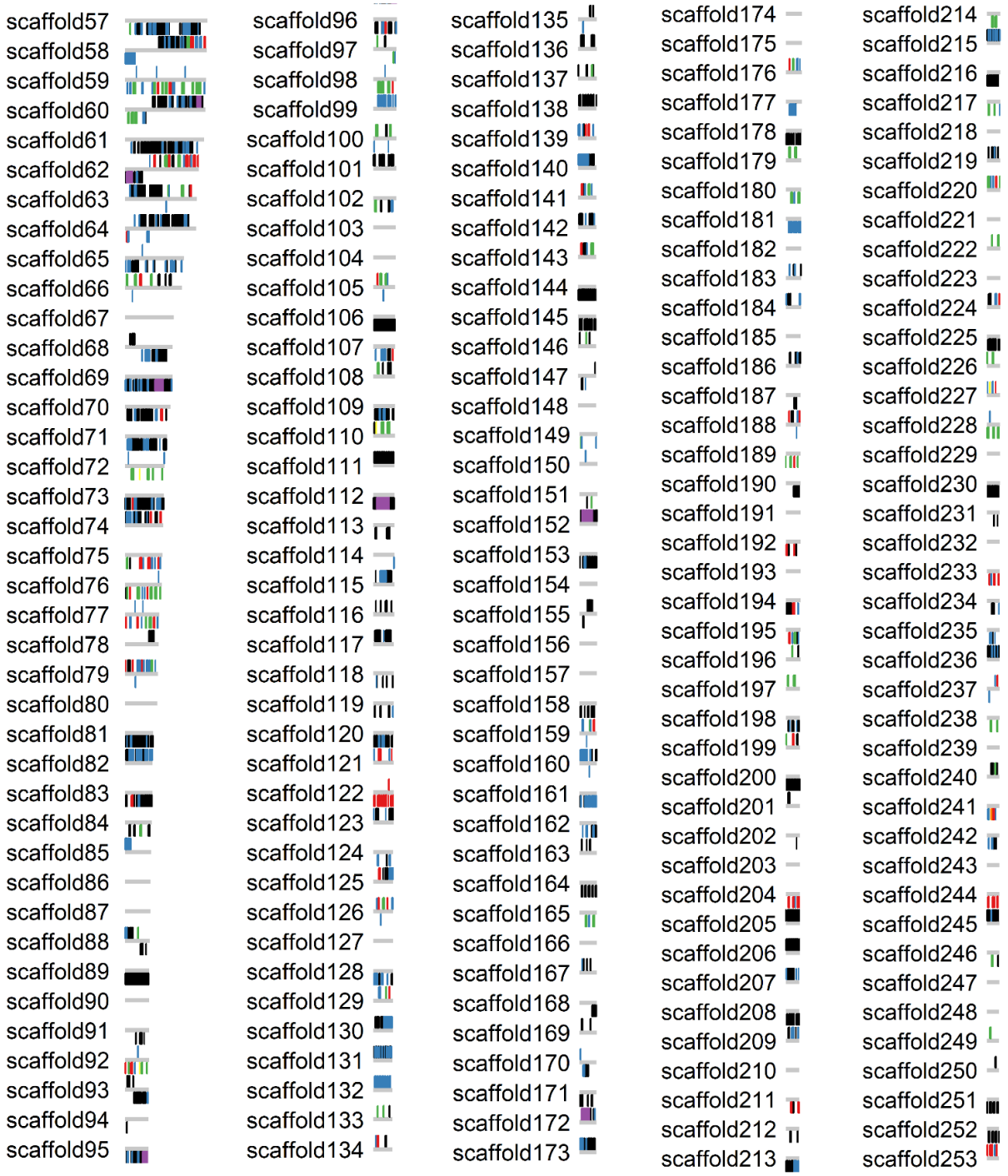
Este apêndice contém o link para a representação gráfica das regiões sintênicas identificadas entre os *scaffolds* da versão 2 do genoma de *Trypanosoma rangeli* SC58 e os cromossomos descritos em *T. cruzi* Sylvio X10/1, obtido do TriTrypDB (versão 57):

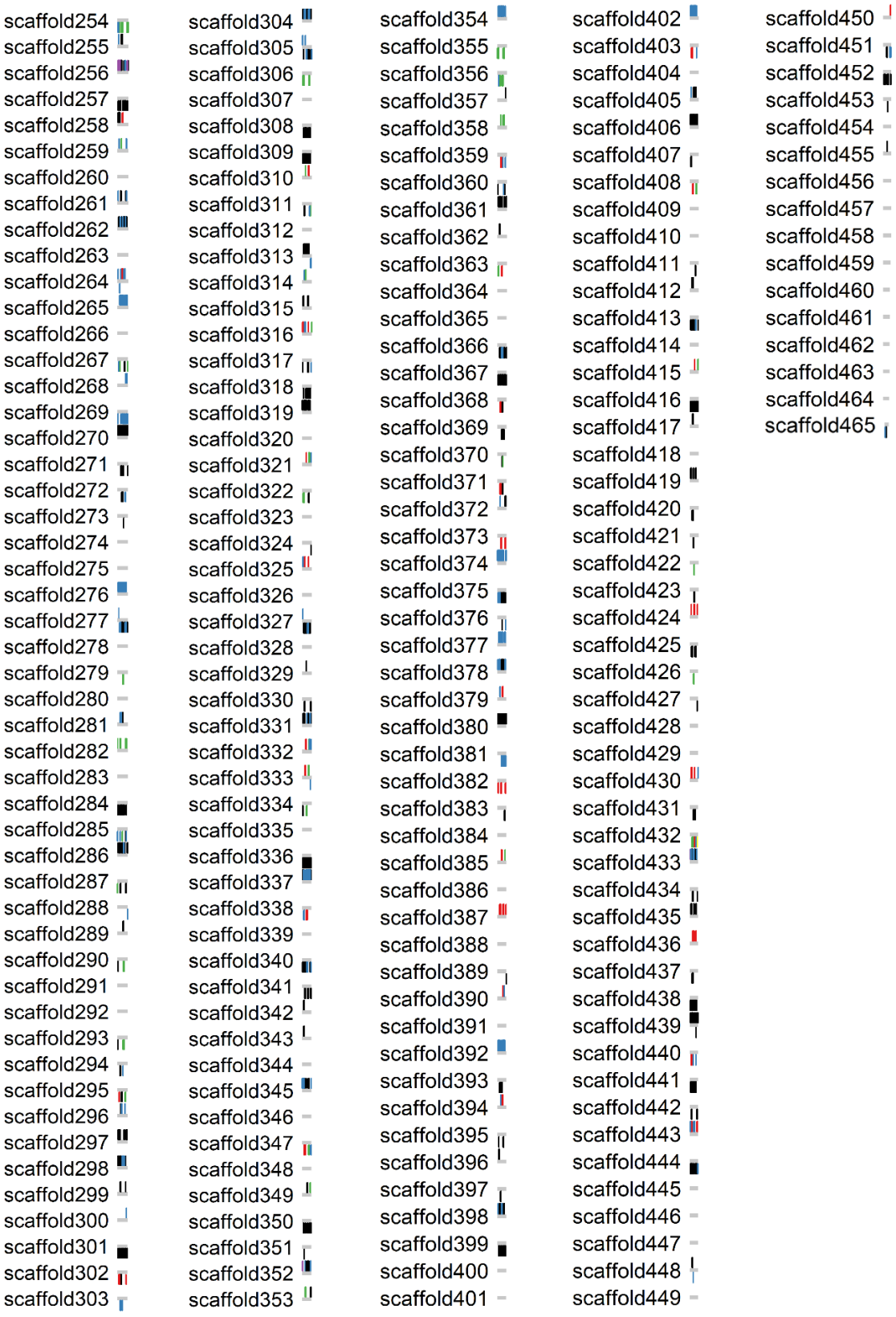
<https://github.com/GuiMaia/Thesis_Shared_Files/blob/be32579f3e3d4257ac7b27a567f0641ee5891526/Apendice_B_Synteny_TrangeliSC58_v2_vs_Tcruzi_SylvioX10-1.svg>.

APÊNDICE C

Este apêndice é dedicado à exposição das CDS nos *scaffolds* com menos de 100 mil pares de bases obtidos na versão 2 do genoma de *Trypanosoma rangeli* SC58.

Figura 14. Distribuição das CDS preditas e anotadas nos *scaffolds* com menos de 100 mil pares de bases obtidos na versão 2 do genoma de *Trypanosoma rangeli* SC58, com destaque para proteínas de superfície.





Legenda: As CDS estão dispostas verticalmente (preto), ao longo da extensão de cada *scaffold* (cinza). Estão destacadas as CDS anotadas como hipotéticas (azul), GP63 (vermelho), sialidases (verde), DGF-1 (roxo), MASP (amarelo), amastinas (marrom) e mucinas (laranja).

APÊNDICE D

Este apêndice contém o link das caracterizações *in silico* realizadas para as CDS preditas e anotadas da versão 2 do genoma de *Trypanosoma rangeli* SC58, na forma de tabela:

<https://github.com/GuiMaia/Thesis_Shared_Files/blob/be32579f3e3d4257ac7b27a567f0641ee5891526/Apendice_D_TrangeliSC58_v2_CDS_Characterization.xlsx>.