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

Eric Kazuo Kawagoe

GENÔMICA COMPARATIVA DE ISOLADOS DE *Leishmania infantum* DE SANTA CATARINA E RIO GRANDE DO NORTE

Florianópolis
2021

Eric Kazuo Kawagoe

GENÔMICA COMPARATIVA DE ISOLADOS DE *Leishmania infantum* DE SANTA CATARINA E RIO GRANDE DO NORTE

Dissertação submetida ao Programa de Pós-Graduação em Biotecnologia e Biociências da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Biotecnologia e Biociências

Orientador: Prof. Dr. Glauber Wagner

Coorientador: Prof. Dr. Guilherme de Toledo e Silva

Florianópolis

2021

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

Kawagoe, Eric Kazuo
Genômica comparativa de isolados de Leishmania infantum de Santa Catarina e Rio Grande do Norte / Eric Kazuo Kawagoe ; orientador, Glauber Wagner, coorientador, Guilherme de Toledo e Silva, 2021.
101 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Programa de Pós Graduação em Biotecnologia e Biociências, Florianópolis, 2021.

Inclui referências.

1. Biotecnologia e Biociências. 2. Genômica comparativa. 3. Pangenoma. 4. Leishmaniose visceral. 5. Leishmania infantum. I. Wagner, Glauber. II. Toledo e Silva, Guilherme de. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Biotecnologia e Biociências. IV. Título.

Eric Kazuo Kawagoe

Título: Genômica comparativa de isolados de *Leishmania infantum* de Santa Catarina e Rio Grande do Norte

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Glauber Wagner, Dr.
Universidade Federal de Santa Catarina

Prof. Rodrigo de Paula Baptista, Dr.
University of Georgia

Prof.ª Patrícia Hermes Stoco, Dr.ª
Universidade Federal de Santa Catarina

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

Documento assinado digitalmente
Glauber Wagner
Data: 16/01/2022 21:40:58-0300
CPF: 007.019.139-59
Verifique as assinaturas em https://v.ufsc.br

_____

Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Glauber Wagner
Data: 16/01/2022 21:38:45-0300
CPF: 007.019.139-59
Verifique as assinaturas em https://v.ufsc.br

_____

Prof. Dr. Glauber Wagner
Orientador

Florianópolis, 2021.

AGRADECIMENTOS

Ciência é plural e se cheguei até aqui foi por não caminhar sozinho. A todos que influenciaram este trabalho, direta ou indiretamente, meu sincero obrigado. Amo todos vocês.

Ao meu orientador, *Glauber*, pelo salto de fé ao aceitar um aluno de outra universidade sem que tivéssemos nos conhecido antes. Obrigado por proporcionar um ambiente de trabalho leve desde nosso primeiro contato. Um ambiente de trabalho é tão bom quanto aqueles que estão nele inseridos e nosso grupo é um exemplo disso.

Ao meu coorientador, *Guilherme*, por aceitar contribuir com este trabalho. Obrigado pela confiança e pelas discussões englobando outras perspectivas que possibilitaram evoluir meu pensamento científico.

À minha família, *Takao*, *Lie*, *Kaory* e *Yoneko*, pelo apoio e incentivo que permitiram me dedicar ao mundo da pesquisa. Cientistas são curiosos e se hoje minha profissão é viver de perguntas foi porque tive a liberdade para questionar.

À *Flavia*, pelo apoio apesar da distância e por ser meu porto seguro em meio às águas desconhecidas do mundo acadêmico. Obrigado pelo companheirismo e por entender os momentos de ausência.

Aos amigos de longa data, *Lucas*, *Yohan*, *Kaio* e *Eric*. Obrigado pelo apoio e pelos encontros virtuais que permitiram esvaziar a mente e descomprimir. Obrigado pelo nosso novo passatempo e que nossa amizade perdure e vença as barreiras da distância.

Aos amigos imbatíveis e tristes, *Guilherme*, *Tatiany*, *Vilmar*, *Renato*, *Karin*, *Jaime*, *Dayane* e *Carolina*, pelo acolhimento. Obrigado pela amizade que vai além das paredes do laboratório. Que possamos nos reunir novamente, em um futuro não tão distante, para dar nossos abraços coletivos.

Aos colegas de corredor, do *Laboratório de Protozoologia* e *Laboratório de Virologia Aplicada*, pelos momentos de descontração e pelas nossas conversas durante o cafezinho. Obrigado pelas pausas necessárias que revigoravam as tardes.

Ao *Pedro* e a *Saori*, pelo ouvido amigo e conselhos durante minha tentativa falha de realizar experimentos de bancada. Obrigado também por me direcionar ao mundo *in silico*, onde continuo a seguir sorrindo.

Ao *Systems Biology and Genomics Laboratory* e, em especial, ao *Ivan*. Toda grande jornada começa com um simples passo e, neste laboratório, dei meus primeiros passos rumo ao mundo da bioinformática. Obrigado pelos ensinamentos que fundamentaram minha caminhada.

À *Alexandra Elbakyan*, criadora do *Sci-Hub*, por instaurar um novo patamar para o conceito de ciência aberta.

À *CAPES* e ao *CNPq* pelo auxílio financeiro que possibilitaram o desenvolvimento deste projeto.

E, por fim, mas não menos importante, aos leitores desta minha narrativa científica. Obrigado pelo tempo e interesse em ler este trabalho. Espero ter retornado ao menos um pouco do que recebi ao longo da minha jornada acadêmica.

“I love deadlines. I like the wooshing sound they make as they fly by.”

(Douglas Adams)

RESUMO

Leishmania infantum é um protozoário unicelular flagelado e parasita obrigatório, que tem como reservatório humanos e canídeos. Este parasito é o agente causador de leishmaniose visceral nas Américas, considerada uma antropozoonose tropical negligenciada. A doença acomete órgãos linfoides como medula óssea, fígado e baço, podendo causar sintomas como febre, perda de peso, hepatoesplenomegalia e anemia. As ciências ômicas podem atuar em estudos para elucidar mecanismos referentes à biologia de parasitos ou para identificar regiões que apresentam, ou não, similaridade entre diferentes organismos. No contexto de similaridade se enquadram estudos de genômica comparativa, que possibilitam identificar, por exemplo, regiões compartilhadas, genes e proteínas essenciais presentes no genoma de organismos de uma mesma espécie. O objetivo deste trabalho foi realizar a montagem, anotação funcional e análise comparativa a partir de dados de sequenciamento de segunda geração de amostras de *L. infantum* coletadas em Santa Catarina e obtidas de bancos de dados públicos para o Rio Grande do Norte. Ambas as etapas de montagem e anotação foram realizadas utilizando dados públicos obtidos do *TriTrypDB* da cepa JPCM5 de *L. infantum* como referência. As montagens preliminares foram realizadas pelo programa *SPAdes* e, então, refinadas pelos programas *SSPACE* e *GapFiller*, para ordenação de *scaffolds* e preenchimento de *gaps*, respectivamente. A ordenação de *scaffolds* em cromossomos foi realizada pelo programa *SAMtools* a partir de alinhamentos contra a referência e refinado pelos programas *BWA-MEM* e *Pilon*. A predição gênica foi realizada pelo preditor *AUGUSTUS* e a anotação funcional pela pipeline *AnnotaPipeline*. As montagens finais apresentaram média de 32,6 Mb totais e 2.790,49 bases não identificadas distribuídas em 36 *scaffolds* contíguos, o que representa 99,5% do tamanho do genoma de referência. As montagens apresentaram média de 8.695 proteínas preditas e anotadas, com média de 2.983 proteínas anotadas como hipotéticas e sem anotação funcional. A partir do resultado destas montagens e predições, o programa *OrthoFinder* foi utilizado para realizar a análise de ortologia entre todas as proteínas anotadas, mas o perfil gênico se mostrou conservado e sem proteínas específicas, que garantem vantagem evolutiva, para uma única amostra. A anotação funcional com base em termos ontológicos permitiu identificar processo biológicos clássicos para as amostras de Santa Catarina, mesmo sem a presença do vetor clássico no município. Polimorfismos de base única foram detectados pelo programa *FreeBayes* e seus impactos foram preditos pelo programa *SnpEff*. Polimorfismos não-sinônimos de alto impacto não foram frequentes e apareceram em poucas amostras. Por fim, este trabalho permitiu uma análise comparativa de diversos genomas altamente sintênicos com base em montagem e anotação de genomas, gerando informações funcionais que podem atuar em conjunto com estudos mais tradicionais de identificação e perfil de variantes.

Palavras-chave: Genômica comparativa. Pangenoma. Leishmaniose visceral. *Leishmania infantum*. Florianópolis.

ABSTRACT

Leishmania infantum is an intracellular parasite that infects mammalian hosts and sandflies. It is the main etiological agent for visceral leishmaniasis, which is considered a neglected tropical disease, in the Americas. Visceral leishmaniasis affects mainly lymphoid organs and manifests in hepatosplenomegaly, weight loss, fever and anaemia. Omics refers to a subfield in bioinformatics focused on biological sequences which can be employed in studies for biological insights. Comparative genomics can involve sequence similarity for genomic discoveries, such as syntenic regions and gene/protein discovery within a species. We employed comparative genomics in assembled and annotated genomes, generated from short reads, obtained in Santa Catarina and from public databases. Both genome assembly and annotation used the current reference genome from *TriTrypDB*, *Leishmania infantum* JPCM5. Draft assemblies were generated by *SPAdes*, and submitted to *SSPACE* and *GapFiller* for scaffolding/gap filling. Scaffolds were aligned to the reference genome and merged into polished chromosomes through *BWA-MEM* and *Pilon*. *AUGUSTUS* was used for gene prediction and *AnnotaPipeline* obtained functional annotations. Final assemblies presented 36 scaffolds, and an average of 32.6 Mb and 2,790.29 unidentified nucleotides. Gene predictions presented and average of 8,695 annotated proteins with 2,983 hypothetical proteins without functional annotations. Functional annotations, based on ontology, allowed the identification of classical biological processes despite the absence of classical vectors in Santa Catarina. Annotated proteins were submitted to *OrthoFinder* for orthology inference, which resulted in conserved gene profiles throughout all samples. Single nucleotide polymorphisms were were detected by *FreeBayes* and annotated by *SnpEff* for potential impacts. High impact non-synonymous mutations were not frequent and not dispersed in multiple samples. In conclusion, comparative genomics resulted in highly syntenic genomes and functional annotations that could be incorporated in traditional studies involving variant analysis.

Keywords: Comparative genomics. Pangenome. Visceral leishmaniasis. *Leishmania infantum*. Florianópolis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Ah	Humano assintomático
CLh	Leishmaniose cutânea humana
DNA	Ácido desoxirribonucleico
LV	Leishmaniose visceral
LVC	Leishmaniose visceral canina
LVH	Leishmaniose visceral humana
GO	*Gene Ontology*
RN	Rio Grande do Norte
RNA	Ácido ribonucleico
RNA-seq	Sequenciamento de RNA
SC	Santa Catarina
SNP	Polimorfismo de nucleotídeo único
SRA	*Sequence Read Archive*
UFSC	Universidade Federal de Santa Catarina
VLd	Leishmaniose visceral canina
VLh	Leishmaniose visceral humana
kDNA	DNA do cinetoplasto
pb	Pares de base
rRNA	RNA ribossomal
tRNA	RNA transportador

SUMÁRIO

1 INTRODUÇÃO

1.1 *LEISHMANIA INFANTUM* E LEISHMANIOSE VISCERAL

Leishmania (Leishmania) infantum é um protozoário da classe Kinetoplastida, ordem Trypanosomatida e família Trypanosomatidae (AKHOUNDI et al., 2017). Protozoários da família Trypanosomatidae são organismos unicelulares flagelados e parasitos obrigatórios de hospedeiros como insetos, répteis e mamíferos (ASLETT et al., 2010). O gênero *Leishmania* possui também uma região citoplasmática característica, rica em DNA extracromossomal, denominada cinetoplasto ou kDNA (ALEMAN, 1969; SUNTER; GULL, 2017). O cinetoplasto fica localizado na base do flagelo e apresenta uma rede característica onde duas estruturas de DNA circular se entremeiam: minicírculos e maxicírculos (LUKEŠ et al., 2002; SHLOMAI, 2004).

Minicírculos apresentam maior número de cópias e sequências heterogêneas enquanto maxicírculos apresentam menor número de cópias e se mostram análogas ao DNA mitocondrial de outros organismos eucariotos – codificando RNAs ribossomais (rRNA) e proteínas relacionadas à cadeia respiratória (APHASIZHEV; APHASIZHEVA, 2014; CAVALCANTI; SOUZA, 2018; LUKEŠ et al., 2002).

Protozoários do gênero *Leishmania* são causadores de leishmanioses, caracterizadas como doenças tropicais e negligenciadas – afetando, principalmente, países em desenvolvimento – e são um grande problema de saúde pública mundial (AKHOUNDI et al., 2017; HERRERA et al., 2017; RUIZ-POSTIGO et al., 2021). Dentre as diferentes formas clínicas de leishmanioses, encontramos a leishmaniose visceral (LV) – tanto humana (LVH) quanto canina (LVC) – em crescente expansão no Brasil e com ocorrências no estado de Santa Catarina (DIAS et al., 2013; FIGUEIREDO et al., 2012; MAZIERO et al., 2014; SOUZA et al., 2010; STEINDEL et al., 2013). Nas Américas, esta zoonose é causada pela *L. infantum* – que tem como hospedeiros diferentes espécies de mamíferos, como canídeos e humanos – e é transmitida por fêmeas infectadas de dípteros hematófagos da família Psychodidae,

popularmente conhecidos como flebotomíneos (HARHAY et al., 2011; SOUZA et al., 2010).

A LV é a forma clínica mais severa das leishmanioses e é uma doença sistêmica que acomete os órgãos linfoides tanto em cães quanto em humanos, além de apresentar como sinais característicos a hepatoesplenomegalia associada a perda de peso. Por afetar órgãos linfoides e o sistema fagocítico mononuclear, a saúde de hospedeiros infectados fica debilitada em decorrência da maior susceptibilidade a infecções bacterianas secundárias. O período de incubação da doença pode variar de 3 a 8 meses e suas manifestações clínicas diferem entre subclínicas, oligossintomáticas, ou completamente estabelecidas (SAKKAS; GARTZONIKA; LEVIDIOTOU, 2016; TORRES-GUERRERO et al., 2017).

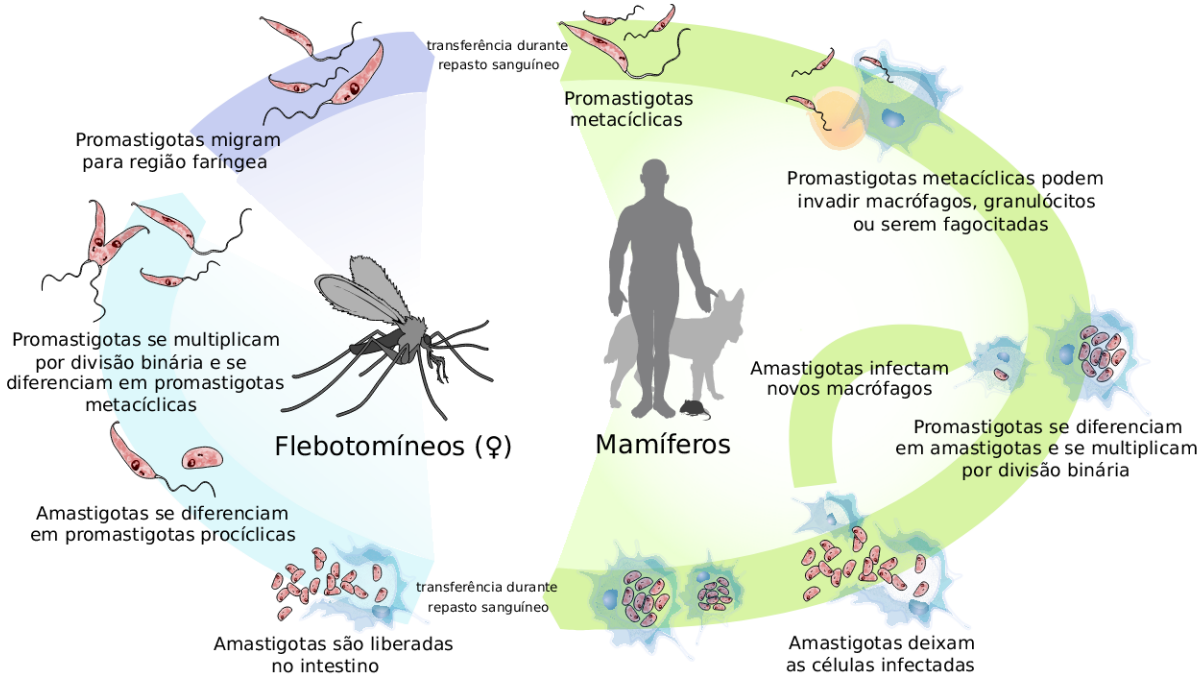
De acordo com a Organização Mundial da Saúde (2021), 79 países são endêmicos para LV, sendo que África, Europa e Américas apresentam o maior número de casos. As Américas representam 16% do número total de casos de LV relatados, distribuídos entre Argentina, Bolívia, Brasil, Colombia, El Salvador, Guatemala, Honduras, México, Paraguai, Uruguai e Venezuela. Em 2019, 97% dos 2.603 casos relatados nas Américas foram atribuídos ao Brasil, o que constitui 2.529 casos e representa 15,52% dos casos mundiais (PAN AMERICAN HEALTH ORGANIZATION, 2020).

O aumento da incidência de casos em diversas regiões brasileiras pode estar relacionada com o fluxo de pessoas entre as diferentes regiões do país, acompanhadas de seus animais, saindo de um ambiente rural para um ambiente urbano (COSTA et al., 2018; OLIVEIRA; MOREIRA, 2021). Desta forma, cães domésticos – ou cães que circulam livremente – podem favorecer a disseminação do parasito em novas regiões e, posteriormente, podem atuar como elo de transmissão entre flebotomíneos e humanos (CARRILLO; MORENO, 2009; OLIVEIRA; MOREIRA, 2021).

A infecção ocorre no momento do repasto sanguíneo (**Figura 1**) onde fêmeas de flebotomíneos infectadas regurgitam promastigotas – forma infectante do parasito – na pele lesada durante a picada no hospedeiro mamífero (GHARBI et al., 2015).

As formas promastigotas são fagocitadas por células fagocíticas mononucleares – principalmente macrófagos – e dentro destas células se diferenciam para formas amastigotas, a forma proliferativa dentro de hospedeiros mamíferos (GHARBI et al., 2015; HARHAY et al., 2011). Flebotomíneos que se alimentam de hospedeiros infectados ingerem as formas amastigotas do parasito, as quais se diferenciam em promastigotas no trato digestório do inseto e se multiplicam por divisão binária, posteriormente se diferenciando em formas promastigotas metacíclicas que serão inoculadas no momento do próximo repasto sanguíneo (HARHAY et al., 2011).

Figura 1. Ciclo de vida de *Leishmania infantum* em hospedeiros vertebrados e invertebrados.



Fonte: Adaptado de HARHAY et al., 2011.

1.2 CONTEXTO GENÔMICO DE *LEISHMANIA*

Organismos pertencentes ao gênero *Leishmania* são caracterizados como seres eucariontes, porém apresentam peculiaridades genômicas: seus genomas são altamente sintênicos e se organizam de tal forma que não apresentam introns (EL-

SAYED, 2005; KAZEMI, 2011; RAVEL, 1999; ROGERS et al., 2011; WINCKER et al., 1996). O tamanho aproximado de genomas para *Leishmania* spp. é de 32,8 Mb, com o número de cromossomos variando entre 34 e 36 (CANTACESSI et al., 2015; KAZEMI, 2011; ROGERS et al., 2011).

O primeiro genoma completo de *Leishmania* referente à cepa Friedlin de *L. major* foi publicado por Ivens e colaboradores (2005) e serviu como etapa inicial para o progresso de estudos genômicos referentes à leishmanioses, sendo que os genomas de *L. infantum* e *L. braziliensis* foram publicados posteriormente (CANTACESSI et al., 2015; PEACOCK et al., 2007; ULIANA; RUIZ; CRUZ, 2008). Com o surgimento das tecnologias de sequenciamento em larga escala, demais espécies de *Leishmania* tiveram seu genoma completamente sequenciado, sendo que o genoma de referência de *L. donovani* foi o primeiro a ser completado utilizando estas tecnologias (CANTACESSI et al., 2015; DOWNING et al., 2011).

Com o aumento do número de genomas disponíveis, um banco de dados voltado para a integração de dados genômicos de tripanossomatídeos foi desenvolvido, o *TriTrypDB*, que continha, inicialmente, conjuntos de dados de *L. braziliensis*, *L. infantum*, *L. major*, *L. tarentolae*, *Trypanosoma brucei* e *T. cruzi* (ASLETT et al., 2010). Desde então, o *TriTrypDB* incorporou dados proteômicos e transcriptômicos que – em conjunto com os dados genômicos – estão em constante expansão e atualização. O *TriTrypDB* v55 (02/12/2021) apresenta 80 genomas distribuídos entre 34 diferentes espécies de tripanossomatídeos, sendo que 30 destes genomas englobam 17 espécies do gênero *Leishmania*.

1.3 GENOMA E ESTUDOS GENÔMICOS

O genoma pode ser definido como todo o conjunto de DNA de uma dada espécie de organismo. Nele estão presentes todos os genes – unidades hereditárias de um ser vivo – que podem se referir às sequências de DNA tanto codificantes quanto não-codificantes da espécie (BARNES, 2007; PEVSNER, 2015; SNYDER; GERSTEIN, 2003). Os genes podem apresentar homologia entre si, sendo

classificados em duas classes: ortólogos e parálogos (ALTENHOFF; DESSIMOZ, 2012). Genes ortólogos são aqueles que possuem ancestralidade comum e divergiram por processo de especiação – normalmente mantendo sua função – enquanto genes parálogos divergem por meio de duplicações gênicas – podendo manter ou não suas funções originais (ALTENHOFF; DESSIMOZ, 2012).

Com o advento das tecnologias de sequenciamento em larga escala, estudos relacionados às sequências de um genoma podem utilizar diferentes abordagens de bioinformática, genética e biologia celular (PEVSNER, 2015). Dentro da bioinformática, duas grandes áreas se destacam: (i) análise de sequências de nucleotídeos e aminoácidos, assim como domínios e estruturas de proteínas; e (ii) desenvolvimento de ferramentas, na forma de algoritmos ou bancos de dados, que possibilitam o manejo ou integração entre dados de diferentes fontes (PAL; BANDYOPADHYAY; RAY, 2006; VERLI, 2014). O intuito deste trabalho é o de análise de sequências, principalmente, nucleotídicas.

Estudos genômicos voltados para análise de sequências podem ser utilizados para elucidar informações gerais acerca de um genoma – como seu tamanho, número de cromossomos e genes (codificantes e não-codificantes) – ou para comparar diferentes genomas por meio de estudos de genômica comparativa, visando identificar elementos e regiões que possuam ortologia ou sintenia (PEVSNER, 2015). Estudos genômicos, além disso, permitem uma visão geral acerca dos componentes do genoma, como os componentes interagem entre si e como funcionam dentro de sistemas biológicos (LOCKHART; WINZELER, 2000).

1.4 SEQUENCIAMENTO E MONTAGEM DE GENOMAS

O sequenciamento do genoma humano, realizado em 2001, instigou a determinação de sequências de outros organismos modelos (BARNES, 2007; INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM et al., 2001; LOCKHART; WINZELER, 2000; REUTER; SPACEK; SNYDER, 2015; VENTER et al., 2001). Inicialmente, os sequenciamentos eram realizados pelo método de

Sanger, altamente custoso e com número limitado de sequenciamento de bases (REUTER; SPACEK; SNYDER, 2015). O declínio de custos e o avanço das tecnologias de sequenciamento em larga escala permitiram, portanto, a caracterização de sequências de organismos de diferentes espécies (ALKAN; SAJJADIAN; EICHLER, 2011; REUTER; SPACEK; SNYDER, 2015).

Atualmente, as diferentes plataformas de sequenciamento podem ser divididas em três tipos: (i) sequenciamento de primeira geração, (ii) sequenciamento de segunda geração, e (iii) sequenciamento de terceira geração. As plataformas de sequenciamento diferem, principalmente, em relação ao tamanho e quantidade de dados (*reads*) gerados, a qualidade destes *reads*, além do tempo e custo necessários para gerá-los (HORNER et al., 2010; POP, 2009).

O sequenciamento de primeira geração se refere ao método desenvolvido por Sanger, Nicklen e Coulson (1977), que consiste na confecção de uma fita complementar de DNA, a partir de uma fita molde, na presença de desoxirribonucleotídeos trifosfatados (dNTPs) – que são utilizados para síntese pela DNA polimerase – e dideoxirribonucleotídeos trifosfatados (ddNTPs) – que param o processo de síntese quando inseridos nas terminações das cadeias dos oligonucleotídeos sintetizados – sendo os ddNTPs terminais utilizados para revelar a sequência de DNA presente na fita molde (MOROZOVA; MARRA, 2008; SANGER; NICKLEN; COULSON, 1977).

Sequenciamentos de segunda geração possibilitam que um maior volume de dados seja gerado com um menor custo, quando comparados ao método de Sanger (DEL ANGEL et al., 2018; METZKER, 2010). A tecnologia de sequenciamento de segunda geração mais comumente empregada é a de sequenciamento por síntese – utilizada pela plataforma Illumina – que gera *reads* com comprimento até 300 pares de base (ALKAN; SAJJADIAN; EICHLER, 2011; HEYDARI et al., 2019). Sequenciamentos por síntese necessitam de uma etapa de amplificação antes do sequenciamento em massa, e ambas etapas utilizam o princípio do método de Sanger – de sequenciamento por extensão – onde bases marcadas complementares a fita molde são adicionadas a nova fita sintetizada (HORNER et al., 2010).

Por fim, métodos de sequenciamento de terceira geração diferem dos demais em algumas características: possibilitam que *reads* mais longos sejam gerados a partir de uma única molécula – com maiores taxas de erro – e não necessitam de uma etapa de amplificação prévia, o que diminui o tempo de preparo para o sequenciamento (DEL ANGEL et al., 2018; LIU et al., 2012). Uma nova característica que está presente nos métodos de sequenciamento de terceira geração é que seus sinais são capturados em tempo real, conforme novos nucleotídeos são adicionados na fita complementar sintetizada (LIU et al., 2012).

Uma vez sequenciado o genoma ocorre uma etapa fundamental para a análise, sua montagem, que pode ocorrer de três formas: (i) montagem *de novo*, (ii) montagem pareada a um genoma de referência, ou (iii) uma combinação dessas duas abordagens (CHIKHI; MEDVEDEV, 2014; MARTIN; WANG, 2011; MILLER; KOREN; SUTTON, 2010). Montagens *de novo* se distinguem de montagens com referência pelo fato de não possuírem uma sequência guia que apresenta similaridade ao genoma do organismo modelo. Portanto, este tipo de abordagem pode ser utilizado para reconstruir genomas de organismos que ainda não possuem uma referência ou que não foram previamente sequenciados (POP, 2009). Já montagens pareadas a uma referência podem ser realizadas para espécies que já possuam um genoma de referência ou quando organismos de espécies filogeneticamente próximas apresentam genoma similar ao modelo de estudo (POP, 2009).

Grande parte das ferramentas de montagem para dados de sequenciamento de segunda geração – que geram *short reads* – são baseados em grafos de *de Bruijn* e utilizam *k-mers* (*substrings* de *reads* com comprimento *k*), que são agrupados e sobrepostos – baseados em sua similaridade – formando sequências contíguas que representam a sequência original (CHIKHI; MEDVEDEV, 2014; MARTIN; WANG, 2011; MILLER; KOREN; SUTTON, 2010). Devido a sobreposição de *reads* com base em similaridade, montagens *de novo* – utilizando dados de sequenciamento de segunda geração – acabam por colapsar sequências repetitivas, causando uma diminuição da complexidade genômica (ALKAN; SAJJADIAN;

EICHLER, 2011; SOHN; NAM, 2016). No caso de montagens comparativas – onde genomas de referência atuam como guia no processo de montagem – é possível que sequências – presentes, por exemplo, em regiões de inserção – sejam excluídas por não apresentarem correspondências com a referência (POP, 2009).

Uma etapa indispensável antes da etapa de montagem é a de controle de qualidade dos *reads* obtidos no processo de sequenciamento. O controle de qualidade se embasa na probabilidade de cada base sequenciada ter sido acuradamente identificada a partir de um valor conhecido como qualidade *phred*.

A expressão para o cálculo de qualidade *phred* é:

$$Q = -10\log 10\left(P\right) \qquad \textbf{(Equação 1)}$$

Onde *Q* se refere ao valor de qualidade *phred* e *P*, a probabilidade do valor obtido condizer com a realidade da base observada. Conforme o valor de qualidade aumenta, menores são as chances de que a base foi erroneamente identificada (**Tabela 1**).

Tabela 1. Valores de qualidade *phred* e suas probabilidades de erro em identificação de bases nucleotídicas durante o sequenciamento.

QUALIDADE *PHRED*	PROBABILIDADE DE ERRO	ACURÁCIA (%)
10	1 a cada 10 bases	90
20	1 a cada 100 bases	99
30	1 a cada 1.000 bases	99,9
40	1 a cada 10.000 bases	99,99
50	1 a cada 100.000 bases	99,999

Fonte: ILLUMINA, 2011

1.5 ANOTAÇÃO DE GENOMAS E PSEUDOGENES

O crescimento do número de genomas completos e dados de sequenciamento gerados levou ao aumento de componentes celulares identificados (REED et al., 2006). Informações a respeito de funções de componentes celulares, suas interações e alterações decorrentes do processo evolutivo podem ser interpretadas em decorrência do processo biológico em que estão envolvidos e podem ser representadas por diferentes tipos de anotações funcionais (REED et al., 2006). Entretanto, anotações de genomas podem estar relacionadas tanto com anotações de genes e proteínas quanto com anotações funcionais envolvendo processos biológicos (REED et al., 2006; STEIN, 2001).

Dois pontos são importantes para a anotação de um genoma: o mapa genético e a predição de genes (STEIN, 2001). Mapas genéticos possibilitam a inferência da localização de genes dentro de um genoma – a partir de informações físicas obtidas de uma referência, como tamanho de cromossomos – enquanto a predição de genes está mais relacionada com a identificação dos genes em si. A anotação de genomas está relacionada com a identificação de elementos funcionais presentes no genoma – normalmente genes codificantes – e a atribuição de funções, conhecidas ou preditas, de seus produtos (ARMSTRONG et al., 2019; REED et al., 2006).

Existem duas diferentes formas para realizar anotações gênicas: predição *ab initio* ou abordagens que envolvem alinhamento entre sequências (ARMSTRONG et al., 2019). Métodos de predição *ab initio* foram os primeiros a serem utilizados por, inicialmente, não necessitarem de informações externas às sequências de DNA, como informações relacionadas ao alinhamento entre proteínas codificadas contra um genoma de referência (YANDELL; ENCE, 2012). Porém, preditores gênicos atuais podem utilizar um conjunto de dados de treinamento criado a partir de sequências de genes conhecidos – obtendo, assim, maiores evidências biológicas – para melhoria dos resultados de predição (EJIGU; JUNG, 2020; YANDELL; ENCE, 2012). Já abordagens que envolvem transferência de anotação entre sequências homólogas providenciam maiores evidências de que certa região pertence a um gene (STEIN, 2001). Entretanto, métodos dependentes de homologia ficam limitados

a sequências de genes e proteínas conhecidas e descritas, já que é necessário que exista similaridade entre a sequência e os dados presentes em bancos de dados (MATHE, 2002).

Genes funcionais que sofrem alterações e se tornam não-codificantes podem ser classificados como pseudogenes. A perda de conservação de função acarreta uma menor pressão seletiva que permite a acumulação de alterações – como mutações, inserções e deleções – no decorrer do processo evolutivo (SALMENA, 2021). No entanto, pseudogenes podem estar relacionados com a subclasse de RNAs longos não-codificantes ou RNAs interferentes pequenos, que podem apresentar papel regulatório na transcrição de RNAs codificantes (SALMENA, 2021; TUTAR, 2012). Além disso, pseudogenes podem agir como reservatórios de diversidade genética, caso os efeitos cumulativos de mutações resultem novamente em um gene codificante (BALAKIREV; AYALA, 2003).

1.6 PANGENOMA E GENÔMICA COMPARATIVA

O termo pangenoma (do grego παν, pan = totalidade) foi utilizado pela primeira vez por Tettelin e colaboradores (2005) em um estudo genômico envolvendo diferentes cepas de *Streptococcus agalactiae* e colocou um importante tópico em pauta: quantos genomas são necessários para caracterizar uma espécie por completo (CHAUDHARI; GUPTA; DUTTA, 2016; ZEKIC; HOLLEY; STOYE, 2018). Portanto, estudos de pangenoma se referem a estudos de genômica comparativa para organismos de uma mesma espécie. Inicialmente, o termo era utilizado apenas para o conjunto completo de genes de espécies microbianas, porém o conceito de pangenoma se expandiu para outros organismos e pode ser utilizado para: estudos evolutivos; análises genômicas comparativas; ou identificação de genes funcionais essenciais de patógenos, relacionados a patogenicidade ou resistência a drogas (CHAUDHARI; GUPTA; DUTTA, 2016; ZEKIC; HOLLEY; STOYE, 2018; ZHAO et al., 2012, 2018).

O pangenoma é constituído por três partes principais: (i) genoma central, (ii) genoma acessório, e (iii) *singletons* (MEDINI et al., 2005; TETTELIN et al., 2005; ZEKIC; HOLLEY; STOYE, 2018). O genoma central se refere aos genes conservados em todas as cepas, enquanto o genoma acessório se refere aos genes que não estão presentes em todas as cepas, porém estão compartilhados em parte delas. Neste sentido, genes centrais estão mais relacionados à aspectos biológicos fundamentais, enquanto genes acessórios à diversidade e vantagens evolutivas (BENTLEY, 2009; MEDINI et al., 2005; TETTELIN et al., 2005; VERNIKOS et al., 2015). Já os *singletons* se referem aos genes únicos, presentes em apenas uma cepa, também relacionados à vantagens evolutivas (CHAUDHARI; GUPTA; DUTTA, 2016; VERNIKOS et al., 2015). Pangenomas podem ser abertos ou fechados, dependendo de seu tamanho. Um pangenoma aberto continua a crescer conforme sequências são adicionadas, enquanto um pangenoma fechado é finito e não se altera com a adição de novas sequências, caracterizando de forma mais completa a espécie estudada (MEDINI et al., 2005; ZEKIC; HOLLEY; STOYE, 2018).

Como genomas de *Leishmania* são altamente sintênicos e conservados (DOWNING et al., 2011; PEACOCK et al., 2007; ROGERS et al., 2011; TEIXEIRA et al., 2017), uma abordagem tradicional de identificação de genoma central, genoma acessório e *singletons* não é aplicada. Ao invés disso, estudos comparativos envolvendo genomas de *Leishmania* são voltados para variação do número de cópias gênicas/cromossômicas, análise de polimorfismos para caracterização de parasitos e identificação de genes espécie-específicos que podem estar relacionados com virulência (ALMEIDA et al., 2021; BUTENKO et al., 2019; CARVALHO et al., 2020; TEIXEIRA et al., 2017; VALDIVIA et al., 2017).

1.7 HIPÓTESE

Genomas altamente sintênicos e conservados de *Leishmania infantum* podem apresentar alterações específicas entre amostras de diferentes regiões e hospedeiros.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Identificar alterações no genoma de *L. infantum* que diferem entre isolados obtidos em Santa Catarina e Rio Grande do Norte a partir de uma abordagem comparativa.

2.2 OBJETIVOS ESPECÍFICOS

- Montar e anotar genomas de isolados de *L. infantum* de Santa Catarina e do Rio Grande do Norte;
- Identificar regiões gênicas conservadas e não conservadas no pangenoma de *L. infantum* de cepas isoladas em Santa Catarina e Rio Grande do Norte;
- Identificar polimorfismos de nucleotídeo único com base em hospedeiro ou região.

3 MATERIAIS E MÉTODOS

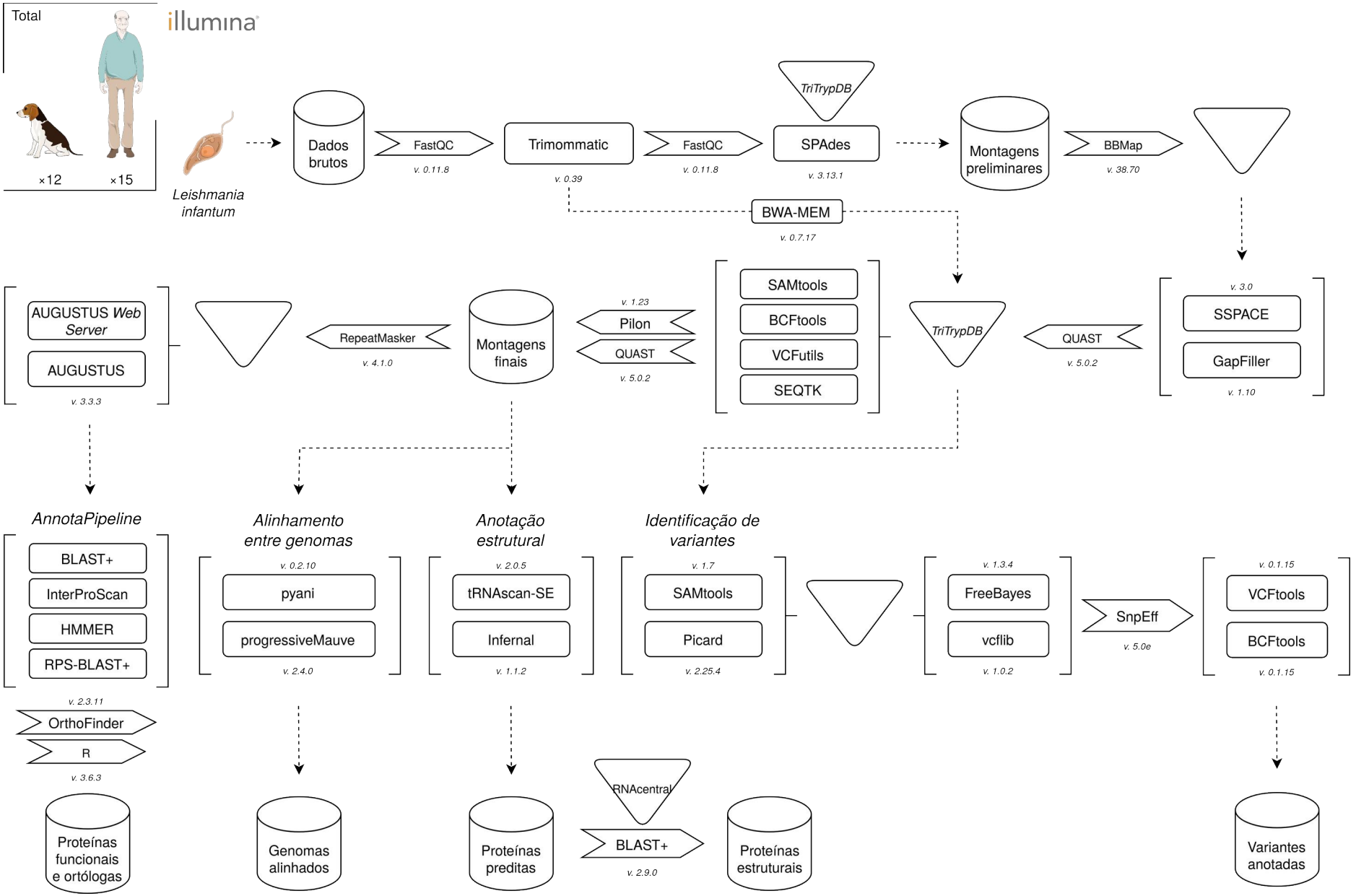
As análises desenvolvidas neste trabalho foram realizadas no Laboratório de Bioinformática do Departamento de Microbiologia, Imunologia e Parasitologia (MIP) – no Centro de Ciências Biológicas (CCB) – da Universidade Federal de Santa Catarina (UFSC). A estrutura do laboratório consta de servidores virtuais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

Os experimentos computacionais foram realizados em Servidores Dell que dispõem de 40 núcleos de processadores (3.2 GHz), 320 GB de RAM (DDR4, 2400 MHz) e armazenamento de 5 TB (HDD SATA 2.5” 7200 RPM).

3.1 DELINEAMENTO EXPERIMENTAL

O delineamento experimental das análises realizadas – incluindo todos os programas utilizados – se encontra na **Figura 2**. As etapas ilustradas correspondem a: (i) obtenção de dados brutos de sequenciamento de dados genômicos, (ii) controle de qualidade, (iii) montagens preliminares de genomas, (iv) refinamento das montagens, (v) ordenação em cromossomos, (vi) predições gênicas, (vii) anotações funcionais e estruturais, (viii) inferência de ploidia, (ix) ontologia genética, (x) ortologia, (xi) identidade entre genomas, e (xii) identificação de variantes.

Figura 2. Desenho experimental para análises de montagem, anotação e genômica comparativa de *Leishmania infantum*.



3.2 OBTENÇÃO DE DADOS

Ao todo, foram utilizadas 27 amostras de *L. infantum*, sendo oito provenientes de Santa Catarina e 19 provenientes do Rio Grande do Norte. Amostras oriundas de canídeos foram identificadas pela letra *D*, enquanto amostras humanas foram identificadas pela letra *H*. Para canídeos, sete amostras foram obtidas em Santa Catarina e cinco no Rio Grande do Norte. Já para humanos, uma amostra foi obtida em Santa Catarina e 14 no Rio Grande do Norte.

Dados de sequenciamento *paired-end* de isolados de *L. infantum* obtidos de humanos e canídeos infectados, oriundos de Santa Catarina, foram gentilmente cedidos pela Profª. Drª. Patrícia Hermes Stoco, pelo Prof. Dr. Edmundo Carlos Grisard – ambos do Laboratório de Protozoologia (MIP/CCB/UFSC) – e pelo Prof. Dr. Bjorn Andersson (Karolinska Institute, Suécia).

Dados brutos de sequenciamento genômico também foram obtidos do banco de dados *Sequence Read Archive* (SRA) (LEINONEN et al., 2011; TEIXEIRA et al., 2017) – através do programa *fastq-dump* v2.8.2 da ferramenta *SRA Toolkit* – a partir dos códigos de acesso presentes na **Tabela 2**.

Dados de referência da cepa JPCM5 de *L. infantum* (GONZÁLEZ-DE LA FUENTE et al., 2017) foram obtidos da versão 46 do banco de dados de acesso público voltado para dados genômicos de tripanosomatídeos, o *TriTrypDB* (ASLETT et al., 2010).

Tabela 2. Caracterização de isolados de *Leishmania infantum* obtidos em Santa Catarina e do *Sequence Read Archive* (SRA).

IDENTIFICADOR	ORGANISMO INFECTADO	CÓDIGO DE ACESSO	REGIÃO	ESTADO	ORIGEM	ANO DE COLETA	TAMANHO DE *READS* (pb)	SEQUENCIAMENTO (Illumina)
D1	Canídeo	—	Canto dos Araças (FLN)	Santa Catarina	—	2010	126	HiSeq 2500
D2	Canídeo	—	Lagoa da Conceição (FLN)	Santa Catarina	—	2015	126	HiSeq 2500
D3	Canídeo	—	Itaguaçu (FLN)	Santa Catarina	—	2016	126	HiSeq 2500
D4	Canídeo	—	Chapecó	Santa Catarina	—	2018	151	NovaSeq 6000
D5	Canídeo	—	Córrego Grande (FLN)	Santa Catarina	—	2018	151	NovaSeq 6000
D6	Canídeo	—	Córrego Grande (FLN)	Santa Catarina	—	2018	151	NovaSeq 6000
D7	Canídeo	—	Pantanal (FLN)	Santa Catarina	—	2018	151	NovaSeq 6000
H1	Humano	—	Florianópolis	Santa Catarina	—	2017	126	HiSeq 2500
1VLh90	Humano	SRR5117900	Natal	Rio Grande do Norte	Medula óssea	1991	101	HiSeq 2000
2VLh90	Humano	SRR5117911	Ielmo Marinho	Rio Grande do Norte	Medula óssea	1992	101	HiSeq 2000
3VLh90	Humano	SRR5117909	Macaíba	Rio Grande do Norte	Medula óssea	1992	101	HiSeq 2000
4VLh90	Humano	SRR5117907	Ceará-Mirim	Rio Grande do Norte	Medula óssea	1993	101	HiSeq 2000
5VLh90	Humano	SRR5117905	São José de Mipibu	Rio Grande do Norte	Medula óssea	1993	101	HiSeq 2000

6CLh	Humano	SRR5117903	Macaíba	Rio Grande do Norte	Pele	2009	101	HiSeq 2000
12VLh	Humano	SRR5117906	Extremoz	Rio Grande do Norte	Medula óssea	2012	101	HiSeq 2000
13VLh	Humano	SRR5117908	Açu	Rio Grande do Norte	Medula óssea	2012	101	HiSeq 2000
14VLh	Humano	SRR5117895	Macaíba	Rio Grande do Norte	Medula óssea	2012	101	HiSeq 2000
19VLh*	Humano	SRR5117897	Sítio Novo	Rio Grande do Norte	Sangue periférico	2013	101	HiSeq 2000
20VLh*	Humano	SRR5117894	Sítio Novo	Rio Grande do Norte	Medula óssea	2013	101	HiSeq 2000
8Ah	Humano	SRR5117901	Natal	Rio Grande do Norte	Sangue periférico	2011	101	HiSeq 2000
9Ah	Humano	SRR5117902	Touros	Rio Grande do Norte	Sangue periférico	2011	101	HiSeq 2000
18Ah	Humano	SRR5117904	Natal	Rio Grande do Norte	Sangue periférico	2012	101	HiSeq 2000
7VLd	Canídeo	SRR5117896	Natal	Rio Grande do Norte	Baço	2010	101	HiSeq 2000
11VLd	Canídeo	SRR5117898	Natal	Rio Grande do Norte	Baço	2011	101	HiSeq 2000
15VLd	Canídeo	SRR5117912	Natal	Rio Grande do Norte	Baço	2012	101	HiSeq 2000
16VLd	Canídeo	SRR5117893	Natal	Rio Grande do Norte	Baço	2012	101	HiSeq 2000
17VLd	Canídeo	SRR5117910	Natal	Rio Grande do Norte	Baço	2012	101	HiSeq 2000

* Amostras obtidas do mesmo indivíduo. FLN: Florianópolis.

3.3 CONTROLE DE QUALIDADE

Os dados brutos obtidos foram submetidos à etapa de controle de qualidade, visando a remoção de pares de base com qualidade *phred* abaixo de um limiar estipulado. O valor mínimo de qualidade foi determinado a partir dos valores observados através dos programas *FastQC* v0.11.8 (ANDREWS, 2010) e *MultiQC* v1.11 (EWELS et al., 2016).

A remoção de bases foi realizada pelo programa *Trimmomatic* v0.39 (BOLGER; LOHSE; USADEL, 2014), sendo que nucleotídeos abaixo do valor de qualidade *phred* esperado são substituídos por bases *N* e *reads* que apresentam inúmeras substituições são removidos do conjunto de dados (LIAO; SATTEN; HU, 2017). Os valores de corte estipulados foram: qualidade *phred* média de 20 (`AVGQUAL:20`), qualidade de bases iniciais de 30 (`LEADING:30`), e qualidade de bases finais de 25 (`TRAILING:25`). O programa também foi utilizado para remoção de adaptadores inerentes a técnica de sequenciamento empregada (`ILLUMINACLIP:all_adapters.fa`), identificados no **Apêndice 1**.

Para as amostras que apresentaram maior cobertura, os parâmetros de remoção de bases gerais (`SLIDINGWINDOW:4:30`) e finais (`TRAILING:30`) foram alterados, assim como a adição de tamanho mínimo para manter fragmentos de sequências limpas (`MINLEN:75`). O método de remoção a partir de janelas de leitura acaba por ser mais rigoroso que o método de qualidade média, sendo fundamentado em janelas intercaladas a cada quatro nucleotídeos que necessitam apresentar qualidade média de 30 para não serem substituídos. As alterações foram realizadas em decorrência da abundância de *reads* disponíveis, que permitiram um controle de qualidade mais rigoroso sem o risco de perda significativa de dados.

O cálculo de cobertura é dependente de um genoma de referência, neste caso a cepa JPCM5 de *L. infantum*, e pode ser obtido através da seguinte equação:

$$cobertura = \frac{total\ de\ reads\left(\times 2\ para\ paired-end\right)\times tamanho\ médio\ dos\ reads}{\left(tamanho\ do\ genoma\ de\ referência\right)} \quad \textbf{(Equação 2)}$$

A cobertura dos genomas foi inferida pelo algoritmo *covstats* do programa *BBMap* v38.70 (BUSHNELL, 2014), que realiza o alinhamento de *reads* para cada amostra contra o genoma de referência.

Os valores mínimos de qualidade e o método de remoção de bases se encontram na **Tabela 3**.

Tabela 3. Parâmetros de controle de qualidade e cobertura média utilizados para limpeza de dados brutos de *Leishmania infantum*.

AMOSTRA	LOCAL	QUALIDADE MÍNIMA	REMOÇÃO DE BASES	COBERTURA MÉDIA APÓS LIMPEZA
D1	SC	20	Qualidade média	111,37
D2	SC	20	Qualidade média	103,05
D3	SC	20	Qualidade média	156,4
D4	SC	30	Janela de leitura	1694,98
D5	SC	30	Janela de leitura	1205,82
D6	SC	30	Janela de leitura	980,43
D7	SC	30	Janela de leitura	1422,4
H1	SC	20	Qualidade média	125,24
16VLd	RN	20	Qualidade média	170,53
20VLh	RN	20	Qualidade média	157,74
14VLh	RN	20	Qualidade média	177,24
7VLd	RN	20	Qualidade média	178,25
19VLh	RN	20	Qualidade média	194,8
11VLd	RN	20	Qualidade média	144,7
1VLh90	RN	20	Qualidade média	180,2
8Ah	RN	20	Qualidade média	204,54
9Ah	RN	20	Qualidade média	180,86
6CLh	RN	20	Qualidade média	161,74
18Ah	RN	20	Qualidade média	178,12
5VLh90	RN	20	Qualidade média	183,8

AMOSTRA	LOCAL	QUALIDADE MÍNIMA	REMOÇÃO DE BASES	COBERTURA MÉDIA APÓS LIMPEZA
12VLh	RN	20	Qualidade média	144,88
4VLh90	RN	20	Qualidade média	183,52
13VLh	RN	20	Qualidade média	166,56
3VLh90	RN	20	Qualidade média	183,35
17VLd	RN	20	Qualidade média	191,13
2VLh90	RN	20	Qualidade média	179,36
15VLd	RN	20	Qualidade média	160,84

SC: Santa Catarina. RN: Rio Grande do Norte.

3.4 MONTAGEM DE GENOMAS

Todas as montagens realizadas neste projeto foram pareadas ao genoma de referência de *L. infantum* JPCM5 obtido no *TriTrypDB* v46 (`--trusted-contigs`). O montador utilizado foi o *SPAdes* v3.13 (BANKEVICH et al., 2012) com os parâmetros: redução de *mismatches* e *indels* (`--careful`), cálculo de cobertura mínima automático (`--cov-cutoff auto`), e tamanho de *k-mer* de 75 (`-k 75`).

Para as amostras com maior cobertura, foram utilizados os programas *SEQTK* v1.2-r94 (LI, 2013a) e *BBNorm* v38.70 (BUSHNELL, 2014) para subamostragem de *reads* e normalização de cobertura, respectivamente. Subamostragens randômicas de 20% dos *reads* totais foram selecionadas e normalizadas em uma cobertura média de 100 vezes antes de serem submetidas ao *SPAdes*, já que coberturas acima deste valor necessitam de maior capacidade de processamento sem melhoras significativas no processo de montagem (DESAI et al., 2013).

Todas as montagens obtidas foram submetidas ao algoritmo *ihist* do programa *BBMap* v38.70 para inferência de tamanhos de inserção – inerentes de sequenciamentos *paired-end* a partir da plataforma Illumina. Os tamanhos de inserção foram adicionados em bibliotecas utilizadas para o refinamento das montagens através dos programas *SSPACE* v3.0 (BOETZER et al., 2011) – para

remoção de fragmentos menores que 1000 pares de base (`-z 1000`) – e *GapFiller* v1-10 (NADALIN; VEZZI; POLICRITI, 2012), para preenchimento de espaços entre *scaffolds*. Para o preenchimento de espaços foram consideradas: (i) interpolações mínimas de 10 nucleotídeos para junção de *scaffolds* adjacentes (`-n 10`), (ii) remoção de 10 nucleotídeos em cada extremidade das sequências que apresentam baixa cobertura (`-t 10`) e (iii) diferença máxima de 50 nucleotídeos entre espaços fechados e nucleotídeos adicionados (`-d 50`), no decorrer de 10 iterações (`-i 10`).

As montagens refinadas foram alinhadas ao genoma de referência (*L. infantum* JPCM5) pelo alinhador *BWA-MEM* v0.7.17 (LI, 2013b) e ordenadas pelo programa *SAMtools* v1.7 (LI et al., 2009), gerando arquivos de alinhamento que relacionam *scaffolds* montados à posições dentro do genoma de referência. A partir dos arquivos de alinhamento, sequências consenso de cromossomos – baseadas na cepa JPCM5 de *L. infantum* – foram obtidas através dos seguintes programas: (i) módulo *mpileup* do programa *SAMtools* v1.7, (ii) módulo *call* (`--consensus-caller`) e *script vcfutils.pl* (`vcf2fq`) do programa *BCFtools* v1.7 (DANECEK et al., 2021), e (iii) *SEQTK* v1.2-r94 configurado para o *soft mask* de pares de base com qualidade de alinhamento *phred* menores que 20 (`seq -q20`).

Os cromossomos obtidos foram polidos – substituindo nucleotídeos erroneamente classificados ou não identificados – pelo programa *Pilon* v1.23 (WALKER et al., 2014) configurado para organismos diploides (`--diploid`) com qualidade *phred* média de 36 (`--defaultqual 36`).

Por fim, todas as montagens foram avaliadas pelo programa *QUAST* v5.0.2 (GUREVICH et al., 2013) para obtenção de suas métricas.

3.5 PREDIÇÃO E ANOTAÇÃO GÊNICA

Antes da etapa de predição gênica e anotação funcional, todos os genomas montados passaram pelo programa *RepeatMasker* v4.1.0 (SMIT; HUBLEY; GREEN, 2015) para identificação de regiões repetitivas sem checagens de inserções bacterianas (`-no_is`). Repetições foram identificadas pela ferramenta `rmblast` e

sofreram *soft mask* (`-xsmall`), que retorna em caracteres minúsculos as possíveis repetições.

O preditor gênico utilizado – para predição de genes e regiões funcionais – foi o *AUGUSTUS* v3.3.3 (STANKE; WAACK, 2003). O modelo preditivo requisitado pelo programa foi gerado através da plataforma *WebAUGUSTUS* (HOFF; STANKE, 2013) a partir de um conjunto de treino formado pelo genoma completo e proteínas anotadas de *L. infantum* da cepa JPCM5.

A etapa de anotação gênica foi realizada através da busca por similaridades contra os bancos de dados *SwissProt* (obtido em 11/2019) (BAIROCH, 1996) e *TriTrypDB* v46. Já a etapa de anotação funcional foi realizada através dos programas *InterProScan* v5.45-80.0 (JONES et al., 2014), *HMMER* v3.1b2 pela ferramenta *hmmscan* (FINN; CLEMENTS; EDDY, 2011; POTTER et al., 2018) e *RPS-BLAST+* v2.9.0 (CAMACHO et al., 2009). Ambas etapas de anotação foram desempenhadas pelo *AnnotaPipeline* (MAIA, 2019) – desenvolvido pelo Laboratório de Bioinformática (UFSC) – com parâmetros padrões.

3.6 PREDIÇÃO DE ELEMENTOS ESTRUTURAIS E PSEUDOGENES

Os elementos estruturais preditos foram RNAs transportadores (tRNA) e RNAs ribossomais (rRNA). A predição de tRNAs foi realizada pelo programa *tRNAscan-SE* v2.0.5 (LOWE; EDDY, 1997), enquanto a predição de rRNAs foi realizada pelo algoritmo *cmscan* do programa *Infernal* v1.1.2 (NAWROCKI; EDDY, 2013), que se baseia na homologia correspondente ao banco de dados *Rfam* v14.2 (KALVARI et al., 2018), de RNAs não-codificantes. Os parâmetros utilizados pelo algoritmo *cmscan* foram: (i) modelos de covariância (CM) em conjunto 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 sequência homóloga (`--cut_ga`), e (iv) dados do *Rfam* – preparados pelo algoritmo *cmpress* – como informações de agrupamentos (`--clanin`).

Todos os elementos estruturais preditos foram validados *in silico* com base em níveis de similaridade e cobertura obtidos a partir do algoritmo *BLASTn* do programa *BLAST+* (CAMACHO et al., 2009). tRNAs e rRNAs que apresentaram níveis de identidade e cobertura acima de 90% contra sequências de *L. infantum* adquiridas do banco de dados *RNAcentral* (RNACENTRAL CONSORTIUM et al., 2019) – de RNAs não-codificantes – foram considerados válidos. O número de sequências de rRNA e tRNA obtidas do *RNAcentral* foram de 108 e 78, respectivamente.

A predição de pseudogenes foi realizada pelo programa *PseudoPipe* (ZHANG et al., 2006) – com configurações padrões – e utilizou os resultados da predição gênica resultantes do *AUGUSTUS* v3.3.3. A localização de exons para cada proteína predita, assim como as montagens cromossômicas, foram utilizadas como entrada para a identificação de possíveis pseudogenes.

A partir das anotações obtidas pela etapa de predição e anotação gênica, a transferência de anotação foi realizada para os possíveis pseudogenes e uma nuvem de palavras foi gerada pelo pacote *wordcloud* v2.6 (FELLOWS, 2018) da linguagem *R* v4.0.5 (R CORE TEAM, 2021) para identificação dos termos mais prevalentes que foram afetados. O termo "*membrane associated protein-like*" foi o mais prevalente e foi removido antes da geração da imagem por apresentar uma proporção dez vezes maior que o segundo termo mais prevalente. Os termos "*putative*" e "*containing protein*" também foram removidos antes da geração da imagem.

3.7 ONTOLOGIA DE GENOMAS ANOTADOS

A partir dos resultados de anotação funcional gerados pelo *AnnotaPipeline*, as informações referentes a termos de ontologia foram filtradas e submetidas a plataforma *WEGO 2.0* (YE et al., 2018) tanto para a referência quanto para as amostras utilizadas neste estudo, a fim de comparar a similaridade entre ontologias do Velho e Novo Mundo.

Os termos de ontologia filtrados também foram separados em grupos de cães e humanos tanto para amostras de Santa Catarina quanto do Rio Grande do Norte. Para cada grupo, os termos ontológicos de cada amostra foram concatenados e suas redundâncias foram removidas para submissão na ferramenta *online jvenn* (BARDOU et al., 2014). Os termos ontológicos presentes em interpolações do diagrama de Venn que não continham todos os grupos foram selecionados e utilizados para criação de *tree maps* na ferramenta *online Revigo* (SUPEK et al., 2011) – com configurações padrões – para visualização de suas ontologias.

3.8 ORTOLOGIA

As proteínas obtidas dos processos de predição e anotação gênica foram submetidas a análise de ortologia pelo programa *OrthoFinder* v2.3.11 (EMMS; KELLY, 2019), com alteração de *e-value* do programa `diamond blastp` para `1e-5`.

Os resultados de ortogrupos – assim como suas respectivas contagens – gerados pelo *OrthoFinder* foram manipulados pela biblioteca *pandas* v1.2.1 (*Python* v3.8.10) (MCKINNEY, 2010) para adição de possíveis anotações para cada ortogrupo e remoção de elementos com possíveis anotações hipotéticas.

O processo de anotação de proteínas ortólogas se baseou na preservação de função para proteínas de um mesmo ortogrupo. Com isso, a anotação da última proteína presente em cada grupo – para todas as amostras – foi utilizada para descrever todas suas proteínas ortólogas.

Ortogrupos gerados que apresentaram contagem zero para alguma amostra foram utilizados para determinar possíveis proteínas não-essenciais da espécie. *Heatmaps* foram gerados a partir das matrizes de contagem zero através dos pacotes *ComplexHeatmap* v2.6.2 (GU; EILS; SCHLESNER, 2016) e *ggplot2* v3.3.5 (WICKHAM, 2016) da linguagem *R* v3.6.3 (R CORE TEAM, 2020).

3.9 ALINHAMENTO MÚLTIPLO DE GENOMAS

O alinhamento múltiplo foi realizado para as montagens em nível cromossômico pelo alinhador *progressiveMauve* v2.4.0 (DARLING; MAU; PERNA, 2010) com parâmetros padrões. Os alinhamentos foram divididos, com base em região de coleta, em quatro grupos para melhor visualização dos resultados: (i) amostras de Santa Catarina, e para o Rio Grande do Norte, (ii) amostras humanas dos anos 90, (iii) amostras humanas, e (iv) amostras de cães.

3.10 IDENTIDADE E AGRUPAMENTO ENTRE GENOMAS

As montagens em nível cromossômico foram submetidas ao módulo *pyani* v0.2.10 (*Python* v3.6.9) (PRITCHARD et al., 2016) para mensurar e criar uma matriz de dissimilaridade referente à identidade entre os genomas montados.

A matriz de dissimilaridade foi utilizada para gerar tanto um *heatmap* ilustrando a porcentagem de identidade quanto o escalonamento multidimensional (MDS, do inglês, *multidimensional scaling*) entre as amostras.

O *heatmap* foi gerado através do pacote *ComplexHeatmap* v2.6.2 (GU; EILS; SCHLESNER, 2016), enquanto o MDS foi inferido com base na matriz de dissimilaridade a partir de distâncias euclidianas e foi gerado pelo pacote *ggplot2* v3.3.5, ambos da linguagem *R* v4.0.5.

3.11 IDENTIFICAÇÃO DE VARIANTES

Os *reads* obtidos pela etapa de controle de qualidade foram alinhados ao genoma de referência pelo alinhador *BWA-MEM* v0.7.17 e ordenados pelo módulo *sort* do programa *SAMtools* v1.7. Os arquivos ordenados de alinhamento foram então processados pela ferramenta *Picard Tools* v2.25.4 (BROAD INSTITUTE, 2019) para identificação e remoção de duplicatas (`MarkDuplicates`) e criação de grupos para os *reads* mapeados (`AddOrReplaceReadGroups`).

Os conjuntos de dados processados foram subdivididos de acordo com o tipo de hospedeiro, humano ou canídeo, e região de coleta, Santa Catarina ou Rio

Grande do Norte, em quatro grupos. Os conjuntos subdivididos foram utilizados pelo programa de detecção de variantes *FreeBayes* v1.3.4 (GARRISON; MARTH, 2012) para organismos diploides.

Os arquivos de variantes passaram por uma etapa de controle de qualidade para remoção de alinhamentos de baixa qualidade e que não apresentaram correspondência em ambas as fitas (`QUAL > 1 & QUAL / AO > 10 & SAF > 0 & SAR > 0 & RPR > 1 & RPL > 1`) pelo comando *filter* da biblioteca *vcflib* v1.0.2 (GARRISON et al., 2021).

As variantes filtradas foram anotadas pelo programa *SnpEff* v5.0e (CINGOLANI et al., 2012) – a partir de um modelo criado pelo genoma de referência – para identificação de possíveis efeitos decorrentes dos polimorfismos identificados.

Os polimorfismos de nucleotídeo único (SNP) anotados de cada grupo foram utilizados pelo módulo *isec* do programa *BCFtools* v0.1.15 (DANECEK et al., 2021) para identificação de interpolações de SNPs entre os diferentes grupos. As interpolações geradas foram parseadas a fim de associar informações referentes aos genes afetados e suas possíveis funções.

A matriz de interpolações e possíveis descrições foi então submetida ao pacote *tidyverse* v1.3.1 (WICKHAM et al., 2019) da linguagem *R* v4.0.5 para a contagem de SNPs ao longo de cada cromossomo.

3.12 NÚMERO DE CÓPIAS CROMOSSÔMICAS

As metodologias descritas tanto por Zhang e colaboradores (2014) quanto por Carvalho e colaboradores (2020) foram empregadas para inferência de ploidia. Portanto, os *reads* mapeados e ordenados obtidos na etapa anterior foram submetidos ao módulo *depth* do programa *SAMtools* v1.7 para a inferência de cobertura para cada base observada.

Para cada amostra, a ploidia foi calculada através da mediana de cobertura ($M_d$chr) para cada cromossomo pelas seguintes equações:

$$d_t = M_d(M_d chr\,1 \ldots M_d chr\,36) \qquad \textbf{(Equação 3)}$$

$$ploidia_{chr} = \frac{(M_d chr)}{(d_t \div 2)} \qquad \textbf{(Equação 4)}$$

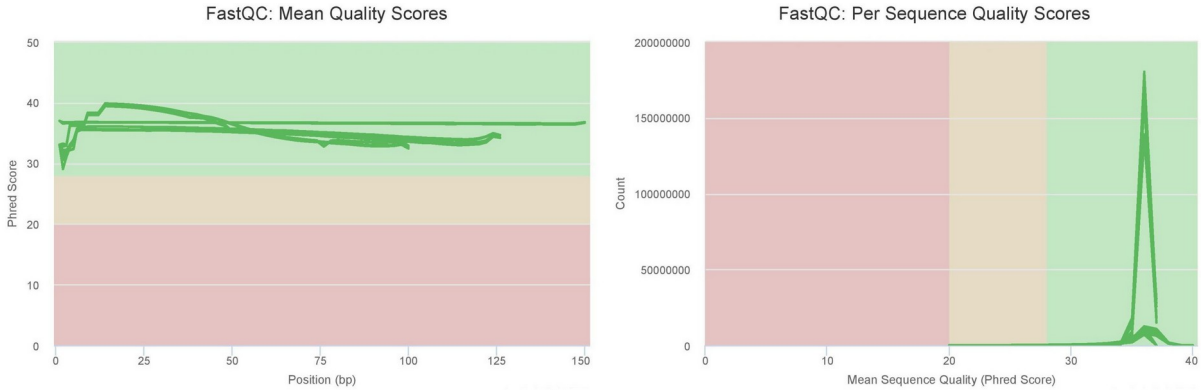
Uma matriz representando o número de cópias cromossômicas – para cada cromossomo em cada amostra – foi criada e utilizada para a geração de um *heatmap* através do pacote *ComplexHeatmap* v2.6.2 da linguagem *R* v4.0.5.

4 RESULTADOS E DISCUSSÃO

4.1 QUALIDADE DOS DADOS UTILIZADOS

Após a etapa de controle de qualidade, todas as amostras apresentaram valores de qualidade *phred* acima de 30 (**Figura 3**) – com qualidade média de 36 – para toda a extensão dos *reads*. A avaliação de qualidade foi realizada pelo programa *MultiQC* v1.11, que agrupou os resultados obtidos pelo programa *FastQC* v0.11.8.

Figura 3. Valores de qualidade *phred* obtidos pelo programa *MultiQC* após etapa de controle de qualidade para dados brutos de *Leishmania infantum*.



A **Tabela 4** apresenta o tamanho médio dos *reads* após a etapa de controle de qualidade. O tamanho dos *reads* influencia diretamente a escolha do tamanho de *k*-mers para a etapa de montagem de *short reads* utilizando grafos de *de Bruijn*. Na etapa de montagem, *k-mers* de tamanhos menores impossibilitam que regiões com maior número de repetições – caso sejam maiores que as *substrings* – sejam incorporadas no momento de interpolação dos caracteres de tamanho *k* (CHIKHI; MEDVEDEV, 2014; SHARIAT et al., 2014). Entretanto, tamanhos maiores abrem margem para que erros de *base calling* sejam perpetuados, gerando uma montagem menos fidedigna (CHIKHI; MEDVEDEV, 2014).

O tamanho médio de *reads* para todas as amostras variou entre 98 pb e 144 pb. Por isso, para as montagens realizadas neste trabalho, o tamanho de *k* escolhido foi de 75. Este valor se enquadra entre 76,53% e 52,08% dos tamanhos dos *reads*, sem que as *substrings* geradas cheguem em valores extremos.

Tabela 4. Total de sequências remanescentes e tamanho médio de *reads* para dados de *Leishmania infantum* após processamento pelo programa *Trimmomatic*.

AMOSTRA	TAMANHO MÉDIO (pb)	TOTAL DE SEQUÊNCIAS INICIAIS (Milhões)	TOTAL DE SEQUÊNCIAS REMANESCENTES (Milhões)
D1	125	17,1	15,1
D2	125	15,5	13,6
D3	125	23,7	20,8
D4	144	465,7	199,9
D5	143	483,6	217,6
D6	143	440	193,8
D7	144	404,5	168,5
H1	125	20	17,1
11VLd	98	25,9	25,4
12VLh	98	27,7	27,1
13VLh	98	31,7	30,9
14VLh	98	31,4	30,9
15VLd	98	30,2	29,6
16VLd	98	29,8	29,4
17VLd	98	33,5	33
18Ah	98	31,9	31,4
19VLh	98	35,3	34,7
1VLh90	98	34,8	34,2
20VLh	98	29,8	29
2VLh90	98	31,5	31
3VLh90	98	31,8	31,1
4VLh90	98	33,2	32,6
5VLh90	98	33,6	32,8
6CLh	98	29,2	28,6

AMOSTRA	TAMANHO MÉDIO (pb)	TOTAL DE SEQUÊNCIAS INICIAIS (Milhões)	TOTAL DE SEQUÊNCIAS REMANESCENTES (Milhões)
7VLd	98	33,1	32,5
8Ah	98	36,2	35,6
9Ah	98	31,8	31,3

4.2 MONTAGEM, ANOTAÇÃO E ONTOLOGIA DE GENOMAS

Em termos comparativos, o genoma de referência de *L. infantum* apresenta 36 *scaffolds* – sendo que cada *scaffold* representa um cromossomo da espécie – e um total de 32.802.969 pb, sem a presença de nucleotídeos não identificados. Um panorama geral sobre as métricas de montagem pode ser encontrado na **Tabela 5**. Devido a metodologia de montagem empregada neste projeto, as montagens apresentaram uma proporção de 1:1 entre *scaffolds* e cromossomos esperados da espécie, além de um tamanho próximo ao genoma de referência. Entretanto, o alto nível de contiguidade só foi possível devido ao nível de qualidade do genoma de *L. infantum* utilizado como referência.

Todas as montagens finalizadas apresentaram 36 *scaffolds* – chegando assim em nível cromossômico – com uma média total de 32.647.077 pb (32,6 Mb). Porém, todas as montagens apresentaram uma média geral de 2.790,49 nucleotídeos não identificados a cada 100 kpb. Uma possibilidade para o número de nucleotídeos não identificados são os tipos de dados brutos utilizados para a montagem. Dados gerados pela plataforma Illumina são caracterizados pelo seu tamanho, entre 100 e 300 pb, e sua taxa de erro, até 2% (HEYDARI et al., 2019; STOLER; NEKRUTENKO, 2021). Por isso, montagens que utilizam somente *short reads* podem sofrer compressões, resultantes de sequências repetitivas que se perdem – ou se fragmentam – por não se enquadrar dentro do tamanho dos *reads* (ADEWALE, 2020; ALKAN; SAJJADIAN; EICHLER, 2011; BAPTISTA; KISSINGER, 2019; GONZÁLEZ-DE LA FUENTE et al., 2017).

Como a etapa final do processo de montagem foi a de ordenação de *scaffolds* em nível cromossômico, durante o mapeamento de *scaffolds* no genoma de referência, as regiões repetitivas podem ter sido preenchidas por *N*, inflando assim o número de nucleotídeos não identificados.

Além disso, cada genoma apresenta uma cobertura de no mínimo 100 vezes (**Tabela 3**), o que está acima do recomendado para montagens de pequenos genomas (DESAI et al., 2013).

Tabela 5. Métricas básicas após processo de montagem de genomas de *Leishmania infantum* utilizando dados de sequenciamento de segunda geração (Illumina).

AMOSTRA	*SCAFFOLDS*	TOTAL DE BASES	CONTEÚDO GC (%)	NÚMERO DE N (a cada 100 kpb)
JPCM5 (Referência)	36	32.802.969	59,74	0
D1	36	32.644.853	59,59	2.920,52
D2	36	32.599.614	59,59	2.855,06
D3	36	32.645.597	59,61	2.691,73
D4	36	32.673.182	59,57	3.363,42
D5	36	32.706.954	59,56	3.664,89
D6	36	32.685.114	59,57	3.557,97
D7	36	32.691.955	59,56	3.560,61
H1	36	32.590.745	59,6	2.593,53
11VLd	36	32.564.616	59,6	2.607,89
12VLh	36	32.627.917	59,61	2.565,26
13VLh	36	32.592.420	59,6	2.571,84
14VLh	36	32.677.639	59,6	2.673,75
15VLd	36	32.649.856	59,6	2.597,84
16VLd	36	32.661.487	59,6	2.790,17
17VLd	36	32.637.122	59,6	2.525,97
18Ah	36	32.699.936	59,59	2.996,23
19VLh	36	32.603.513	59,61	2.477,54
1VLh90	36	32.663.143	59,6	2.685,82
20VLh	36	32.584.668	59,6	2.496,19

AMOSTRA	*SCAFFOLDS*	TOTAL DE BASES	CONTEÚDO GC (%)	NÚMERO DE N (a cada 100 kpb)
2VLh90	36	32.663.811	59,61	2.652,16
3VLh90	36	32.624.839	59,6	2.630,24
4VLh90	36	32.699.510	59,61	2.844,71
5VLh90	36	32.610.201	59,6	2.581,76
6CLh	36	32.667.118	59,61	2.627,31
7VLd	36	32.652.304	59,61	2.625,51
8Ah	36	32.670.356	59,62	2.485,07
9Ah	36	32.682.617	59,6	2.700,29

A partir dos genomas montados, uma média de 8.695 proteínas funcionais foram preditas, sendo 2.983 proteínas hipotéticas sem qualquer tipo de anotação e 367 proteínas hipotéticas com pelo menos um tipo de anotação funcional. Com isso, as 2.983 proteínas hipotéticas sem anotação correspondem a 34,31% da média total de proteínas preditas para as amostras deste estudo. Já o genoma de referência de *L. infantum* apresenta um conjunto de 8.591 proteínas preditas, sendo 3.476 proteínas hipotéticas, o que corresponde a 40,46% do conjunto de proteínas totais. Um panorama geral das predições gênicas pode ser encontrado na **Tabela 6**. Dentre a média de 8.695 proteínas preditas e anotadas, uma média de 5.041 proteínas apresentaram algum tipo de anotação funcional provenientes de ontologia genética, em contraste às 4.609 proteínas com anotações funcionais do genoma de referência.

Tabela 6. Proteínas preditas e anotadas para genomas montados de *Leishmania infantum*.

AMOSTRA	PROTEÍNAS TOTAIS		PROTEÍNAS HIPOTÉTICAS	
	Totais	Anotação funcional	Totais	Anotação funcional
JPCM5	8.591	4.609	3.476	1.129
D1	8.685	5.029	3.347	369
D2	8.665	5.014	3.338	367

AMOSTRA	PROTEÍNAS TOTAIS		PROTEÍNAS HIPOTÉTICAS	
	Totais	Anotação funcional	Totais	Anotação funcional
D3	8.704	5.029	3.358	362
D4	8.647	5.007	3.343	373
D5	8.624	5.017	3.337	371
D6	8.633	4.993	3.333	366
D7	8.626	5.021	3.337	371
H1	8.702	5.043	3.348	369
11VLd	8.684	5.037	3.342	368
12VLh	8.717	5.050	3.358	370
13VLh	8.702	5.042	3.350	368
14VLh	8.707	5.043	3.352	360
15VLd	8.724	5.045	3.361	365
16VLd	8.704	5.070	3.348	368
17VLd	8.713	5.052	3.350	369
18Ah	8.690	5.058	3.353	369
19VLh	8.715	5.052	3.350	367
1VLh90	8.722	5.042	3.363	361
20VLh	8.708	5.053	3.347	370
2VLh90	8.721	5.068	3.363	360
3VLh90	8.689	5.038	3.338	362
4VLh90	8.708	5.047	3.356	366
5VLh90	8.702	5.043	3.346	370
6CLh	8.706	5.056	3.345	371
7VLd	8.715	5.043	3.361	369
8Ah	8.733	5.063	3.361	370
9Ah	8.708	5.058	3.355	361

Proteínas hipotéticas com anotações funcionais apresentam termos relacionados com ontologia genética (do inglês, *Gene Ontology*) nos domínios componente celular, função molecular e/ou processo biológico.

O preditor gênico *AUGUSTUS* realiza predições *ab initio* para organismos eucariotos (STANKE; WAACK, 2003), neste caso, a partir de um modelo preditivo obtido das proteínas do genoma de referência, que possibilita uma predição mais acurada e robusta (EJIGU; JUNG, 2020).

Seria possível refinar os resultados obtidos utilizando dados de transcriptômica em conjunto aos dados genômicos – mais especificamente as montagens cromossômicas – já que os mapeamentos de *reads* de RNA-seq possibilitaria a identificação de genes e transcritos nos genomas montados, corroborando ou complementando as predições (CHEN; SHI; SHI, 2017; JUNG et al., 2020). A utilização de dados de RNA-seq também possibilitaria uma melhor compreensão na identificação de funções biológicas a partir de anotações funcionais (JUNG et al., 2020; VAN DEN BERGE et al., 2019).

Anotações funcionais envolvendo ontologia genética podem ser divididas em três domínios: componente celular, função molecular e processo biológico. A criação de um novo tipo de anotação partiu do pressuposto de que existe um número finito de genes e proteínas, que são conservadas em grande parte dos seres vivos (ASHBURNER et al., 2000). Esta conservação fomentou a descoberta de informações relacionadas a genes e proteínas compartilhadas entre diferentes organismos (ASHBURNER et al., 2000). Este compartilhamento de informações é possível devido aos três domínios de anotação, que permitem a classificação de produtos gênicos dentro de um conjunto estruturado e controlado de termos (BADA et al., 2004; SMITH; WILLIAMS; SCHULZE-KREMER, 2003).

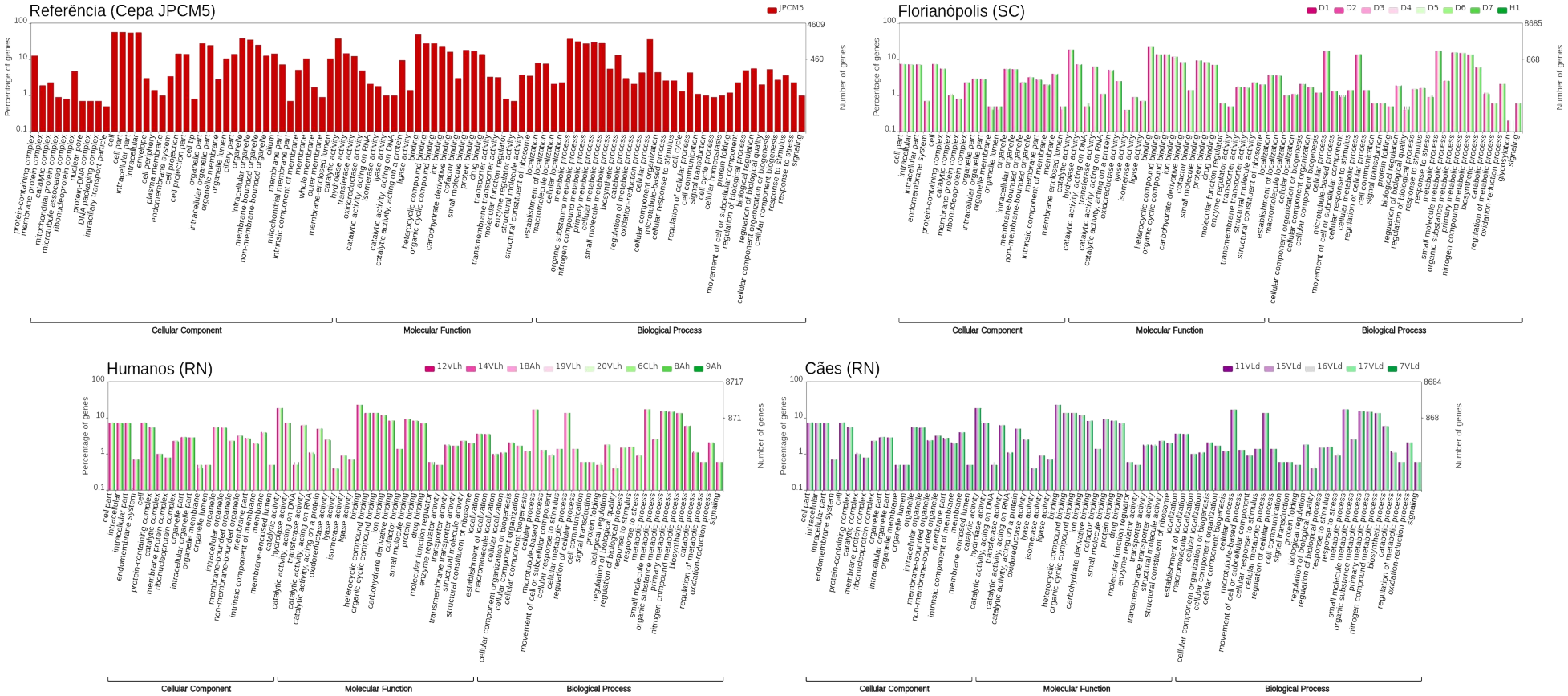
Todas as amostras de Santa Catarina e Rio Grande do Norte apresentaram um perfil semelhante, como ilustrado na **Figura 4**. Uma possibilidade para a semelhança de ontologia, para todas as proteínas com anotação funcional, entre todas as amostras é a alta identidade – acima de 99%, incluindo o genoma de referência – que diferentes cepas de *L. infantum* apresentam entre si (TEIXEIRA et al., 2017).

Uma etapa primordial para a infecção de parasitos é a de ligação em algum receptor de seu hospedeiro (MENEZES; SARAIVA; ROCHA-AZEVEDO, 2016). A

partir da **Figura 4** – tanto para a referência quanto para as amostras analisadas neste estudo – a função molecular mais expressiva foi a de ligação celular (do inglês, *binding*), corroborando com a importância da adesão para a infecção e sobrevivência do parasito (DI-BLASI et al., 2015; KILLICK-KENDRICK, 1990).

Além disso, outras anotações funcionais mais expressivas e de interesse foram: (i) composição celular e de organelas, (ii) atividade catalítica e de transporte, (iii) função molecular relacionada à localização, (iv) processos metabólicos e celulares, e (v) sinalização. Algumas dessas anotações também foram relatadas por Ottino (2021) – onde o enriquecimento de termos ontológicos para cromossomos com maior número de cópias foi realizado – e por Andrade e colaboradores (2020), especialmente, em genes envolvidos em processos relacionados à manutenção e sobrevivência do parasito dentro das células hospedeiras.

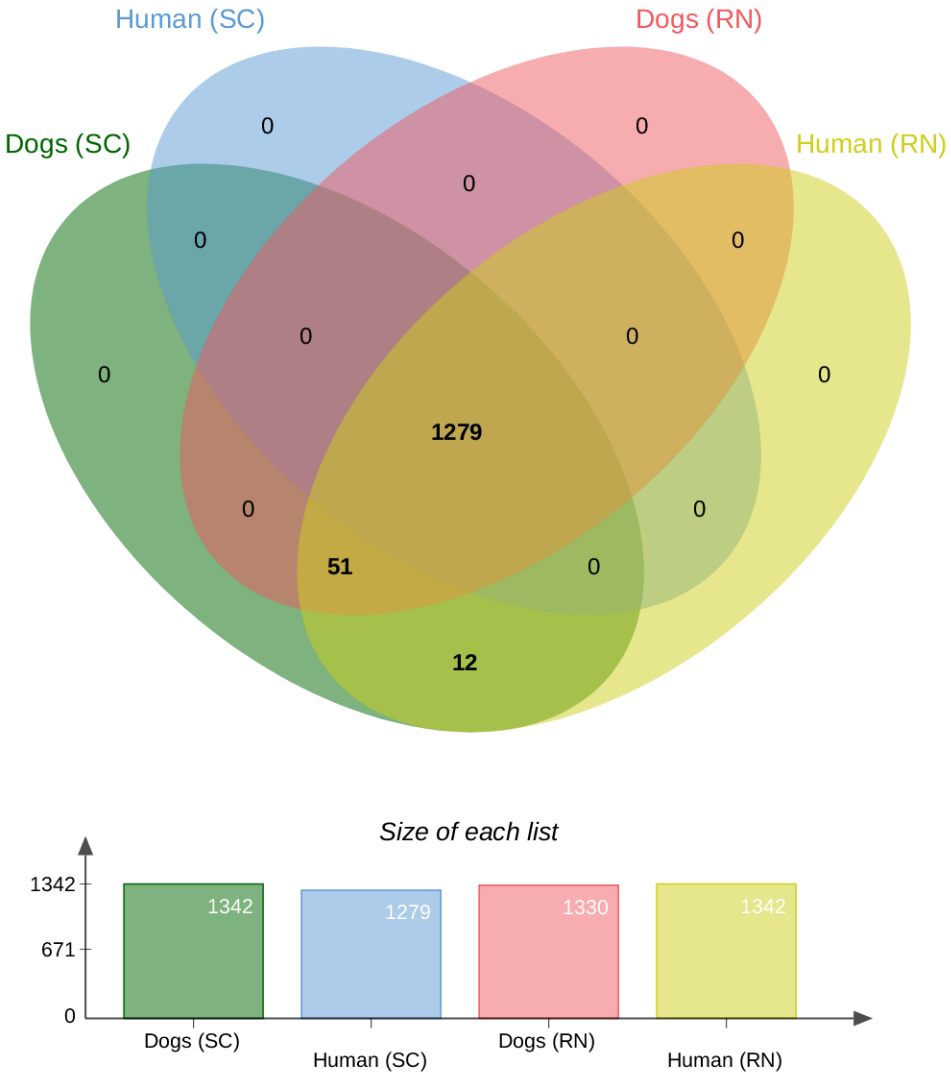
Figura 4. Termos mais abundantes de ontologia genética para proteínas anotadas de *Leishmania infantum*.



Amostras obtidas de humanos entre o período de 1991-1993 se apresentaram altamente similares as amostras de humanos obtidas entre o período de 2009-2013. Por este motivo, não foram incluídas na imagem para melhor visualização dos outros grupos. A amostra *13VLh* não foi reconhecida pelo programa *WEGO 2.0*.

Em sua grande maioria, os termos ontológicos representados na **Figura 4** se mostraram presentes nos quatro subgrupos divididos (**Figura 5**). Apenas para amostras entre duas combinações apresentaram termos únicos: (i) cães de Santa Catarina e cães/humanos do Rio Grande do Norte, e (ii) cães de Santa Catarina e humanos do Rio Grande do Norte. Uma possível explicação para a não distinção de termos ontológicos para humanos de Santa Catarina pode estar relacionado com a amostragem disponível (n = 1).

Figura 5. Número de termos ontológicos de *Leishmania infantum* compartilhados pelos grupos de diferentes localidades.



SC: Santa Catarina. RN: Rio Grande do Norte.

As anotações funcionais para os termos ontológicos únicos obtidos pelo *Revigo* para as combinações (i) e (ii) estão representadas pela **Figura 6** e **Figura 7**, respectivamente. Entretanto, como anotações funcionais estão relacionadas a um conjunto finito de termos enquadrados em três diferentes domínios, é possível que a alta sintenia tenha tornado mais evidentes termos gerais relacionados à sobrevivência do parasito (**Figura 4**) que estão presentes em todas as amostras. Por

isso, termos mais específicos, como os ilustrados pela **Figura 6** e **Figura 7**, podem ter sido subestimados por apresentarem menor quantidade de GOs.

Para a combinação (i), o processo biológico mais representado foi o de catabolismo de prolina. Para vetores de tripanossomatídeos, a prolina é utilizada como principal fonte de carbono e energia circulante na hemolinfa (BURSELL, 1981). Já para formas promastigostas de *Leishmania*, a prolina é acumulada ativamente – contra o gradiente de concentração – e utilizada para o metabolismo energético do parasito (BRINGAUD; BARRETT; ZILBERSTEIN, 2012; KRASSNER, 1969; KRASSNER; FLORY, 1972; MAZAREB; FU; ZILBERSTEIN, 1999; WESTROP et al., 2015). O cromossomo 31 de *Leishmania* apresenta anotações funcionais relacionadas ao transportador de arginina de alta afinidade, responsável pela captação do aminoácido em hospedeiros vertebrados (OTTINO, 2021). Tripanossomatídeos não sintetizam arginina, mas a utilizam para a biossíntese de compostos necessários – sendo um deles a prolina – para seu crescimento e diferenciação (MUXEL et al., 2018).

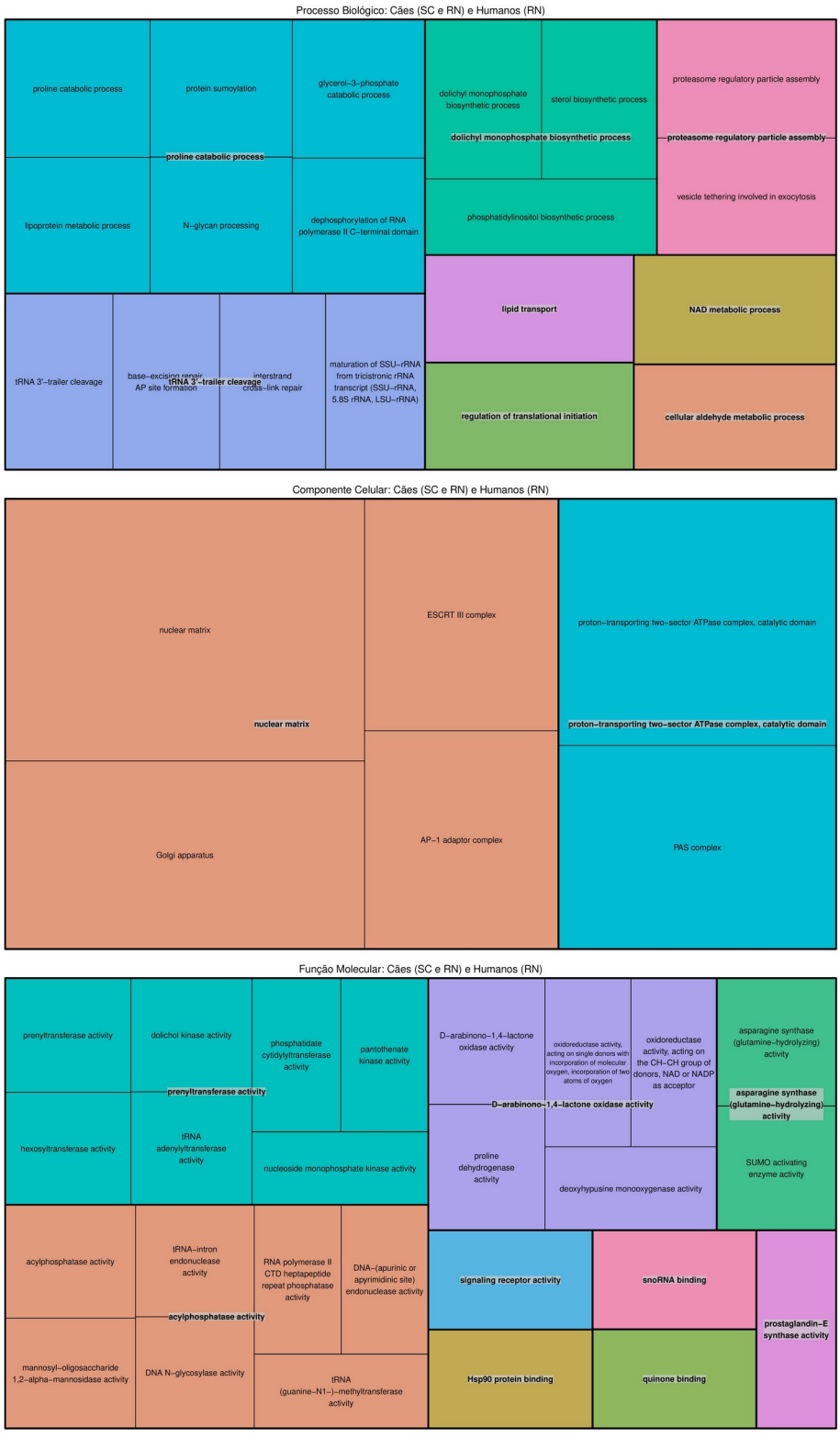
Outro processo biológico relevante presente nos GOs para a combinação (i) é o de regulação da síntese de proteínas. Ao contrário de outros organismos eucariontes, tripanossomatídeos apresentam transcrição constitutiva policistrônica. Isso significa que RNAs são constantemente transcritos e o controle de expressão gênica ocorre de forma pós-transcricional – a partir da degradação de moléculas de RNAs ou de proteínas traduzidas – além de permitir resposta imediata frente a estresses no microambiente dos parasitos (GRÜNEBAST; CLOS, 2020; LEIFSO et al., 2007; PABLOS; FERREIRA; WALRAD, 2016). Alterações no processo de tradução são responsáveis por desencadear a diferenciação celular da forma promastigota para amastigota, e ocorrem quando os parasitos são expostos a mudanças de pH e, especialmente, temperatura após sua transmissão para hospedeiros vertebrados (KARAMYSHEVA; GUARNIZO; KARAMYSHEV, 2020).

Dentre as diferentes formas de *Leishmania*, a forma amastigota é a que apresenta menor número de proteínas, sendo a diminuição decorrente da redução do volume celular nesta etapa de diferenciação (PABLOS; FERREIRA; WALRAD,

2016). A fosforilação e, consequentemente, ativação da subunidade alfa do fator de iniciação eucariótico 2 (eIF2α, do inglês, *alpha-subunit of eukaryotic initiation factor 2*) que está relacionada com o desencadeamento do processo de tradução e alteração da síntese proteica durante a amastigogênese (GRÜNEBAST; CLOS, 2020; PABLOS; FERREIRA; WALRAD, 2016; SONENBERG; HINNEBUSCH, 2009).

Um terceiro processo biológico que fica em evidência é o de formação do proteassoma, um complexo proteico dependente de ATP que apresenta atividade proteolítica. Proteassomas são compostos por duas subunidades: uma relacionada ao processo de degradação em si (20S) e outra, ao processo de regulação e ativação do complexo (19S) (TANAKA, 2009). Em células eucarióticas, o *turnover* proteico – relacionado com renovação proteica a partir de processos de síntese e degradação – é mediado através da via de ubiquitina-proteassoma (GOLDBERG et al., 1997; SILVA-JARDIM; FÁTIMA HORTA; RAMALHO-PINTO, 2004). A sinalização através do processo de ubiquitinação permite que a proteólise ocorra de forma ordenada, já que somente proteínas marcadas são degradadas pelo proteassoma (SINHA; SARKAR, 2013). Portanto, as alterações celulares que ocorrem em *Leishmania* – não somente morfológicas, mas também a nível proteico – necessitam de controle através da via proteolítica para manutenção da homeostase proteica nas diferentes formas do parasito (MUÑOZ et al., 2015; SINHA; SARKAR, 2013).

Figura 6. Ontologia para termos únicos presentes em amostras de cães (SC e RN) e humanos (RN).



Para a combinação (ii), o processo biológico mais representado foi o de transporte intracelular de colesterol, que potencialmente atua em conjunto com o transporte de lipídeos (**Figura 6**). É possível que o colesterol seja extraído da membrana plasmática – a partir de microdomínios de membrana ricos em lipídeos – de macrófagos para que a resposta imune do hospedeiro seja prejudicada, assim como relatado em *L. donovani* (GHOSH et al., 2012; PUCADYIL et al., 2004; SEMINI et al., 2017; SVIRIDOV; BUKRINSKY, 2014). Além disso, os lipídeos e colesterol também podem ser utilizados pelo parasito para produção de energia e biossíntese de sua própria membrana plasmática (MARTÍNEZ; RUIZ, 2019). Portanto, a infecção afeta tanto o perfil lipídico quanto o metabolismo de colesterol dos hospedeiros vertebrados e, com isso, a síntese de colesterol fica comprometida, especialmente para lipoproteínas de alta densidade (HDL) (CARVALHO et al., 2014; DESCOTEAUX; MORADIN; DUQUE, 2013; MARTÍNEZ; RUIZ, 2019; SOARES et al., 2010).

O segundo processo biológico mais representado foi o relacionado ao metabolismo de inositol trifosfato. Parasitos do gênero *Leishmania* apresentam, em sua maioria, inositol em suas moléculas de superfície (DREW et al., 1995). O inositol trifosfato atua como mensageiro secundário para mobilização intracelular de íons de cálcio e está localizado no retículo endoplasmático de células eucarióticas (BERRIDGE, 1993; BERRIDGE; IRVINE, 1984). A modulação da concentração de cálcio intracelular permite a regulação de diferentes processos celulares, como: (i) contração muscular, (ii) transmissão de sinapses, (iii) motilidade, (iv) divisão e diferenciação celular, (v) expressão gênica, e (vi) apoptose (BERRIDGE, 2016; FOSKETT et al., 2007). Em sua forma promastigota, fosfolipídeos – que contém inositol – ancorados na membrana plasmática do parasito sofrem alterações que permitem a infecção de macrófagos e adesão ao trato intestinal de vetores (DREW et al., 1995). Células infectadas apresentam menor concentração de inositol trifosfato afetando, consequentemente, vias dependentes de sinalização através de íons de cálcio (DOCAMPO; MORENO, 1996; OLIVIER; BAIMBRIDGE; REINER, 1992).

Figura 7. Ontologia para termos únicos presentes em amostras de cães (SC) e humanos (RN).

As combinações (i) e (ii) enquadram as amostras de *L. infantum* obtidas de cães em Santa Catarina. Embora não existam relatos do vetor clássico, *Lutzomyia longipalpis*, para LV em Santa Catarina (DIAS et al., 2013), os parasitos apresentaram anotações funcionais para processos biológicos condizentes com infecções clássicas, apesar de serem submetidas a pressões seletivas diferentes por infectar um vetor não-clássico.

O número de rRNAs e tRNAs preditos e validados pode ser encontrado na **Tabela 7**. Exceto para rRNAs do genoma de referência e uma predição de rRNA 28S para *D5*, todas as amostras apresentaram validação para os genes não-codificantes preditos, quando comparados com as sequências de *L. infantum* presentes no *RNAcentral*.

Uma possível explicação para a quase totalidade de validações pode estar relacionada a contiguidade das montagens submetidas aos preditores de elementos estruturais. Montagens com maior contiguidade e mais completas possibilitam que mais elementos sejam preditos corretamente (MOLINA-MORA et al., 2020).

Como as montagens chegaram a proporção de 1:1 entre *scaffolds* e cromossomos, sequências preditas podem ter ficado mais completas e, consequentemente, mais acuradas, o que possibilitou passarem pelos valores de corte estabelecidos na etapa de validação.

Tabela 7. Predição de genes não-codificantes para genomas montados de *Leishmania infantum*.

AMOSTRA	rRNAs		tRNAs	
	Preditos	**Validados**	**Preditos**	**Validados**
JPCM5 (Referência)	20	17	84	84
D1	14	14	84	84
D2	14	14	84	84
D3	14	14	84	84
D4	18	18	85	85
D5	16	15	79	79
D6	14	14	82	82

AMOSTRA	rRNAs		tRNAs	
	Preditos	Validados	Preditos	Validados
D7	18	18	80	80
H1	14	14	84	84
11VLd	14	14	83	83
12VLh	14	14	84	84
13VLh	14	14	84	84
14VLh	14	14	84	84
15VLd	14	14	83	83
16VLd	17	17	83	83
17VLd	13	13	81	81
18Ah	14	14	83	83
19VLh	14	14	83	83
1VLh90	17	17	84	84
20VLh	13	13	81	81
2VLh90	14	14	84	84
3VLh90	14	14	83	83
4VLh90	14	14	83	83
5VLh90	14	14	83	83
6CLh	14	14	83	83
7VLd	14	14	84	84
8Ah	14	14	84	84
9Ah	13	13	81	81

As predições de rRNAs agrupam genes de 28S, 18S, 5.8S e 5S. Produtos preditos foram validados *in silico* com base em valores de cobertura e identidade acima de 90% em relação às sequências de *L. infantum* depositadas no banco de dados *RNAcentral*.

Genes não-codificantes de rRNA se referem às sequências responsáveis pela formação das subunidades que constituem o ribossomo e apresentam, normalmente, repetições em *tandem* (GOODFELLOW; ZOMERDIJK, 2013; SOLLNER-WEBB; MOUGEY, 1991). As subunidades de um ribossomo podem ser classificadas em: 18S, 5.8S e 28S. A subunidade 18S faz parte da pequena subunidade (SSU, do inglês, *small subunit*), enquanto 5.8S e 28S fazem parte da grande subunidade (LSU, do inglês, *large subunit*) dos ribossomos (DECUYPERE et

al., 2005). Para *L. donovani*, os genes de rRNA estão presentes no cromossomo 27 (YAN et al., 1999), assim como ocorreu para todas as predições validadas – tanto para LSU quanto para SSU – deste estudo.

Sequências de SSU rRNA e outros genes *housekeeping* já foram utilizados em estudos filogenéticos para inferência da história evolutiva de tripanossomatídeos (BORGHESAN et al., 2013; MARCILI et al., 2014; ZHANG et al., 2013). Para *L. infantum*, a partir de dados de variantes genéticas, relações filogenéticas foram obtidas com base em amostras oriundas do sudeste da Espanha (ORTUÑO et al., 2019). Uma abordagem similar poderia ser utilizada para possível inferência da história evolutiva de *L. infantum* no Brasil, a partir de sequências depositadas em bancos de dados públicos em conjunto com as amostras deste estudo (CARVALHO et al., 2020; SCHWABL et al., 2021). Com isso, talvez fosse possível identificar a região de origem dos parasitos introduzidos em Santa Catarina.

A **Tabela 8** apresenta os resultados de predição de pseudogenes obtidos pelo programa *PseudoPipe*. As predições podem ser classificadas como: (i) fragmentos de genes (FRAG), (ii) duplicações gênicas (DUP), ou pseudogenes produzidos por retrotransposição (PSSD).

Tabela 8. Pseudogenes de *Leishmania infantum* identificados e classificados pelo programa *PseudoPipe*.

AMOSTRA	PREDIÇÕES		CLASSIFICAÇÃO		
	Totais	Anotadas*	FRAG	DUP	PSSD
D1	847	152	572	14	261
D2	842	136	566	10	266
D3	851	154	565	10	276
D4	822	141	555	7	260
D5	829	135	560	5	264
D6	886	157	603	9	274
D7	820	133	557	5	258
H1	848	152	569	10	269
11VLd	850	156	574	12	264

12VLh	841	138	564	11	266
13VLh	871	158	580	9	282
14VLh	859	159	575	8	276
15VLd	875	153	586	7	282
16VLd	858	155	583	10	265
17VLd	856	157	580	8	268
18Ah	845	128	566	13	266
19VLh	858	154	577	10	271
1VLh90	860	155	577	10	273
20VLh	862	158	575	6	281
2VLh90	842	131	562	8	272
3VLh90	853	157	579	7	267
4VLh90	874	162	584	10	280
5VLh90	855	163	584	6	265
6CLh	860	157	585	9	266
7VLd	839	136	567	8	264
8Ah	865	150	574	12	279
9Ah	860	169	590	11	259

***** Possíveis anotações não-hipotéticas. FRAG: pseudogenes identificados como fragmento de genes. DUP: pseudogenes identificados como duplicação gênica. PSSD: pseudogenes produzidos por retrotransposição.

Pseudogenes são classificados – de acordo com sua estrutura e origem – de quatro formas: (i) processados, (ii) duplicados, (iii) unitários, ou (iv) polimórficos (SALMENA, 2021). Pseudogenes processados recebem este nome por derivar de mRNAs processados a partir de retrotransposição (SALMENA, 2021). Portanto, mRNAs transcritos são incorporados novamente ao genoma através de transcrição reversa (CHEETHAM; FAULKNER; DINGER, 2020). Por se originarem de mRNAs processados, também são considerados fragmentos de genes, já que derivam de exons e não apresentam sequências promotoras. Pseudogenes duplicados, também conhecidos como não-processados, resultam de duplicações parciais ou completas de genes em conjunto com mutações que impossibilitam sua transcrição (CHEETHAM; FAULKNER; DINGER, 2020; SALMENA, 2021). Estas duas classes são as mais comuns em termos de pseudogenes. Portanto, para os resultados do

PseudoPipe, as classificações FRAG e PSSD se referem a pseudogenes processados e DUP, a não-processados. Pseudogenes unitários são menos comuns e se originam de genes codificantes que acabam por perder sua funcionalidade em decorrência de mutações e, ao fim, não apresentam mais cópias funcionais no genoma (CHEETHAM; FAULKNER; DINGER, 2020). Por fim, a classe mais rara é a polimórfica. Pseudogenes polimórficos são inativados, por conta de mutações, em genomas de referência, mas apresentam cópias ativas em outros genomas (SALMENA, 2021). Portanto, podem originar proteínas truncadas por consequência de mutações não-sinônimas ou mudança de quadro de leitura.

A **Figura 8** apresenta uma nuvem de palavras – onde termos mais frequentes apresentam tamanhos maiores – com as anotações não-hipotéticas mais frequentes para os pseudogenes preditos pelo *PseudoPipe*.

Figura 8. Nuvem de palavras representando possíveis anotações relacionadas à pseudogenes preditos de *Leishmania infantum*.

O termo *“membrane associated protein-like protein”* foi removido por apresentar uma proporção dez vezes maior que o segundo termo mais prevalente (*“tuzin”*). Os termos *“putative”* e *“containing protein”* também foram removidos.

Anotações relacionadas aos termos *tuzina* e *amastina* foram mais evidentes e chamam atenção por serem proteínas transmembrana exclusivas de tripanossomatídeos (TEIXEIRA; KIRCHHOFF; DONELSON, 1999). Portanto, em conjunto com a anotação mais representada (*“membrane associated protein-like*

protein"), os pseudogenes mais frequentes partiram de proteínas de membrana. Para amastinas e proteínas-*like* associadas à membrana, a classificação foi relacionada à pseudogenes processados. Já para tuzinas, também houve classificação de pseudogenes duplicados, que representam 17,55% – 43 predições de 245 caracterizadas como DUP – de todas as ocorrências de pseudogenes originados por duplicação.

Genes que codificam para amastina e tuzina, normalmente, estão dispostos de forma contígua no genoma de tripanossomatídeos e apresentam múltiplas cópias gênicas (LAKSHMI; WANG; MADHUBALA, 2014; PAIVA et al., 2015). Por este motivo, é provável que as chances de sofrerem fragmentações aumentem, acarretando, assim, em um maior número de pseudogenes relacionados a estas famílias gênicas. Produtos destas famílias gênicas ainda não apresentam funções conhecidas, porém estão relacionadas com infecção e patogenicidade em hospedeiros (JACKSON, 2010; LAKSHMI; WANG; MADHUBALA, 2014; PAIVA et al., 2015).

Como mencionado anteriormente, pseudogenes podem apresentar papel regulatório na transcrição de genes e, por tripanossomatídeos apresentarem transcrição constitutiva, o controle da regulação deve ocorrer de forma pós-transcricional. Pseudogenes que atuam na regulação pós-transcricional podem regular a expressão de seus genes parentais funcionais (MURO; MAH; ANDRADE-NAVARRO, 2011).

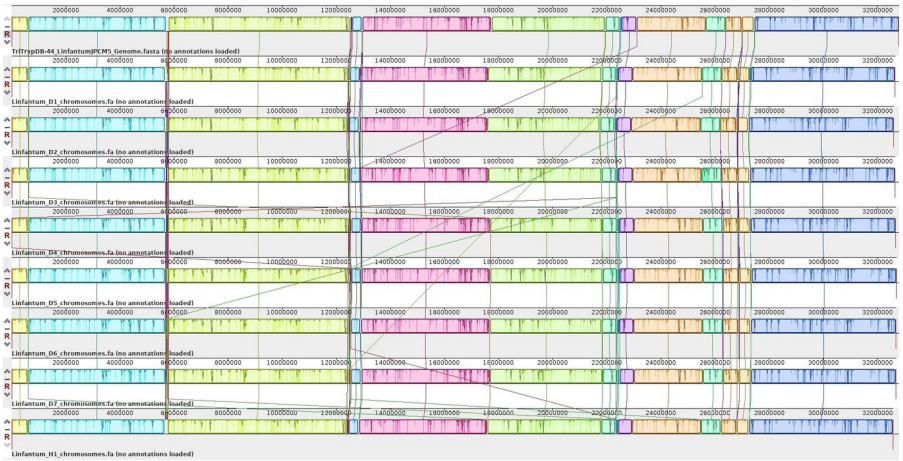
Os pseudogenes preditos mais frequentes se originaram de genes exclusivos de tripanossomatideos que estão presentes, de forma geral, por toda extensão da família Trypanosomatidae. Mas quando pseudogenes derivam de genes espécie-específicos de outras espécies de parasitos do mesmo gênero – caso se tornem codificantes novamente – podem aumentar as chances de sobrevivência ou perfil de virulência do parasito (MCCALL; MATLASHEWSKI, 2010; ZHANG; MATLASHEWSKI, 2010, 2012). Portanto, pseudogenes polimórficos, em tripanossomatídeos, podem estar relacionados com vantagens evolutivas em termos de infecção e proliferação.

4.3 ALINHAMENTO E ORTOLOGIA DE PROTEÍNAS ANOTADAS

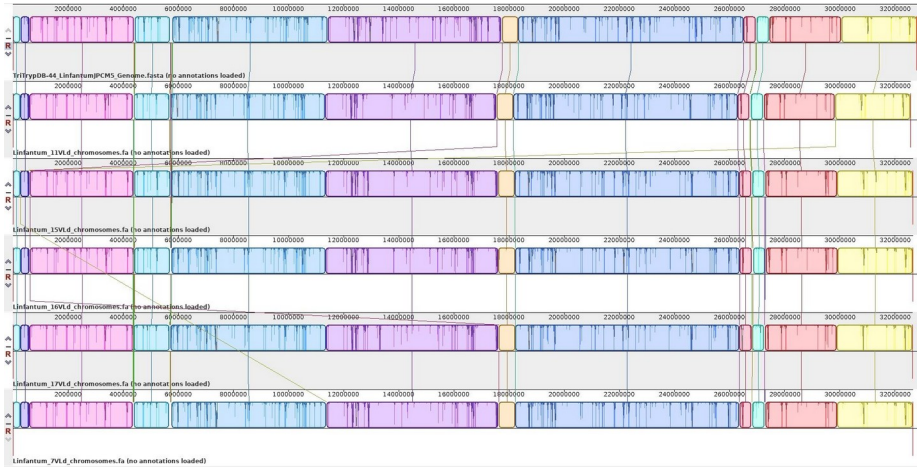
Os resultados obtidos do *progressiveMauve* apresentam alto nível de similaridade (**Figura 9**). Blocos sintênicos são representados pela mesma cor e possíveis rearranjos ou variações são ilustrados pelas linhas que conectam diferentes genomas. Já regiões de menor similaridade apresentam um declive de coloração mais escura. Como genomas de *Leishmania* apresentam um maior número de sequências repetitivas (ADEWALE, 2020; ALKAN; SAJJADIAN; EICHLER, 2011; GONZÁLEZ-DE LA FUENTE et al., 2017), é possível que os declives da **Figura 9** representem as regiões de repetição que não foram contempladas pelo processo de montagem.

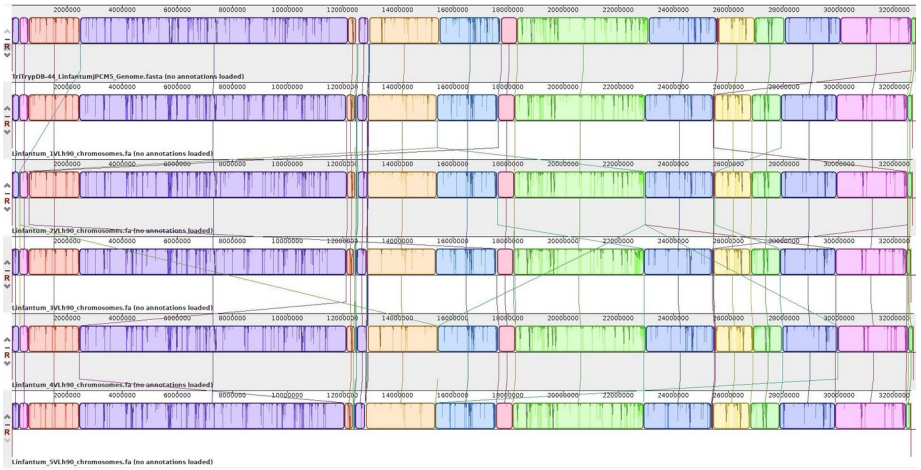
Figura 9. Alinhamentos múltiplos entre genomas de *Leishmania infantum* gerados pelo alinhador *progressiveMauve*.

Florianópolis (SC)

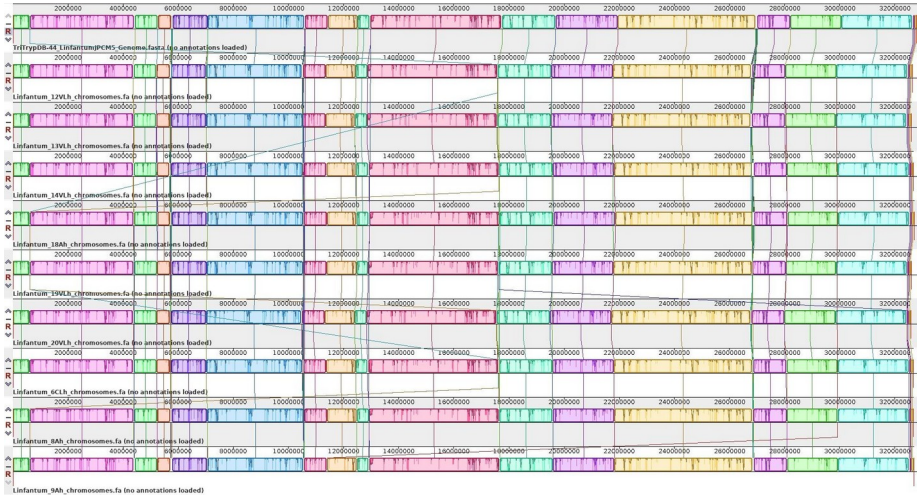


Rio Grande do Norte (Cães)





Rio Grande do Norte (Humanos - Anos 90)



Rio Grande do Norte (Humanos)

Para cada agrupamento, o primeiro alinhamento se refere ao genoma de referência de *Leishmania infantum* (cepa JPCM5).

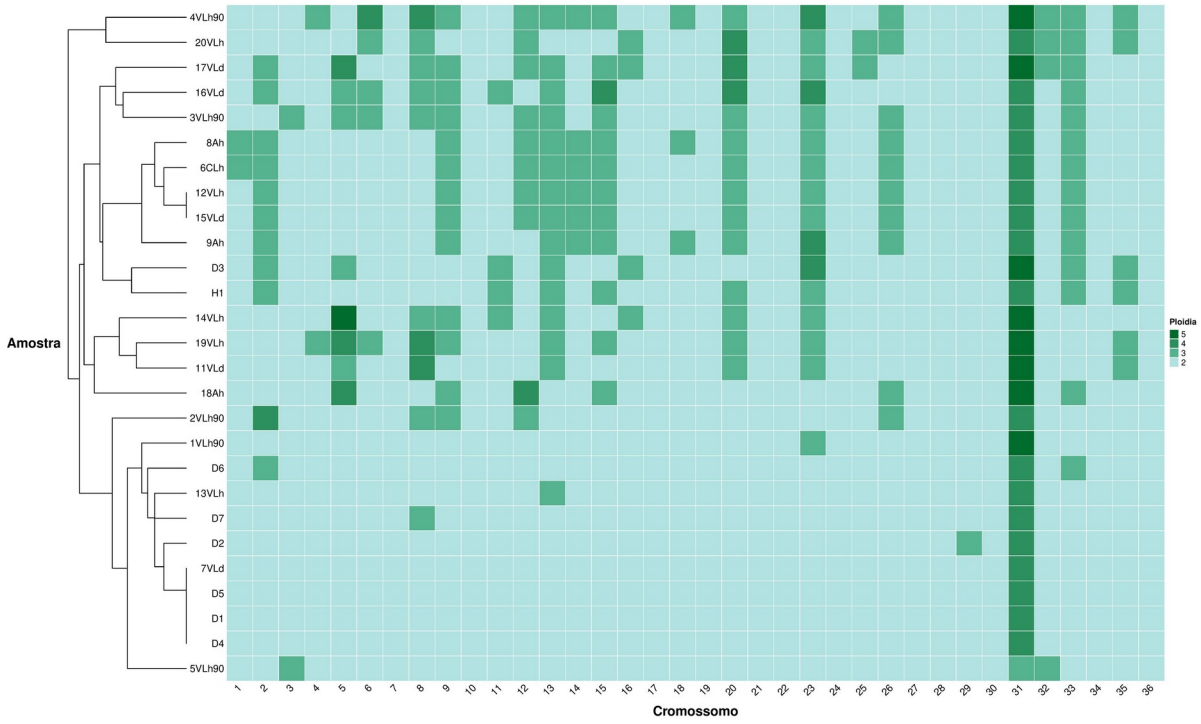
Para os alinhamentos referentes as amostras de Santa Catarina e amostras obtidas de humanos no Rio Grande do Norte durante os anos 90, os rearranjos entre as amostras ficam mais evidentes. Uma possível explicação pode estar relacionada com o período de introdução dos parasitos nas regiões. Parasitos recém-introduzidos possuem maior plasticidade genética e, frente as pressões seletivas exercidas pelo sistema imune dos hospedeiros e o novo ambiente, podem perpetuar variações que garantem vantagens adaptativas (CASANOVA; ABEL, 2013; NEWPORT; FINAN, 2011; ROGERS et al., 2011).

Uma vez que os parasitos se adaptam tanto ao novo ambiente quanto aos hospedeiros da região, os perfis apresentam menos rearranjos, como mostram os alinhamentos entre amostras do Rio Grande do Norte após os anos 90. Uma possibilidade para a menor quantidade de rearranjos pode estar relacionada com a forma como parasitos do gênero *Leishmania* se adaptam: alteram o número de cópias gênicas ou cromossômicas (IVENS, 2005; PABLOS; FERREIRA; WALRAD, 2016).

O número de cópias cromossômicas está representado na **Figura 10**, sendo que a ploidia variou de duas a cinco cópias cromossômicas. O cromossomo 31 foi o único que apresentou unanimemente polissomia para todas as amostras, o que está de acordo com outros estudos de aneuploidia em *Leishmania* (CARVALHO et al., 2020; DOWNING et al., 2011; ROGERS et al., 2011; SAMARASINGHE et al., 2018; TEIXEIRA et al., 2017; ZHANG et al., 2014).

A partir de análise de enriquecimento de termos ontológicos para cromossomos poliploides (OTTINO, 2021), o cromossomo 31 codifica proteínas essenciais para o funcionamento primordial do parasito. Quanto a processos biológicos, está envolvido com metabolismo de lipídeos e transporte de íons. Já para funções moleculares, está associado a atividades catalíticas, a segunda função molecular mais expressiva para as ontologias presentes na **Figura 4**. O cromossomo 31 também está relacionado – para cepas de diferentes regiões brasileiras – com a resistência ao medicamento com atividade antileishmania, miltefosina (SCHWABL et al., 2021).

Figura 10. Número de cópias cromossômicas para amostras de *Leishmania infantum* de Santa Catarina e Rio Grande do Norte.



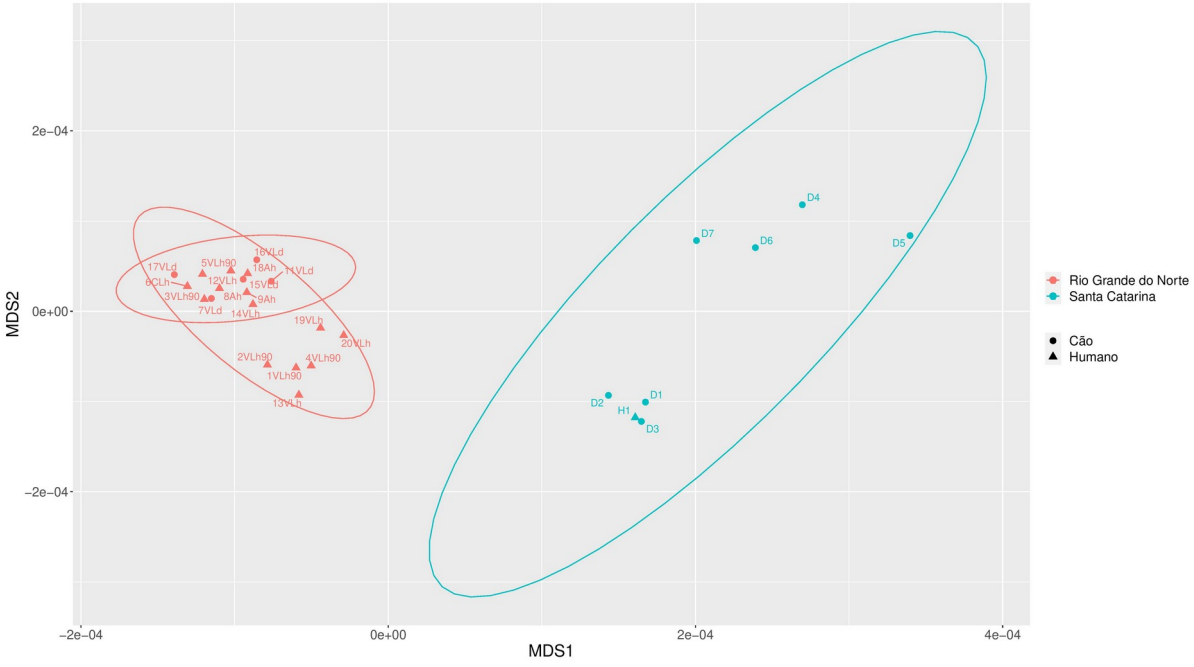
Amostras que contém a letra *D* foram obtidas de cães enquanto amostras com a letra *H* foram obtidas de humanos.

De forma geral, amostras do Rio Grande do Norte apresentaram um perfil de aneuploidia mais acentuado quando comparadas com as amostras de Santa Catarina. Aneuplodia em *Leishmania* pode ser uma estratégia de adaptação frente a pressões seletivas. Estudos envolvendo inativação de genes acarretaram em alteração do número de cópias cromossômicas, fazendo com que parasitos aumentassem a ploidia de cromossomos que continham genes essenciais para sobrevivência e virulência (DUMAS et al., 1997; LEPROHON et al., 2015; MARTÍNEZ-CALVILLO; STUART; MYLER, 2005; ZHANG et al., 2014). O grau de aneuploidia pode variar entre as diferentes espécies, mas *L. infantum* e *L. donovani* parecem apresentar o maior número de cromossomos poliploides (LEPROHON et al., 2015; ROGERS et al., 2011).

Uma possibilidade para o maior número de cromossomos poliploides no Rio Grande do Norte pode ser o tempo que o parasito está presente na região. O Rio Grande do Norte apresentou o primeiro surto de LV em 1989, enquanto os primeiros surtos relatados em Santa Catarina foram mais recentes (FIGUEIREDO et al., 2012; JERONIMO et al., 1994; STEINDEL et al., 2013). Portanto, adaptações frente as pressões seletivas ao longo do espaço temporal desde o primeiro surto no Rio Grande do Norte podem ter acarretado nas aneuploidias além do cromossomo 31.

Escalonamentos multidimensionais permitem o agrupamento e visualização de amostras similares em um conjunto de dados. O agrupamento entre as amostras de Santa Catarina e Rio Grande do Norte se encontra na **Figura 11**. As amostras foram agrupadas em dois grandes grupos, onde o fator divisor foi a região de origem.

Figura 11. Escalonamento multidimensional (MDS) de amostras de *Leishmania infantum* de Santa Catarina e Rio Grande do Norte.



Assim como relatado por Ottino (2021), as amostras de Santa Catarina demonstraram um maior nível de divergência intrapopulacional. As amostras *D1*, *D2*,

D3 e *H1* foram coletadas no período de 2010-2017, enquanto as amostras *D4*, *D5*, *D6* e *D7* foram coletadas entre 2017-2018. Este intervalo temporal presente entre as coletas pode ter acarretado tanto em mutações intrapopulacionais quanto diferentes cepas circulantes que causaram maior divergência entre as amostras de Santa Catarina. Parasitos do gênero *Leishmania* possuem alta plasticidade cariotípica que possibilitam sua adaptação ao longo de inúmeras infecções (SCHWABL et al., 2021).

Outra possibilidade pode estar relacionada aos diferentes vetores circulantes para LV em Florianópolis. Como relatado por Catecati (2018), três possíveis vetores podem estar relacionados com a transmissão de LV em Florianópolis: *Pintomyia fischeri*, *Migonemyia migonei* e *Nyssomyia neivai*. Estes flebotomíneos já foram relatados como potenciais vetores de *Leishmania* em outras regiões do Brasil (CARVALHO et al., 2010; GALVIS-OVALLOS et al., 2017, 2021; RÊGO et al., 2020) e no norte da Argentina (SALOMÓN et al., 2010). Exceto para amostra *D4* – coletada em Chapecó – é possível que dentre os potenciais vetores identificados em Florianópolis apenas um seja responsável pela transmissão de LV, assim como ocorre em Porto Alegre (Rio Grande do Sul), onde o vetor mais provável é *Pi. fischeri* (MAHMUD et al., 2019; PITA-PEREIRA et al., 2011; RÊGO et al., 2019). Portanto, a divergência intrapopulacional também pode ser decorrente de mutações que possibilitaram a infecção e sobrevivência dentro de um novo vetor.

A hipótese da Rainha Vermelha é uma hipótese evolutiva que recebe este nome em decorrência da personagem de mesmo nome (CARROLL, 2013) que diz: *"você precisa correr o máximo que puder para permanecer no lugar"*. No caso de organismos digenéticos – que necessitam de diferentes hospedeiros para completar seu ciclo de vida – como é o caso de *Leishmania*, a pressão seletiva nos diferentes hospedeiros é diferente (BENTON, 2010; HOLMGREN; MCCONKEY; SHIN, 2017). Portanto, tanto a proliferação dentro de um novo vetor quanto sua capacidade de diferenciação após o repasto sanguíneo são essenciais para que o parasito produza linhagens que se perpetuem ao longo de múltiplas infecções (EBERT, 1998). Caso os parasitos não consigam se proliferar no novo vetor – ou consigam se proliferar

sem que as adaptações adquiridas possibilitem sua diferenciação em mamíferos – acabariam por se perder no decorrer do processo evolutivo.

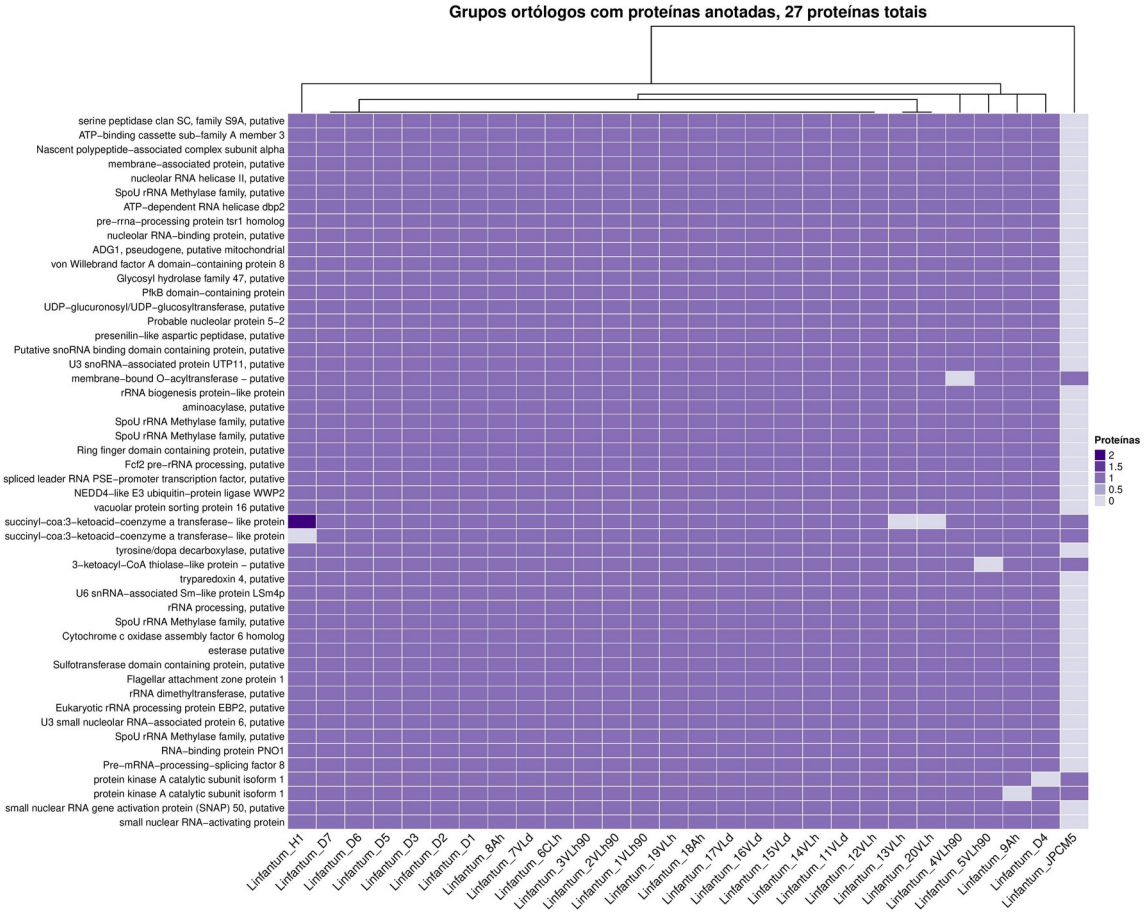
Em relação à ortologia, os ortogrupos gerados foram divididos por amostra – incluindo o genoma de referência – em um total de 8.563 grupos. O total de proteínas dentro de cada ortogrupo variou entre 2 e 285 proteínas, sendo que ortogrupos com 28 proteínas foram os mais frequentes. Ortogrupos com 28 proteínas, provavelmente, são os que possuem uma proteína ortóloga por amostra. Informações gerais a respeito dos resultados gerados pelo programa *OrthoFinder* podem ser encontradas na **Tabela 9**.

Tabela 9. Métricas gerais de resultados de ortologia de *Leishmania infantum* gerados pelo programa *OrthoFinder*.

Total de amostras	28
Total de proteínas	243.345
Total de ortogrupos	8.563
Total de proteínas em ortogrupos	243.068 (99,9%)
Média de proteínas em ortogrupos	28
Ortogrupos com proteínas de todas as amostras	7.906 (92,33%)

Dos 8.563 ortogrupos totais, 101 grupos apresentaram anotações não-hipotéticas e pelo menos uma contagem zero para alguma amostra. A **Figura 12** ilustra proteínas com contagem zero que não estão presentes, em sua maior parte, no genoma de referência.

Figura 12. Ortogrupos com 27 proteínas anotadas que estão ausentes, principalmente, na cepa JPCM5 de *Leishmania infantum*.



De acordo com a navalha de Occam (ou lei da parcimônia), a explicação mais simples para um problema tende a ser a mais correta. Por isso, é provável que as proteínas do genoma de referência – isolada em Madrid (Espanha) a partir de uma amostra de cão infectado com LV (MORENO et al., 2007) – que apresentam

contagem zero possuam menor similaridade em relação às proteínas das amostras brasileiras.

Isso quer dizer que as proteínas presentes na **Figura 12** devem estar presentes no genoma de referência, mas em ortogrupos diferentes e com menor nível de similaridade. Embora parasitos do genêro *Leishmania* apresentem genomas altamente sintênicos, diferentes populações sofrem diferentes pressões seletivas que podem ocasionar em alterações genéticas ao longo do processo evolutivo (PETIT, 2011; SCHWABL et al., 2021). Essas alterações genéticas podem alterar a similaridade entre proteínas de mesma função, mas os parasitos ainda apresentam conservação de conteúdo gênico entre as espécies (DOWNING et al., 2011, 2012; ROGERS et al., 2011; WESTROP et al., 2015).

Um exemplo é a proteína da zona de adesão flagelar, que é essencial para o processo de invasão em células de hospedeiros e divisão/diferenciação celular (BASTIN et al., 2000; SUNTER; GULL, 2016). Ela está ausente no genoma de referência (**Figura 12**) – dentro da matriz de contagem zero filtrada a partir dos resultados do programa *OrthoFinder* – mas isso causaria implicações negativas para o processo de infecção do parasito. O mais provável é que ela esteja presente em um ortogrupo onde todas as amostras apresentem pelo menos um ortólogo para proteína da zona de adesão flagelar ou que esteja presente na referência sem ortólogos – devido à baixa similaridade – no restante das amostras.

4.4 ANÁLISE DE VARIANTES

A distribuição de SNPs ao longo de cada cromossomo – dividido em intervalos de 10% do tamanho total para cada cromossomo – está ilustrado na **Figura 13**. A imagem mostra a distribuição a partir da contagem de dados da matriz de interpolações, independente de anotações hipotéticas ou não.

Os cromossomos que apresentaram maior número de SNPs foram os cromossomos 12, 27, 22, 36 e 34, respectivamente. Este perfil de distribuição vai de encontro com o perfil de variantes encontrado em amostras de *L. infantum* coletadas

de humanos em Teresina (Piauí) (CARVALHO et al., 2020). E, assim como as variantes de Teresina, a distribuição ao longo de cada cromossomo não ocorre de forma proporcional, ficando mais evidente nos cromossomos 12, 22 e 27. A partir do perfil de distribuição, é possível notar que a quantidade de SNPs é independente do tamanho e ploidia dos cromossomos, já que os cromossomos 22, 27 e 34 são diploides para todas as amostras (**Figura 10**).

Ao todo, 1.918 variantes foram obtidas. Deste total, 922 foram identificadas em genes com anotações não-hipotéticas e mutações não-sinônimas. Apenas sete destas variantes apresentaram efeito de alto impacto. A **Tabela 10** apresenta os genes com SNPs de alto impacto que resultaram em um códon de terminação. Os impactos têm caráter negativo – já que originam proteínas truncadas – e causam tradução de uma proteína não funcional. Porém, a inserção de códons de terminação em proteínas funcionais podem acarretar no surgimento de pseudogenes não-processados ao longo do processo evolutivo, que possam atuar no controle de expressão dos genes parentais (MURO; MAH; ANDRADE-NAVARRO, 2011; SAMARASINGHE et al., 2018).

Exceto para o SNP presente no cromossomo 8, que afetou todas as amostras em todos os grupos, o restante dos resultados afetou somente uma amostra para cada grupo. As amostras afetadas foram: (i) *D4* para cães de Santa Catarina; (ii) *16VLd* para cães do Rio Grande do Norte; e (iii) *4VLh90* e *12VLh* para humanos do Rio Grande do Norte, para os cromossomos 19 e 21, respectivamente.

Figura 13. Distribuição de polimorfismos de nucleotídeo único (SNP) ao longo de cada cromossomo de *Leishmania infantum*.

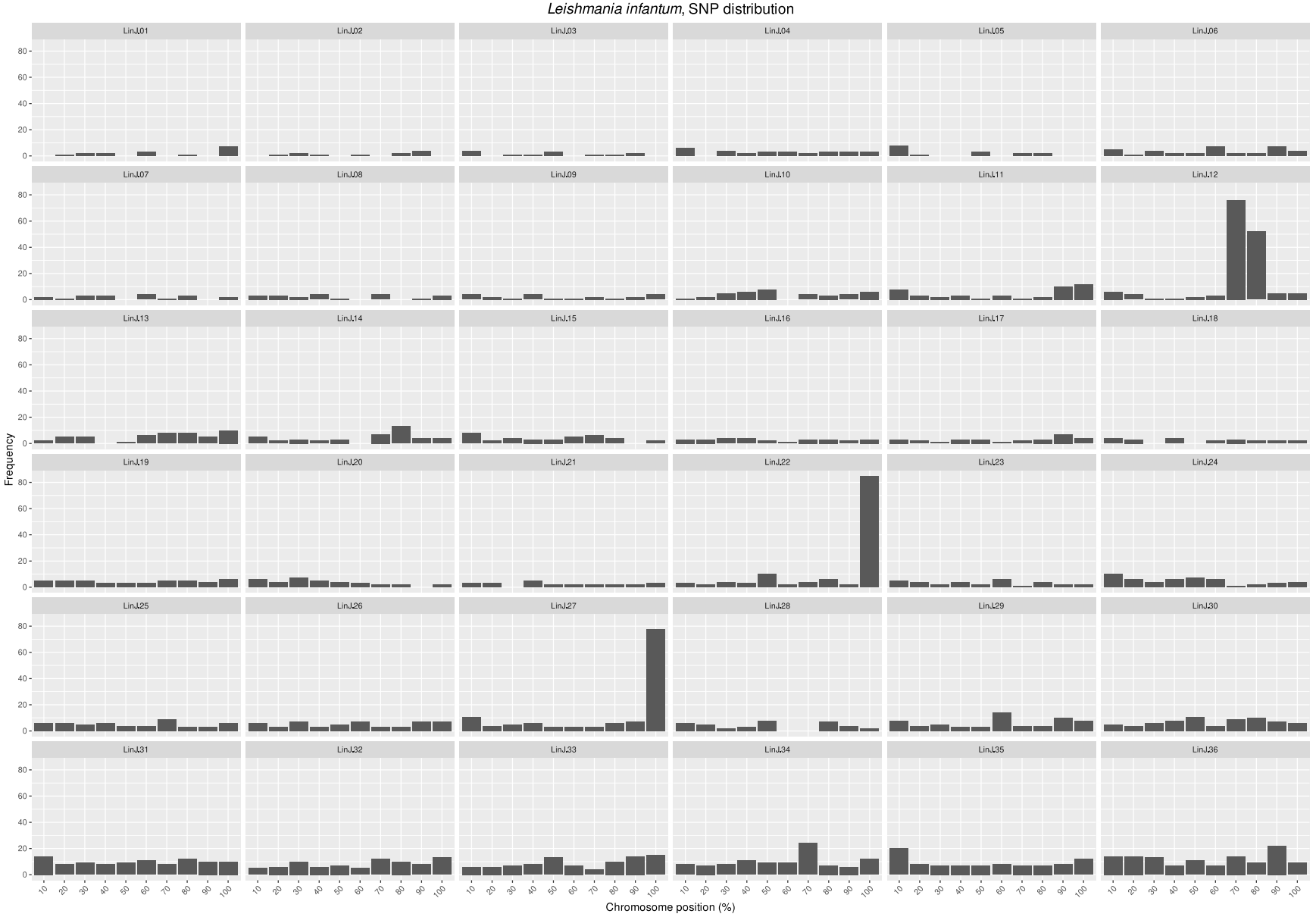


Tabela 10. Polimorfismos de nucleotídeo único (SNP) de alto impacto que resultam em códon de terminação em *Leishmania infantum*.

CROMOSSOMO	POSIÇÃO	REFERÊNCIA	SNP	CÃES (SC)	CÃES (RN)	HUMANO (SC)	HUMANOS (RN)	GENE	ANOTAÇÃO
8	437.460	C	A	✓	✓	✓	✓	LINF_080015600	histone deacetylase - putative
19	120.322	C	T	✗	✗	✗	✓	LINF_190008100	intraflagellar transport protein 52 - putative
21	272.091	C	T	✗	✗	✗	✓	LINF_210014100	temperature dependent protein affecting M2 dsRNA replication - putative
27	545.509	G	A	✓	✗	✗	✗	LINF_270018000	retinoic acid induced 16-like protein - putative
31	541.283	C	A	✗	✓	✗	✗	LINF_310019600	pentamidine resistance protein 1
34	1.816.484	G	A	✓	✗	✗	✗	LINF_340051000	NADH-ubiquinone oxidoreductase complex I subunit - putative
36	158.217	G	T	✓	✗	✗	✗	LINF_360010600	protein phosphatase 2C-like protein

Anotações retiradas do genoma de referência de *Leishmania infantum* (cepa JPCM5) obtido do banco de dados *TriTrypDB* v46.

✓ SNP presente no grupo

✗ SNP ausente no grupo

Dentre os resultados da **Tabela 10**, a proteína de resistência à pentamidina (PRP1) chama a atenção, por se tratar de uma proteína que garante vantagem evolutiva, mas que apresenta um polimorfismo que acarreta no enfraquecimento do mecanismo de resistência do parasito. O tratamento tradicional para LV ocorre por meio de administração intravenosa de antimoniais pentavalentes, mas a pentamidina pode ser utilizada como tratamento alternativo (TORRES-GUERRERO et al., 2017). Para a amostra *16VLd*, houve uma mutação que acarretou em uma proteína truncada de PRP1, que é uma glicoproteína P associada à transportadores ABC (do inglês, ATP-*binding cassette*) (COELHO et al., 2007; COELHO; BEVERLEY; COTRIM, 2003). Transportadores ABC realizam o transporte transmembrana de moléculas por importação ou exportação – neste caso, transportam pentamidina do interior para o exterior da célula – utilizando energia resultante da hidrólise de ATP (HOLLAND; BLIGHT, 1999; SAURIN; HOFNUNG; DASSA, 1999).

É de se imaginar que a mutação não se perpetuou, já que está presente em apenas uma amostra, por não conferir vantagens evolutivas ao parasito, além de proporcionar menor resistência ao tratamento, mesmo que secundário, para a doença. Também é possível que, por afetar um cromossomo poliploide, a inativação da PRP1 não acarretou em alterações significativas. Como mencionado anteriormente, a adaptabilidade de parasitos do gênero *Leishmania* está relacionado com o aumento do número de cópias gênicas ou cromossômicas, o que implicaria na compensação da glicoproteína P truncada por proteínas funcionais originadas de outros genes que codificam para PRP1.

5 CONCLUSÃO

Genomas de *L. infantum* são altamente sintênicos e apresentam perfil gênico conservado em amostras de diferentes regiões obtidas de diferentes hospedeiros, sem proteínas específicas marcantes para uma única amostra. Entretanto, amostras de isolados obtidos em Santa Catarina apresentaram maior divergência intrapopulacional, que pode estar relacionado com introduções distintas a partir de parasitos de outras regiões brasileiras, principalmente em Florianópolis.

A montagem de genomas resultou em cromossomos com métricas altamente similares e a anotação funcional de genomas possibilitou a agregação de processos biológicos à variantes para inferências de potenciais impactos no ciclo de vida dos parasitos. Polimorfismos de alto impacto que causam efeitos negativos tenderam a não afetar grupos, divididos por regiões e hospedeiros, como um todo.

Os resultados deste trabalho foram obtidos a partir de uma única plataforma de sequenciamento com amostras de dois estados brasileiros. Para refinar os resultados e complementar as análises, seriam necessários dados híbridos de sequenciamento e um maior número de amostras oriundas de diversas regiões e estados. Além disso, para inferência de genes essenciais para a sobrevivência de *L. infantum* poderia ser realizado o cálculo para identificação de genes que apresentam maior número de cópias.

REFERÊNCIAS

ADEWALE, B. A. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? **African Journal of Laboratory Medicine**, v. 9, n. 1, 26 nov. 2020.

AKHOUNDI, M. et al. *Leishmania* infections: molecular targets and diagnosis. **Molecular Aspects of Medicine**, v. 57, p. 1–29, out. 2017.

ALEMAN, C. Finestructure of cultured *Leishmania brasiliensis*. **Experimental Parasitology**, v. 24, n. 2, p. 259–264, abr. 1969.

ALKAN, C.; SAJJADIAN, S.; EICHLER, E. E. Limitations of next-generation genome sequence assembly. **Nature Methods**, v. 8, n. 1, p. 61–65, jan. 2011.

ALMEIDA, L. V. DE et al. Comparative genomics of *Leishmania* isolates from Brazil confirms the presence of *Leishmania major* in the Americas. **International Journal for Parasitology**, v. 51, n. 12, p. 1047–1057, nov. 2021.

ALTENHOFF, A. M.; DESSIMOZ, C. Inferring orthology and paralogy. In: ANISIMOVA, M. (Ed.). . **Evolutionary Genomics**. Totowa, NJ: Humana Press, 2012. v. 855p. 259–279.

ANDRADE, J. M. et al. Comparative transcriptomic analysis of antimony resistant and susceptible *Leishmania infantum* lines. **Parasites & Vectors**, v. 13, n. 1, p. 600, dez. 2020.

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

APHASIZHEV, R.; APHASIZHEVA, I. Mitochondrial RNA editing in trypanosomes: small RNAs in control. **Biochimie**, v. 100, p. 125–131, maio 2014.

ARMSTRONG, J. et al. Whole-genome alignment and comparative annotation. **Annual Review of Animal Biosciences**, v. 7, n. 1, p. 41–64, 15 fev. 2019.

ASHBURNER, M. et al. Gene Ontology: tool for the unification of biology. **Nature Genetics**, v. 25, n. 1, p. 25–29, maio 2000.

ASLETT, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic Acids Research**, v. 38, n. suppl_1, p. D457–D462, jan. 2010.

BADA, M. et al. A short study on the success of the Gene Ontology. **Journal of Web Semantics**, v. 1, n. 2, p. 235–240, fev. 2004.

BAIROCH, A. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. **Nucleic Acids Research**, v. 24, n. 1, p. 21–25, 1 jan. 1996.

BALAKIREV, E. S.; AYALA, F. J. Pseudogenes: are they “junk” or functional DNA? **Annual Review of Genetics**, v. 37, n. 1, p. 123–151, dez. 2003.

BANKEVICH, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. **Journal of Computational Biology**, v. 19, n. 5, p. 455–477, maio 2012.

BAPTISTA, R. P.; KISSINGER, J. C. Is reliance on an inaccurate genome sequence sabotaging your experiments? **PLOS Pathogens**, v. 15, n. 9, p. e1007901, 12 set. 2019.

BARDOU, P. et al. jvenn: an interactive Venn diagram viewer. **BMC Bioinformatics**, v. 15, n. 1, p. 293, dez. 2014.

BARNES, M. R. (ED.). **Bioinformatics for geneticists: a bioinformatics primer for the analysis of genetic data**. 2nd edition ed. Chichester, England; Hoboken, NJ: Wiley, 2007.

BASTIN, P. et al. Inside and outside of the trypanosome flagellum:a multifunctional organelle. **Microbes and Infection**, v. 2, n. 15, p. 1865–1874, dez. 2000.

BENTLEY, S. Sequencing the species pan-genome. **Nature Reviews Microbiology**, v. 7, n. 4, p. 258–259, abr. 2009.

BENTON, M. J. New take on the Red Queen. **Nature**, v. 463, n. 7279, p. 306–307, 21 jan. 2010.

BERRIDGE, M. J. Inositol trisphosphate and calcium signalling. **Nature**, v. 361, n. 6410, p. 315–325, jan. 1993.

BERRIDGE, M. J. The inositol trisphosphate/calcium signaling pathway in health and disease. **Physiological Reviews**, v. 96, n. 4, p. 1261–1296, out. 2016.

BERRIDGE, M. J.; IRVINE, R. F. Inositol trisphosphate, a novel second messenger in cellular signal transduction. **Nature**, v. 312, n. 5992, p. 315–321, nov. 1984.

BOETZER, M. et al. Scaffolding pre-assembled contigs using SSPACE. **Bioinformatics**, v. 27, n. 4, p. 578–579, 15 fev. 2011.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014.

BORGHESAN, T. C. et al. Molecular phylogenetic redefinition of herpetomonas (Kinetoplastea, Trypanosomatidae), a genus of insect parasites associated with flies. **Protist**, v. 164, n. 1, p. 129–152, jan. 2013.

BRINGAUD, F.; BARRETT, M. P.; ZILBERSTEIN, D. Multiple roles of proline transport and metabolism in trypanosomatids. **Frontiers in Bioscience**, v. 17, n. 1, p. 349, 2012.

BROAD INSTITUTE. **Picard Toolkit**, 2019. Disponível em: <http://broadinstitute.github.io/picard>

BURSELL, E. The role of proline in energy metabolism. In: DOWNER, R. G. H. (Ed.). . **Energy Metabolism in Insects**. Boston, MA: Springer US, 1981. p. 135–154.

BUSHNELL, B. BBMap: a fast, accurate, splice-aware aligner. mar. 2014.

BUTENKO, A. et al. Comparative genomics of *Leishmania (Mundinia)*. **BMC Genomics**, v. 20, n. 1, p. 726, dez. 2019.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC Bioinformatics**, v. 10, n. 1, p. 421, 2009.

CANTACESSI, C. et al. The past, present, and future of *Leishmania* genomics and transcriptomics. **Trends in Parasitology**, v. 31, n. 3, p. 100–108, mar. 2015.

CARRILLO, E.; MORENO, J. Cytokine profiles in canine visceral leishmaniasis. **Veterinary Immunology and Immunopathology**, v. 128, n. 1–3, p. 67–70, mar. 2009.

CARROLL, L. **Through the Looking-Glass**. London: Harper Press, 2013.

CARVALHO, M. R. DE et al. Natural *Leishmania infantum* infection in *Migonemyia migonei* (França, 1920) (Diptera:Psychodidae:Phlebotominae) the putative vector of visceral leishmaniasis in Pernambuco State, Brazil. **Acta Tropica**, v. 116, n. 1, p. 108–110, out. 2010.

CARVALHO, K. S. S. et al. Application of next generation sequencing (NGS) for descriptive analysis of 30 genomes of *Leishmania infantum* isolates in Middle-North Brazil. **Scientific Reports**, v. 10, n. 1, p. 12321, dez. 2020.

CARVALHO, M. D. T. et al. Lipoprotein lipase and PPAR alpha gene polymorphisms, increased very-low-density lipoprotein levels, and decreased high-density lipoprotein levels as risk markers for the development of visceral leishmaniasis by *Leishmania infantum*. **Mediators of Inflammation**, v. 2014, p. 1–10, 2014.

CASANOVA, J.-L.; ABEL, L. The genetic theory of infectious diseases: a brief history and selected illustrations. **Annual Review of Genomics and Human Genetics**, v. 14, n. 1, p. 215–243, 31 ago. 2013.

CATECATI, T. **Leishmaniose visceral em Florianópolis: caracterização molecular das cepas de *Leishmania infantum* isoladas de casos locais e pesquisa vetorial**. Dissertação (Mestrado em Biotecnologia e Biociências)— Florianópolis: Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, 2018.

CAVALCANTI, D. P.; SOUZA, W. DE. The kinetoplast of trypanosomatids: from early studies of electron microscopy to recent advances in atomic force microscopy. **Scanning**, v. 2018, p. 1–10, 19 jun. 2018.

CHAUDHARI, N. M.; GUPTA, V. K.; DUTTA, C. BPGA- an ultra-fast pan-genome analysis pipeline. **Scientific Reports**, v. 6, n. 1, p. 24373, abr. 2016.

CHEETHAM, S. W.; FAULKNER, G. J.; DINGER, M. E. Overcoming challenges and dogmas to understand the functions of pseudogenes. **Nature Reviews Genetics**, v. 21, n. 3, p. 191–201, mar. 2020.

CHEN, G.; SHI, T.; SHI, L. Characterizing and annotating the genome using RNA-seq data. **Science China Life Sciences**, v. 60, n. 2, p. 116–125, fev. 2017.

CHIKHI, R.; MEDVEDEV, P. Informed and automated *k*-mer size selection for genome assembly. **Bioinformatics**, v. 30, n. 1, p. 31–37, 1 jan. 2014.

CINGOLANI, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. **Fly**, v. 6, n. 2, p. 80–92, abr. 2012.

COELHO, A. C. et al. Role of the ABC transporter PRP1 (ABCC7) in pentamidine resistance in *Leishmania* amastigotes. **Antimicrobial Agents and Chemotherapy**, v. 51, n. 8, p. 3030–3032, ago. 2007.

COELHO, A. C.; BEVERLEY, S. M.; COTRIM, P. C. Functional genetic identification of *PRP1*, an ABC transporter superfamily member conferring pentamidine resistance in *Leishmania major*. **Molecular and Biochemical Parasitology**, v. 130, n. 2, p. 83–90, ago. 2003.

COSTA, D. N. C. C. et al. Human visceral leishmaniasis and relationship with vector and canine control measures. **Revista de Saúde Pública**, v. 52, p. 92, 14 nov. 2018.

DANECEK, P. et al. Twelve years of SAMtools and BCFtools. **GigaScience**, v. 10, n. 2, p. giab008, 29 jan. 2021.

DARLING, A. E.; MAU, B.; PERNA, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. **PLoS ONE**, v. 5, n. 6, p. e11147, 25 jun. 2010.

DECUYPERE, S. et al. Differential polyadenylation of ribosomal RNA during post-transcriptional processing in *Leishmania*. **Parasitology**, v. 131, n. 3, p. 321–329, set. 2005.

DEL ANGEL, V. D. et al. Ten steps to get started in genome assembly and annotation. **F1000Research**, v. 7, p. 148, 5 fev. 2018.

DESAI, A. et al. Identification of optimum sequencing depth especially for *de novo* genome assembly of small genomes using next generation sequencing data. **PLoS ONE**, v. 8, n. 4, p. e60204, 12 abr. 2013.

DESCOTEAUX, A.; MORADIN, N.; DUQUE, G. A. *Leishmania* dices away cholesterol for survival. **Cell Host & Microbe**, v. 13, n. 3, p. 245–247, mar. 2013.

DIAS, E. S. et al. Detection of *Leishmania infantum* , the etiological agent of visceral leishmaniasis, in *Lutzomyia neivai* , a putative vector of cutaneous leishmaniasis. **Journal of Vector Ecology**, v. 38, n. 1, p. 193–196, jun. 2013.

DI-BLASI, T. et al. The flagellar protein FLAG1/SMP1 is a candidate for *Leishmania*–sand fly interaction. **Vector-Borne and Zoonotic Diseases**, v. 15, n. 3, p. 202–209, mar. 2015.

DOCAMPO, R.; MORENO, S. N. J. The role of Ca^{2+} in the process of cell invasion by intracellular parasites. **Parasitology Today**, v. 12, n. 2, p. 61–65, fev. 1996.

DOWNING, T. et al. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. **Genome Research**, v. 21, n. 12, p. 2143–2156, 1 dez. 2011.

DOWNING, T. et al. Genome-wide SNP and microsatellite variation illuminate population-level epidemiology in the *Leishmania donovani* species complex. **Infection, Genetics and Evolution**, v. 12, n. 1, p. 149–159, jan. 2012.

DREW, M. E. et al. Functional expression of a *myo*-inositol/H^+ symporter from *Leishmania donovani*. **Molecular and Cellular Biology**, v. 15, n. 10, p. 5508–5515, out. 1995.

DUMAS, C. et al. Disruption of the trypanothione reductase gene of Leishmania decreases its ability to survive oxidative stress in macrophages. **The EMBO journal**, v. 16, n. 10, p. 2590–2598, 15 maio 1997.

EBERT, D. Experimental evolution of parasites. **Science**, v. 282, n. 5393, p. 1432–1436, 20 nov. 1998.

EJIGU, G. F.; JUNG, J. Review on the computational genome annotation of sequences obtained by next-generation sequencing. **Biology**, v. 9, n. 9, p. 295, 18 set. 2020.

EL-SAYED, N. M. Comparative genomics of trypanosomatid parasitic protozoa. **Science**, v. 309, n. 5733, p. 404–409, 15 jul. 2005.

EMMS, D. M.; KELLY, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. **Genome Biology**, v. 20, n. 1, p. 238, dez. 2019.

EWELS, P. et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, v. 32, n. 19, p. 3047–3048, 1 out. 2016.

FELLOWS, I. **wordcloud: Word Clouds**. [s.l: s.n.].

FIGUEIREDO, F. B. et al. Leishmaniose visceral canina: dois casos autóctones no município de Florianópolis, estado de Santa Catarina. **Acta Scientiae Veterinariae**, v. 40, n. 1, 2012.

FINN, R. D.; CLEMENTS, J.; EDDY, S. R. HMMER web server: interactive sequence similarity searching. **Nucleic Acids Research**, v. 39, n. suppl, p. W29–W37, 1 jul. 2011.

FOSKETT, J. K. et al. Inositol trisphosphate receptor Ca^{2+} release channels. **Physiological Reviews**, v. 87, n. 2, p. 593–658, abr. 2007.

GALVIS-OVALLOS, F. et al. Canine visceral leishmaniasis in the metropolitan area of São Paulo: *Pintomyia fischeri* as potential vector of *Leishmania infantum*. **Parasite**, v. 24, p. 2, 2017.

GALVIS-OVALLOS, F. et al. Detection of *Pintomyia fischeri* (Diptera: Psychodidae) with *Leishmania infantum* (Trypanosomatida: Trypanosomatidae) promastigotes in a focus of visceral leishmaniasis in brazil. **Journal of Medical Entomology**, v. 58, n. 2, p. 830–836, 12 mar. 2021.

GARRISON, E. et al. **Vcflib and tools for processing the VCF variant call format**. [s.l.] Bioinformatics, 23 maio 2021. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2021.05.21.445151>. Acesso em: 10 set. 2021.

GARRISON, E.; MARTH, G. Haplotype-based variant detection from short-read sequencing. **arXiv:1207.3907 [q-bio]**, 20 jul. 2012.

GHARBI, M. et al. Leishmaniosis (*Leishmania infantum* infection) in dogs. **Revue Scientifique et Technique de l'OIE**, v. 34, n. 2, p. 613–626, 1 ago. 2015.

GHOSH, J. et al. Hyperlipidemia offers protection against *Leishmania donovani* infection: role of membrane cholesterol. **Journal of Lipid Research**, v. 53, n. 12, p. 2560–2572, dez. 2012.

GOLDBERG, A. L. et al. New insights into the mechanisms and importance of the proteasome in intracellular protein degradation. **Biological Chemistry**, v. 378, n. 3–4, p. 131–140, abr. 1997.

GONZÁLEZ-DE LA FUENTE, S. et al. Resequencing of the *Leishmania infantum* (strain JPCM5) genome and de novo assembly into 36 contigs. **Scientific Reports**, v. 7, n. 1, p. 18050, dez. 2017.

GOODFELLOW, S. J.; ZOMERDIJK, J. C. B. M. Basic mechanisms in RNA polymerase I transcription of the ribosomal RNA genes. In: KUNDU, T. K. (Ed.). . **Epigenetics: Development and Disease**. Subcellular Biochemistry. Dordrecht: Springer Netherlands, 2013. v. 61p. 211–236.

GRÜNEBAST, J.; CLOS, J. *Leishmania*: responding to environmental signals and challenges without regulated transcription. **Computational and Structural Biotechnology Journal**, v. 18, p. 4016–4023, 2020.

GU, Z.; EILS, R.; SCHLESNER, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. **Bioinformatics**, 2016.

GUREVICH, A. et al. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072–1075, 15 abr. 2013.

HARHAY, M. O. et al. Urban parasitology: visceral leishmaniasis in Brazil. **Trends in Parasitology**, v. 27, n. 9, p. 403–409, set. 2011.

HERRERA, G. et al. Evaluation of a Multilocus Sequence Typing (MLST) scheme for *Leishmania (Viannia) braziliensis* and *Leishmania (Viannia) panamensis* in Colombia. **Parasites & Vectors**, v. 10, n. 1, p. 236, dez. 2017.

HEYDARI, M. et al. Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. **BMC Bioinformatics**, v. 20, n. 1, p. 298, dez. 2019.

HOFF, K. J.; STANKE, M. WebAUGUSTUS – a web service for training AUGUSTUS and predicting genes in eukaryotes. **Nucleic Acids Research**, v. 41, n. W1, p. W123–W128, 1 jul. 2013.

HOLLAND, I. B.; BLIGHT, M. A. ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. **Journal of Molecular Biology**, v. 293, n. 2, p. 381–399, out. 1999.

HOLMGREN, A. M.; MCCONKEY, C. A.; SHIN, S. Outrunning the Red Queen: bystander activation as a means of outpacing innate immune subversion by intracellular pathogens. **Cellular & Molecular Immunology**, v. 14, n. 1, p. 14–21, jan. 2017.

HORNER, D. S. et al. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. **Briefings in Bioinformatics**, v. 11, n. 2, p. 181–197, 1 mar. 2010.

ILLUMINA, I. Quality scores for next-generation sequencing. **Technical Note: Informatics**, v. 31, 2011.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 15 fev. 2001.

IVENS, A. C. The genome of the kinetoplastid parasite, *Leishmania major*. **Science**, v. 309, n. 5733, p. 436–442, 15 jul. 2005.

JACKSON, A. P. The evolution of amastin surface glycoproteins in trypanosomatid parasites. **Molecular Biology and Evolution**, v. 27, n. 1, p. 33–45, 1 jan. 2010.

JERONIMO, S. M. B. et al. An urban outbreak of visceral leishmaniasis in Natal, Brazil. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 88, n. 4, p. 386–388, jul. 1994.

JONES, P. et al. InterProScan 5: genome-scale protein function classification. **Bioinformatics**, v. 30, n. 9, p. 1236–1240, 1 maio 2014.

JUNG, H. et al. Twelve quick steps for genome assembly and annotation in the classroom. **PLOS Computational Biology**, v. 16, n. 11, p. e1008325, 12 nov. 2020.

KALVARI, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. **Nucleic Acids Research**, v. 46, n. D1, p. D335–D342, 4 jan. 2018.

KARAMYSHEVA, Z. N.; GUARNIZO, S. A. G.; KARAMYSHEV, A. L. Regulation of translation in the protozoan parasite *Leishmania*. **International Journal of Molecular Sciences**, v. 21, n. 8, p. 2981, 23 abr. 2020.

KAZEMI, B. Genomic organization of *Leishmania* species. **Iranian Journal of Parasitology**, v. 6, n. 3, p. 1–18, ago. 2011.

KILLICK-KENDRICK, R. The life-cycle of *Leishmania* in the sandfly with special reference to the form infective to the vertebrate host. **Annales de Parasitologie Humaine et Comparée**, v. 65, p. 37–42, 1990.

KRASSNER, S. M. Proline metabolism in *Leishmania tarentolae*. **Experimental Parasitology**, v. 24, n. 3, p. 348–363, jun. 1969.

KRASSNER, S. M.; FLORY, B. Proline metabolism in *Leishmania donovani* promastigotes. **The Journal of Protozoology**, v. 19, n. 4, p. 682–685, nov. 1972.

LAKSHMI, B. S.; WANG, R.; MADHUBALA, R. *Leishmania* genome analysis and high-throughput immunological screening identifies tuzin as a novel vaccine candidate against visceral leishmaniasis. **Vaccine**, v. 32, n. 30, p. 3816–3822, jun. 2014.

LEIFSO, K. et al. Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: the *Leishmania* genome is constitutively expressed. **Molecular and Biochemical Parasitology**, v. 152, n. 1, p. 35–46, mar. 2007.

LEINONEN, R. et al. The Sequence Read Archive. **Nucleic Acids Research**, v. 39, n. Database, p. D19–D21, 1 jan. 2011.

LEPROHON, P. et al. Drug resistance analysis by next generation sequencing in *Leishmania*. **International Journal for Parasitology: Drugs and Drug Resistance**, v. 5, n. 1, p. 26–35, abr. 2015.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 15 ago. 2009.

LI, H. **Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences**. Disponível em: <https://github.com/lh3/seqtk>.

LI, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **arXiv:1303.3997 [q-bio]**, 26 maio 2013b.

LIAO, P.; SATTEN, G. A.; HU, Y.-J. PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. **Genetic Epidemiology**, v. 41, n. 5, p. 375–387, jul. 2017.

LIU, L. et al. Comparison of next-generation sequencing systems. **Journal of Biomedicine and Biotechnology**, v. 2012, p. 1–11, 2012.

LOCKHART, D. J.; WINZELER, E. A. Genomics, gene expression and DNA arrays. **Nature**, v. 405, n. 6788, p. 827–836, jun. 2000.

LOWE, T. M.; EDDY, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic Acids Research**, v. 25, n. 5, p. 955–964, 1 mar. 1997.

LUKEŠ, J. et al. Kinetoplast DNA network: evolution of an improbable structure. **Eukaryotic Cell**, v. 1, n. 4, p. 495–502, ago. 2002.

MAHMUD, I. C. et al. Epidemiological aspects of the first human autochthonous visceral leishmaniosis cases in Porto Alegre, Brazil. **The Brazilian Journal of Infectious Diseases**, v. 23, n. 2, p. 124–129, mar. 2019.

MAIA, G. A. **Ferramenta integrada para anotação de proteínas hipotéticas: estudo de caso utilizando análises proteogenômicas em *Trypanosoma rangeli***. Dissertação (Mestrado em Biotecnologia e Biociências)—Florianópolis: Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, 2019.

MARCILI, A. et al. Phylogenetic relationships of Leishmania species based on trypanosomatid barcode (SSU rDNA) and gGAPDH genes: taxonomic revision of *Leishmania (L.) infantum chagasi* in South America. **Infection, Genetics and Evolution**, v. 25, p. 44–51, jul. 2014.

MARTIN, J. A.; WANG, Z. Next-generation transcriptome assembly. **Nature Reviews Genetics**, v. 12, n. 10, p. 671–682, out. 2011.

MARTÍNEZ, C. R.; RUIZ, C. J. Alterations in host lipid metabolism produced during visceral leishmaniasis infections. **Current Tropical Medicine Reports**, v. 6, n. 4, p. 250–255, dez. 2019.

MARTÍNEZ-CALVILLO, S.; STUART, K.; MYLER, P. J. Ploidy changes associated with disruption of two adjacent genes on *Leishmania major* chromosome 1. **International Journal for Parasitology**, v. 35, n. 4, p. 419–429, abr. 2005.

MATHE, C. Current methods of gene prediction, their strengths and weaknesses. **Nucleic Acids Research**, v. 30, n. 19, p. 4103–4117, 1 out. 2002.

MAZAREB, S.; FU, Z. Y.; ZILBERSTEIN, D. Developmental regulation of proline transport in *Leishmania donovani*. **Experimental Parasitology**, v. 91, n. 4, p. 341–348, abr. 1999.

MAZIERO, N. et al. Rural–urban focus of canine visceral leishmaniosis in the far western region of Santa Catarina State, Brazil. **Veterinary Parasitology**, v. 205, n. 1–2, p. 92–95, set. 2014.

MCCALL, L.-I.; MATLASHEWSKI, G. Localization and induction of the A2 virulence factor in *Leishmania*: evidence that A2 is a stress response protein. **Molecular Microbiology**, v. 77, n. 2, p. 518–530, 24 maio 2010.

MCKINNEY, W. **Data structures for statistical computing in Python**. (S. van der Walt, J. Millman, Eds.)Proceedings of the 9th Python in Science Conference. **Anais**...2010.

MEDINI, D. et al. The microbial pan-genome. **Current Opinion in Genetics & Development**, v. 15, n. 6, p. 589–594, dez. 2005.

MENEZES, J. P. DE; SARAIVA, E. M.; ROCHA-AZEVEDO, B. DA. The site of the bite: *Leishmania* interaction with macrophages, neutrophils and the extracellular matrix in the dermis. **Parasites & Vectors**, v. 9, n. 1, p. 264, dez. 2016.

METZKER, M. L. Sequencing technologies — the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31–46, jan. 2010.

MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. **Genomics**, v. 95, n. 6, p. 315–327, jun. 2010.

MOLINA-MORA, J. A. et al. High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 *Pseudomonas aeruginosa* genome: benchmark of hybrid and non-hybrid assemblers. **Scientific Reports**, v. 10, n. 1, p. 1392, dez. 2020.

MORENO, J. et al. Immunization with H1, HASPB1 and MML Leishmania proteins in a vaccine trial against experimental canine leishmaniasis. **Vaccine**, v. 25, n. 29, p. 5290–5300, jul. 2007.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. **Genomics**, v. 92, n. 5, p. 255–264, nov. 2008.

MUÑOZ, C. et al. Role of the ubiquitin-proteasome systems in the biology and virulence of protozoan parasites. **BioMed Research International**, v. 2015, p. 1–13, 2015.

MURO, E. M.; MAH, N.; ANDRADE-NAVARRO, M. A. Functional evidence of post-transcriptional regulation by pseudogenes. **Biochimie**, v. 93, n. 11, p. 1916–1921, nov. 2011.

MUXEL, S. M. et al. Arginine and polyamines fate in *Leishmania* infection. **Frontiers in Microbiology**, v. 8, p. 2682, 15 jan. 2018.

NADALIN, F.; VEZZI, F.; POLICRITI, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. **BMC Bioinformatics**, v. 13, n. S14, p. S8, set. 2012.

NAWROCKI, E. P.; EDDY, S. R. Infernal 1.1: 100-fold faster RNA homology searches. **Bioinformatics**, v. 29, n. 22, p. 2933–2935, 15 nov. 2013.

NEWPORT, M. J.; FINAN, C. Genome-wide association studies and susceptibility to infectious diseases. **Briefings in Functional Genomics**, v. 10, n. 2, p. 98–107, 1 mar. 2011.

OLIVEIRA, L. C. DE; MOREIRA, N. M. Epidemiological aspects of visceral leishmaniasis in Brazil and in international border regions. **Research, Society and Development**, v. 10, n. 12, p. e549101220684, 1 out. 2021.

OLIVIER, M.; BAIMBRIDGE, K. G.; REINER, N. E. Stimulus-response coupling in monocytes infected with *Leishmania*. Attenuation of calcium transients is related to defective agonist-induced accumulation of inositol phosphates. **Journal of Immunology (Baltimore, Md.: 1950)**, v. 148, n. 4, p. 1188–1196, 15 fev. 1992.

ORTUÑO, M. et al. Genetic diversity and phylogenetic relationships between *Leishmania infantum* from dogs, humans and wildlife in south-east Spain. **Zoonoses and Public Health**, v. 66, n. 8, p. 961–973, dez. 2019.

OTTINO, J. **Caracterização comparativa da variabilidade genômica de isolados de *Leishmania* spp. obtidos de cães com leishmaniose visceral em diferentes áreas endêmicas no Brasil**. Tese (Doutorado em Parasitologia)—Belo Horizonte: Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 2021.

PABLOS, L. D.; FERREIRA, T.; WALRAD, P. Developmental differentiation in *Leishmania* lifecycle progression: post-transcriptional control conducts the orchestra. **Current Opinion in Microbiology**, v. 34, p. 82–89, dez. 2016.

PAIVA, R. M. C. DE et al. Amastin knockdown in *Leishmania braziliensis* affects parasite-macrophage interaction and results in impaired viability of intracellular amastigotes. **PLOS Pathogens**, v. 11, n. 12, p. e1005296, 7 dez. 2015.

PAL, S. K.; BANDYOPADHYAY, S.; RAY, S. S. Evolutionary computation in bioinformatics: a review. **IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)**, v. 36, n. 5, p. 601–615, set. 2006.

PAN AMERICAN HEALTH ORGANIZATION. **Leishmaniasis: epidemiological report of the Americas**. [s.l: s.n.]. Disponível em: <https://iris.paho.org/handle/10665.2/53090>.

PEACOCK, C. S. et al. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. **Nature Genetics**, v. 39, n. 7, p. 839–847, jul. 2007.

PETIT, R. J. Early insights into the genetic consequences of range expansions. **Heredity**, v. 106, n. 2, p. 203–204, fev. 2011.

PEVSNER, J. **Bioinformatics and functional genomics**. 2nd edition ed. [s.l.] Wiley, 2015.

PITA-PEREIRA, D. DE et al. *Lutzomyia (Pintomyia) fischeri* (Diptera: Psychodidae: Phlebotominae), a probable vector of American Cutaneous Leishmaniasis: Detection of natural infection by *Leishmania (Viannia)* DNA in specimens from the municipality

of Porto Alegre (RS), Brazil, using multiplex PCR assay. **Acta Tropica**, v. 120, n. 3, p. 273–275, dez. 2011.

POP, M. Genome assembly reborn: recent computational challenges. **Briefings in Bioinformatics**, v. 10, n. 4, p. 354–366, 1 jul. 2009.

POTTER, S. C. et al. HMMER web server: 2018 update. **Nucleic Acids Research**, v. 46, n. W1, p. W200–W204, 2 jul. 2018.

PRITCHARD, L. et al. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. **Analytical Methods**, v. 8, n. 1, p. 12–24, 2016.

PUCADYIL, T. J. et al. Cholesterol is required for *Leishmania donovani* infection: implications in leishmaniasis. **Molecular and Biochemical Parasitology**, v. 133, n. 2, p. 145–152, fev. 2004.

R CORE TEAM. **R: a language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2020.

R CORE TEAM. **R: a language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2021.

RAVEL, C. High conservation of the fine-scale organisation of chromosome 5 between two pathogenic *Leishmania* species. **Nucleic Acids Research**, v. 27, n. 12, p. 2473–2477, 15 jun. 1999.

REED, J. L. et al. Towards multidimensional genome annotation. **Nature Reviews Genetics**, v. 7, n. 2, p. 130–141, fev. 2006.

RÊGO, F. D. et al. Ecology and molecular detection of *Leishmania infantum* Nicolle, 1908 (Kinetoplastida: Trypanosomatida) in wild-caught sand flies (Psychodidae: Phlebotominae) collected in Porto Alegre, Rio Grande do Sul: a new focus of visceral leishmaniasis in Brazil. **Journal of Medical Entomology**, v. 56, n. 2, p. 519–525, 25 fev. 2019.

RÊGO, F. D. et al. Potential vectors of *Leishmania* parasites in a recent focus of visceral leishmaniasis in neighborhoods of Porto Alegre, State of Rio Grande do Sul, Brazil. **Journal of Medical Entomology**, v. 57, n. 4, p. 1286–1292, 4 jul. 2020.

REUTER, J. A.; SPACEK, D. V.; SNYDER, M. P. High-throughput sequencing technologies. **Molecular Cell**, v. 58, n. 4, p. 586–597, maio 2015.

RNACENTRAL CONSORTIUM et al. RNAcentral: a hub of information for non-coding RNA sequences. **Nucleic Acids Research**, v. 47, n. D1, p. D221–D229, 8 jan. 2019.

ROGERS, M. B. et al. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. **Genome Research**, v. 21, n. 12, p. 2129–2142, 1 dez. 2011.

RUIZ-POSTIGO, J. A. et al. **Global leishmaniasis surveillance: 2019–2020, a baseline for the 2030 roadmap**. [s.l: s.n.]. Disponível em: <https://www.who.int/publications/i/item/who-wer9635-401-419>.

SAKKAS, H.; GARTZONIKA, C.; LEVIDIOTOU, S. Laboratory diagnosis of human visceral leishmaniasis. **Journal of Vector Borne Diseases**, v. 53, n. 1, p. 8–16, mar. 2016.

SALMENA, L. Pseudogenes: Four Decades of Discovery. In: POLISENO, L. (Ed.). . **Pseudogenes**. Methods in Molecular Biology. New York, NY: Springer US, 2021. v. 2324p. 3–18.

SALOMÓN, O. D. et al. *Lutzomyia migonei* as putative vector of visceral leishmaniasis in La Banda, Argentina. **Acta Tropica**, v. 113, n. 1, p. 84–87, jan. 2010.

SAMARASINGHE, S. R. et al. Genomic insights into virulence mechanisms of *Leishmania donovani*: evidence from an atypical strain. **BMC Genomics**, v. 19, n. 1, p. 843, dez. 2018.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences**, v. 74, n. 12, p. 5463–5467, 1 dez. 1977.

SAURIN, W.; HOFNUNG, M.; DASSA, E. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. **Journal of Molecular Evolution**, v. 48, n. 1, p. 22–41, jan. 1999.

SCHWABL, P. et al. Colonization and genetic diversification processes of Leishmania infantum in the Americas. **Communications Biology**, v. 4, n. 1, p. 139, dez. 2021.

SEMINI, G. et al. Changes to cholesterol trafficking in macrophages by *Leishmania* parasites infection. **MicrobiologyOpen**, v. 6, n. 4, p. e00469, ago. 2017.

SHARIAT, B. et al. HyDA-Vista: towards optimal guided selection of k-mer size for sequence assembly. **BMC Genomics**, v. 15, n. S10, p. S9, dez. 2014.

SHLOMAI, J. The structure and replication of kinetoplast DNA. **Current Molecular Medicine**, v. 4, n. 6, p. 623–647, 1 set. 2004.

SILVA-JARDIM, I.; FÁTIMA HORTA, M.; RAMALHO-PINTO, F. J. The *Leishmania chagasi* proteasome: role in promastigotes growth and amastigotes survival within murine macrophages. **Acta Tropica**, v. 91, n. 2, p. 121–130, jul. 2004.

SINHA, A.; SARKAR, S. Ubiquitin-proteasome system – a target to control pathogenic protozoa. **Microbial Pathogens and Strategies for Combating Them: Science, Technology and Education**, v. 1, p. 764–73, 2013.

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

SMITH, B.; WILLIAMS, J.; SCHULZE-KREMER, S. The ontology of the gene ontology. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, p. 609–613, 2003.

SNYDER, M.; GERSTEIN, M. Genomics: defining genes in the genomics era. **Science**, v. 300, n. 5617, p. 258–260, 11 abr. 2003.

SOARES, N. M. et al. Plasma lipoproteins in visceral leishmaniasis and their effect on *Leishmania* -infected macrophages. **Parasite Immunology**, v. 32, n. 4, p. 259–266, abr. 2010.

SOHN, J.; NAM, J.-W. The present and future of *de novo* whole-genome assembly. **Briefings in Bioinformatics**, p. bbw096, 14 out. 2016.

SOLLNER-WEBB, B.; MOUGEY, E. B. News from the nucleolus: rRNA gene expression. **Trends in Biochemical Sciences**, v. 16, p. 58–62, jan. 1991.

SONENBERG, N.; HINNEBUSCH, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. **Cell**, v. 136, n. 4, p. 731–745, fev. 2009.

SOUZA, N. P. et al. *Leishmania (Leishmania) infantum chagasi* em canídeos silvestres mantidos em cativeiro, no Estado de Mato Grosso. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 43, n. 3, p. 333–335, jun. 2010.

STANKE, M.; WAACK, S. Gene prediction with a hidden Markov model and a new intron submodel. **Bioinformatics**, v. 19, n. Suppl 2, p. ii215–ii225, 27 set. 2003.

STEIN, L. Genome annotation: from sequence to biology. **Nature Reviews Genetics**, v. 2, n. 7, p. 493–503, jul. 2001.

STEINDEL, M. et al. Outbreak of autochthonous canine visceral leishmaniasis in Santa Catarina, Brazil. **Pesquisa Veterinária Brasileira**, v. 33, n. 4, p. 490–496, abr. 2013.

STOLER, N.; NEKRUTENKO, A. Sequencing error profiles of Illumina sequencing instruments. **NAR Genomics and Bioinformatics**, v. 3, n. 1, p. lqab019, 6 jan. 2021.

SUNTER, J. D.; GULL, K. The Flagellum Attachment Zone: “the cellular ruler” of trypanosome morphology. **Trends in Parasitology**, v. 32, n. 4, p. 309–324, abr. 2016.

SUNTER, J.; GULL, K. Shape, form, function and *Leishmania* pathogenicity: from textbook descriptions to biological understanding. **Open Biology**, v. 7, n. 9, p. 170165, set. 2017.

SUPEK, F. et al. REVIGO summarizes and visualizes long lists of Gene Ontology terms. **PLoS ONE**, v. 6, n. 7, p. e21800, 18 jul. 2011.

SVIRIDOV, D.; BUKRINSKY, M. Interaction of pathogens with host cholesterol metabolism. **Current Opinion in Lipidology**, v. 25, n. 5, p. 333–338, out. 2014.

TANAKA, K. The proteasome: overview of structure and functions. **Proceedings of the Japan Academy, Series B**, v. 85, n. 1, p. 12–36, 2009.

TEIXEIRA, D. G. et al. Comparative analyses of whole genome sequences of *Leishmania infantum* isolates from humans and dogs in northeastern Brazil. **International Journal for Parasitology**, v. 47, n. 10–11, p. 655–665, set. 2017.

TEIXEIRA, S. M. R.; KIRCHHOFF, L. V.; DONELSON, J. E. *Trypanosoma cruzi*: suppression of tuzin gene expression by Its 5′-UTR and spliced leader addition site. **Experimental Parasitology**, v. 93, n. 3, p. 143–151, nov. 1999.

TETTELIN, H. et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. **Proceedings of the National Academy of Sciences**, v. 102, n. 39, p. 13950–13955, 27 set. 2005.

TORRES-GUERRERO, E. et al. Leishmaniasis: a review. **F1000Research**, v. 6, p. 750, 26 maio 2017.

TUTAR, Y. Pseudogenes. **Comparative and Functional Genomics**, v. 2012, p. 1–4, 2012.

ULIANA, S. R. B.; RUIZ, J. C.; CRUZ, A. K. *Leishmania* genomics: where do we stand? In: GRUBER, A. et al. (Eds.). . **Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach**. Bethesda, MD: NCBI, 2008. p. 24.

VALDIVIA, H. O. et al. Comparative genomics of canine-isolated *Leishmania (Leishmania) amazonensis* from an endemic focus of visceral leishmaniasis in Governador Valadares, southeastern Brazil. **Scientific Reports**, v. 7, n. 1, p. 40804, fev. 2017.

VAN DEN BERGE, K. et al. RNA sequencing data: hitchhiker's guide to expression analysis. **Annual Review of Biomedical Data Science**, v. 2, n. 1, p. 139–173, 20 jul. 2019.

VENTER, J. C. et al. The sequence of the human genome. **Science**, v. 291, n. 5507, p. 1304–1351, 16 fev. 2001.

VERLI, H. **Bioinformática: da biologia à flexibilidade molecular**. 1st edition ed. [s.l.] Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

VERNIKOS, G. et al. Ten years of pan-genome analyses. **Current Opinion in Microbiology**, v. 23, p. 148–154, fev. 2015.

WALKER, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. **PLoS ONE**, v. 9, n. 11, p. e112963, 19 nov. 2014.

WESTROP, G. D. et al. Metabolomic analyses of *Leishmania* reveal multiple species differences and large differences in amino acid metabolism. **PLOS ONE**, v. 10, n. 9, p. e0136891, 14 set. 2015.

WICKHAM, H. ***ggplot2*: elegant graphics for data analysis**. [s.l.] Springer-Verlag New York, 2016.

WICKHAM, H. et al. Welcome to the Tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 21 nov. 2019.

WINCKER, P. et al. The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. **Nucleic Acids Research**, v. 24, n. 9, p. 1688–1694, 1 maio 1996.

YAN, S. et al. Characterization of the *Leishmania donovani* ribosomal RNA promoter. **Molecular and Biochemical Parasitology**, v. 103, n. 2, p. 197–210, out. 1999.

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

YE, J. et al. WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. **Nucleic Acids Research**, v. 46, n. W1, p. W71–W75, 2 jul. 2018.

ZEKIC, T.; HOLLEY, G.; STOYE, J. Pan-genome storage and analysis techniques. In: SETUBAL, J. C.; STOYE, J.; STADLER, P. F. (Eds.). . **Comparative Genomics**. Methods in Molecular Biology. New York, NY: Springer New York, 2018. v. 1704p. 29–53.

ZHANG, C.-Y. et al. Phylogenetic and evolutionary analysis of chinese <>Leishmania</i> isolates based on multilocus sequence typing. **PLoS ONE**, v. 8, n. 4, p. e63124, 30 abr. 2013.

ZHANG, W. W. et al. Genetic analysis of *Leishmania donovani* tropism using a naturally attenuated cutaneous strain. **PLoS Pathogens**, v. 10, n. 7, p. e1004244, 3 jul. 2014.

ZHANG, W.-W.; MATLASHEWSKI, G. Screening *Leishmania donovani*-specific genes required for visceral infection. **Molecular Microbiology**, v. 77, n. 2, p. 505–517, 1 jun. 2010.

ZHANG, W.-W.; MATLASHEWSKI, G. Deletion of an ATP-binding cassette protein subfamily C transporter in *Leishmania donovani* results in increased virulence. **Molecular and Biochemical Parasitology**, v. 185, n. 2, p. 165–169, out. 2012.

ZHANG, Z. et al. PseudoPipe: an automated pseudogene identification pipeline. **Bioinformatics**, v. 22, n. 12, p. 1437–1439, 15 jun. 2006.

ZHAO, Y. et al. PGAP: pan-genomes analysis pipeline. **Bioinformatics**, v. 28, n. 3, p. 416–418, 1 fev. 2012.

ZHAO, Y. et al. PGAP-X: extension on pan-genome analysis pipeline. **BMC Genomics**, v. 19, n. S1, p. 36, jan. 2018.

APÊNDICE A – Adaptadores removidos na etapa de controle de qualidade

A **Tabela 11** apresenta as sequências dos potenciais adaptadores da plataforma Illumina (`ILLUMINACLIP:all_adapters.fa`) removidos pelo programa *Trimmomatic* v0.39.

Tabela 11. Sequências presentes no arquivo de adaptadores `all_adapters.fa` removidos dos dados brutos pelo programa *Trimmomatic*.

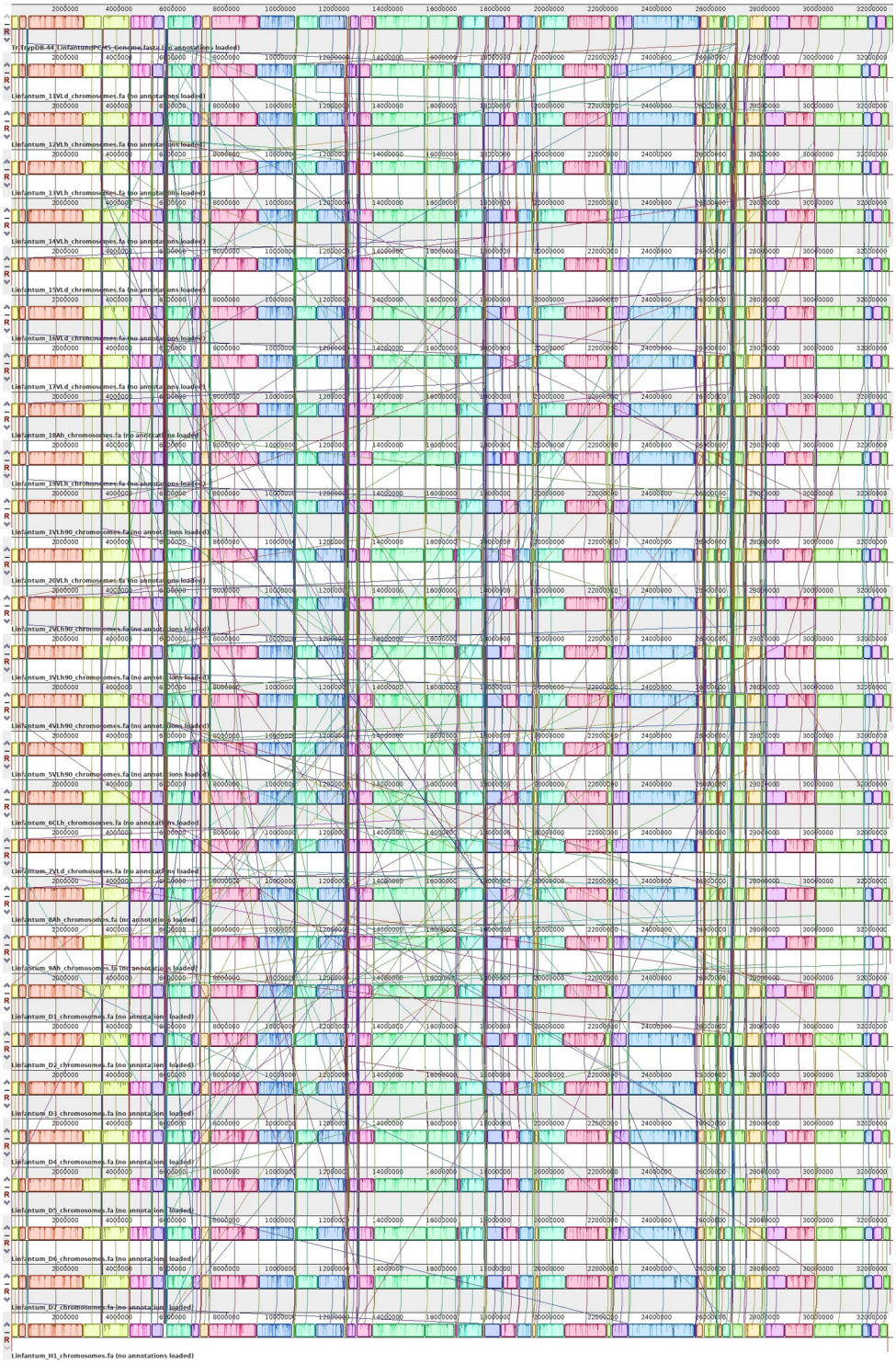
ADAPTADOR	SEQUÊNCIA
PrefixNX-1	AGATGTGTATAAGAGACAG
PrefixNX-2	AGATGTGTATAAGAGACAG
Trans1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Trans1_rc	CTGTCTCTTATACACATCTGACGCTGCCGACGA
Trans2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
Trans2_rc	CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
PrefixPE-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PrefixPE-2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PCR_Primer1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PCR_Primer1_rc	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
PCR_Primer2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PCR_Primer2_rc	AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG
FlowCell1	TTTTTTTTTTAATGATACGGCGACCACCGAGATCTACAC
FlowCell2	TTTTTTTTTTCAAGCAGAAGACGGCATACGA
TruSeq2_SE	AGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
TruSeq2_PE_f	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

ADAPTADOR	SEQUÊNCIA
TruSeq2_PE_r	AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PrefixPE/1	TACACTCTTTCCCTACACGACGCTCTTCCGATCT
PrefixPE/2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
PE1	TACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE1_rc	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA
PE2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
PE2_rc	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
PrefixPE-1	TACACTCTTTCCCTACACGACGCTCTTCCGATCT
PrefixPE-2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq3_IndexedAdapter	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
TruSeq3_UniversalAdapter	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA
TruSeq_Index9	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGC CGTCTTCTGCTTG

APÊNDICE B – Alinhamentos múltiplos para todas as amostras de *Leishmania infantum* resultantes do *progressiveMauve*

A **Figura 14** é resultante do alinhamento completo entre os cromossomos montados e o genoma de referência. Blocos sintênicos apresentam a mesma coloração enquanto possíveis rearranjos ou variações são indicados por linhas que ligam as diferentes amostras.

Figura 14. Alinhamento de genomas completos para montagens de *Leishmania infantum*.



O primeiro genoma se refere ao genoma de referência da espécie (cepa JPCM5).