UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Anderson Coqueiro dos Santos

MONTAGEM HÍBRIDA E ANÁLISES DE ANEUPLOIDIAS EM GENOMAS COMPLEXOS

Trypanosoma cruzi CL Brener como modelo

Belo Horizonte – MG
2023

Anderson Coqueiro dos Santos

MONTAGEM HÍBRIDA E ANÁLISES DE ANEUPLOIDIAS EM GENOMAS COMPLEXOS

Trypanosoma cruzi CL Brener como modelo

Tese apresentada ao Programa de Pós-graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito para obtenção do título de Doutor em Bioinformática.

Orientadora: Daniella Castanheira Bartholomeu
Coorientador: João Luís Reis Cunha

Belo Horizonte – MG
2023

043 Santos, Anderson Coqueiro dos.
Montagem híbrida e análises de aneuploidias em genomas complexos: Trypanosoma cruzi CL Brener como modelo [manuscrito] / Anderson Coqueiro dos Santos. – 2023.
135 f. : il. ; 29,5 cm.

Orientadora: Daniella Castanheira Bartholomeu. Coorientador: João Luís Reis Cunha.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Trypanosoma cruzi. 3. Genoma. 4. Aneuploidia. I. Bartholomeu, Daniella Castanheira. II. Cunha, João Luís Reis. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DA DEFESA DE TESE

ANDERSON COQUEIRO DOS SANTOS

Às nove horas do dia **25 de agosto de 2023**, reuniu-se, no aplicativo Zoom, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Montagem híbrida e análises de aneuploidias em genomas complexos: Trypanosoma cruzi CL Brener como modelo"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, a Presidente da Comissão, **Dra. Daniella Castanheira Bartholomeu**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	**Instituição**	**Indicação**
Dra. Daniella Castanheira Bartholomeu	Universidade Federal de Minas Gerais	**Aprovado**
Dr. João Luís Reis Cunha	University of York	**Aprovado**
Dr. Francisco Pereira Lobo	Universidade Federal de Minas Gerais	**Aprovado**
Dr. Rodrigo de Paula Baptista	Houston Methodist Research Institute	**Aprovado**
Dr. Leonardo Koerich	Universidade Federal de Minas Gerais	**Aprovado**
Dr. Fabiano Sviatopolk-Mirsky Pais	University of York	**Aprovado**

Pelas indicações, o candidato foi considerado: **Aprovado**.

O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 25 de agosto de 2023.

Documento assinado eletronicamente por **Fabiano Sviatopolk Mirsky Pais**, **Usuário Externo**, em 25/08/2023, às 12:25, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rodrigo de Paula Baptista**, **Usuário Externo**, em 25/08/2023, às 13:23, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Daniella Castanheira Bartholomeu**, **Professora do Magistério Superior**, em 25/08/2023, às 15:03, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Joao Luis Reis Cunha**, **Cidadão**, em 31/08/2023, às 11:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 05/09/2023, às 11:50, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Barbosa Koerich**, **Professor do Magistério Superior**, em 05/09/2023, às 19:51, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2555473** e o código CRC **331469F3**.

AGRADECIMENTOS

Durante toda a jornada em que se desenvolveu esse trabalho e os demais que aqui não estão presentes eu pude conhecer muitas pessoas. Eu tenho a todas, mesmo aqui não citadas, que agradecer pelo conhecimento e confiança. Dentre todas o meu maior agradecimento vai a minha orientadora Dani. Não seria 1/5 do que sou hoje sem os ensinamentos e oportunidades que ela confiou a mim. Além disso, os conselhos não profissionais ou acadêmicos, que me propiciou a ver um mundo de outra forma e que me ajudaram a me moldar como pessoa. Graças a ela, e ao Leandro Martins orientador da graduação que a me apresentou, pude sair de uma cidade pequena no interior da Bahia e trabalhar com aquilo que gosto e ser conhecido por isso.

Outro agradecimento que tenho a fazer é ao João. Meu coorientador, um grande ser humano, uma das melhores pessoas que pude conhecer tanto no âmbito profissional quanto no pessoal. Até hoje me lembro das conversas que tínhamos durante o almoço, e como cada vez fica mais claro para mim os conselhos que me deu e conversas que banhadas e risadas e nerdices me ajudaram, me tornei uma pessoa melhor graças a ele.

Agradeço ainda a Laila por ter tornado meus dias mais leves, sem falar do mundo musical que me apresentou. A Michele por momentos magníficos. A Mari Cardoso, Vanessa, Ana Leão, Samuel, Lucas, Agostinho, Ana Clara, e todos membros do laboratório que pude conviver. Desses há aqueles que se tornaram uma família para mim em Belo Horizonte, Maria Climaco e Thais Eloi que apesar de nos encontramos todos os dias vemos formas de se ver mais. Além é claro do Breno e do Douglas que me proporcionaram risos e mais risos. Eu gostaria de agradecer as pessoas que pude

conhecer em minha nova jornada. Ao Gustavo, Pedro, Nara, Ilária, Amanda e a Rose por fazerem cada dia único.

Não poderia esquecer de todos os professores em sala de aula e aqueles com quem pude conviver, e que proporcionaram minha evolução acadêmica. Aos membros do colegiado pelo atendimento e ajuda quando precisei. Aos membros convidados a participar da banca pelo aceite. É uma honra tê-los presente. As agências de fomento e instituições que permitiram que este trabalho fosse realizado.

E é claro a minha família, aquela que me fez chegar até onde estou como minha mãe, pai e tios e tias, e aquela que pude formar. Dessa eu tenho uma agradecimento enorme a “minha senhora”, a Denise por me presentear com três presentes magníficos: o Otto, Ravi e ela mesmo. Sem vocês não saberia dizer quem eu sou. Muito obrigado pelo carinho, apoio, auxílio e amor.

RESUMO

Trypanosoma cruzi, o agente etiológico da doença de Chagas, é um protozoário unicelular flagelado que teve a primeira versão do seu genoma sequenciado e publicado em 2005. A cepa selecionada foi a CL Brener, de linhagem híbrida entre os DTUs TcII e TcIII. Essa característica híbrida, além da grande quantidade de membros de famílias multigênicas, associado a outros elementos repetitivos do genoma de *T. cruzi* comprometeu a qualidade da montagem. Devido a isso, neste trabalho uma nova montagem foi realizada utilizando de sequenciamento de reads longas (PacBio) combinado a reads curtas (Illumina), e reads de Sanger de BACs e Fosmídeos geradas pelo projeto genoma de 2005 para auxiliar a montagem. Para tal, diferentes montadores foram testados, como o Canu e HGAP, para construção dos contigs. O scaffolding foi realizado de modo interativo, reduzindo a cada iteração o número de reads usado para juntar contigs, permitindo a montagem de regiões com melhor suporte primeiro. Um total de 446 sequências foram obtidas ao fim da montagem, seguida da correção das mesmas utilizando reads curtas. Uma anotação *de novo* desta nova montagem foi realizada utilizando o programa Augustus tendo como base dados a anotação disponível no TritrypDB da cepa CL Brener, bem como de outras cepas já anotadas. Além disso, as regiões teloméricas e subteloméricas foram avaliadas, tendo sido obtidas 24 sequências com telômeros. A montagem do genoma da cepa CL Brener de *T. cruzi* utilizando a combinação de diferentes métodos de sequenciamento apresentou bons resultados quando comparado com a atual montagem de CL Brener do TritrypDB e outras montagens mais recentes de outras linhagens que foram também montadas utilizando reads longas. Nós também avaliamos a ocorrência de pontos de

recombinação no genoma híbrido de CL Brener utilizando reads curtas de Illumina de cepas representantes das linhagens parentais (Y TcII e 231 TcIII). Foram identificados possíveis pontos de recombinação exclusivos de CL Brener, bem como sítios de recombinação compartilhados entre CL Brener e TCC, uma outra cepa híbrida também da DTU TcVI. Por fim, foi desenvolvida a ferramenta CADIn, destinada a inferir ploidia genômica e variações de somias cromossômicas com base em dados NGS com um único comando. Para tal, CADIn usa frequências alélicas de SNPs heterozigotos e análises de cobertura de profundidade de reads. CADIn remove regiões cromossômicas com coberturas atípicas, as quais podem comprometer as análises de profundidade de reads, e válida variações de polida estatisticamente. Através desta ferramenta, foram detectadas aneuploidias no genoma de CL Brener, bem em outros genomas de diferentes complexidades como *Leismania sp.* e *Saccharomyces cerevisiae*. Além disso, dados simulados demonstraram a capacidade de CADIn de usar reads com diferentes comprimentos e obtidas por diferentes métodos de sequenciamento.

Palavras-chaves: Trypanosoma cruzi; montagem hibrida; CADIn; aneuploidia.

ABSTRACT

Trypanosoma cruzi, the etiological agent of Chagas disease, is a flagellated unicellular protozoan parasite, whose the first version of its genome was sequenced and published in 2005. The strain selected was CL Brener, a hybrid lineage between DTUs TcII and TcIII. This hybrid characteristic, in addition to the large number of members of multigenic families associated with other repetitive elements of the *T. cruzi* genome, compromised the quality of assembly. Because of this, in this work, a new assembly was performed using sequencing of long reads (PacBio) combined with short reads (Illumina), and Sanger reads of BACs and Fosmids generated by the genome project of 2005 were also used to aid assembly. For this, different assemblers were tested, such as Canu and HGAP, for the construction of contigs. Scaffolding was performed interactively, reducing the number of reads used to join contigs at each iteration and therefore allowing the assembly of regions with better support first. A total of 446 sequences were obtained at the end of assembly, followed by their correction using short reads. A *de novo* annotation of this new assembly was performed using the Augustus program based on the CL Brener annotation available in the TritrypDB, as well as that of other strains already annotated. In addition, the telomeric and subtelomeric regions were evaluated, obtaining 24 sequences with telomeres. Compared to the public CL Brener assembly and other recent assemblies of different strains that also used long reads, this new genome assembly of the CL Brener showed good results. We have also evaluated the occurrence of recombination in the CL Brener genome using short Illumina reads from strains representative of the parental lineages (Y TcII and 231 TcIII). We detected possible recombination sites exclusive of CL Brener, as well as common recombination

sites between CL Brener and TCC, another hybrid strain of DTU TcVI. Finally, we have developed CADIn, a tool intended to infer genomic ploidy and chromosomal somy variations based on NGS data with a single command. To this end, CADIn uses both allele frequencies of heterozygous SNPs and depth coverage analysis of reads. CADIn removes chromosomal regions with atypical coverage that may complicate read depth analysis and statistically validates ploidy variations. Through this tool, aneuploidies were detected in the CL Brener genome as well as in other genomes with distinct levels of complexity such as *Leismania* sp., and *Saccharomyces cerevisiae*. In addition, simulated data demonstrated CADIn's ability to use reads with different lengths and obtained by different sequencing methods.

Keywords: Trypanosoma cruzi; Hybrid assembly; CADIn; aneuploidy.

Lista de abreviaturas e siglas

BLAST	Ferramenta de Busca Local Básica de Alinhamento (Basic Local Alignment Search Tool)
BLASTN	Alinhamento de Nucleotídeos BLAST
BLASTP	Alinhamento de Proteínas BLAST
DNA	Ácido Desoxirribonucleico
Illumina	Sequenciamento por Illumina
INDEL	Inserção/Deleção (Insertion/Deletion)
MASP	Proteína de superfície associada à mucina
NCBI	National Center for Biotechnology Information
NGS	Sequenciamento de Nova Geração (Next-Generation Sequencing)
OLC	Consenso de Sobreposição de Layout (Overlap Layout Consensus)
RNA	Ácido Ribonucleico
Sanger	Sequenciamento de Sanger
SNP	Polimorfismo de Nucleotídeo Único (Single Nucleotide Polymorphism)
TS	Transialidase
WGS	Sequenciamento de genoma completo

SUMÁRIO

INTRODUÇÃO

Trypanosoma cruzi é um protozoário unicelular flagelado da família Trypanosomatidae, a qual pertence outros parasitos de importância em saúde pública como *Leishmania spp.* e *Trypanosoma brucei.* O protozoário é o agente etiológico da doença de Chagas, uma infecção de alta morbidade que aflige 6-7 milhões de pessoas nas Américas, sendo considerada uma doença tropical negligenciada e endêmica na América Latina (LIDANI *et al.*, 2019; WORLD HEALTH ORGANIZATION, 2022).

T. cruzi é um táxon extremamente polimórfico, sendo atualmente dividido em seis subgrupos ou DTUs (*discrete typing units*) (ZINGALES *et al.*, 2012). Um subgrupo adicional isolado de morcegos e denominado TcBat foi recentemente identificado (SZPEITER *et al.*, 2017). Os DTUs apresentam diferentes características fenotípicas, tais como: virulência, tropismo tecidual, resistência a drogas e adaptação a diferentes espécies de vetores (ZINGALES, 2018). Uma característica dos subgrupos TcV e TcVI é o fato de ambos serem compostos por cepas híbridas entre representantes dos grupos parentais TcII e TcIII. A cepa CL Brener pertencente ao TcVI, e foi o primeiro genoma de *T. cruzi* a ser sequenciado (EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005).

O sequenciamento de *T. cruzi* CL Brener permitiu a identificação de mais de 10 mil genes em cada haplótipo, Esmeraldo-like (derivado do ancestral TcII) e non-Esmeraldo-like (derivado do ancestral TcIII). Grande parte destes genes corresponde a regiões repetitivas que perfazem cerca de 50% do genoma sequenciado, sendo sua maioria elementos transponíveis, sequências satélites, e genes codificadores de proteínas de superfície, como as famílias GP63, MASP, Mucina, e Trans-sialidase, além de DGF-1 e RHS (EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005). Esses

genes de famílias multigênicas não se organizam em clusters gênicos codificados pela mesma fita de DNA como se observa nas regiões contendo genes que codificam proteínas estruturais e do metabolismo basal. Ao invés disso, há uma disposição não ordenada com uma constantemente mudança da fita codificadora e alternância entre genes de diferentes famílias (BARTHOLOMEU *et al.*, 2009; WEATHERLY; BOEHLKE; TARLETON, 2009). Acredita-se que essa organização gênica com alternância de genes de diferentes famílias evite a homogeneização das sequências através de conversão gênica, que pode acontecer quando genes de uma mesma família assumem a organização em tandem (BARTHOLOMEU *et al.*, 2009). A natureza híbrida, aneuploidias e o grande conteúdo repetitivo do genoma de CL Brener dificultaram a montagem do genoma em cromossomos completos, levando a uma montagem inicial que continha aproximadamente 5.000 scaffolds/contigs (EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005). Posteriormente, com o auxílio de BAC-end sequences e mapas de sintenia com *Trypanosoma brucei (BERRIMAN et al., 2005)* e *Leishmania major (IVENS et al., 2005)*, os contigs do clone CL Brener foram montados em 41 cromossomos hipotéticos. Entretanto, 1/3 do genoma de CL Brener ainda não foi anotado e cerca de metade dos membros das famílias multigênicas que codificam proteínas de superfície ainda não foi incorporada na montagem, limitando estudos voltados para aspectos da interação parasito-hospedeiro (WEATHERLY; BOEHLKE; TARLETON, 2009).

Essas dificuldades de montagem são inerentes à estratégia de sequenciamento de *whole genome shotgun,* usada no projeto genoma de *T. cruzi* e método de escolha dos projetos genomas desde o início dos anos 2000. Apesar dos avanços das técnicas de sequenciamento de long reads, ainda não existe uma metodologia que permita a

obtenção da sequência completa de mega-cromossomos em uma única rodada de sequenciamento. Desta forma, para obtenção de genomas completos pela técnica de shotgun, o DNA genômico é fragmentado e depois todos os fragmentos de sequências gerados são comparados entre si para reconstrução dos genomas completos. Este processo é muito mais complexo em genomas altamente repetitivos como o de *T. cruzi*, e ainda mais para cepas híbridas como a CL Brener, composto por dois genomas parentais divergentes. Outro fator complicador é a comum ocorrência de cromossomos supranumerários no táxon (REIS-CUNHA *et al.*, 2018). Todos estes fatores inerentes às características do genoma de CL Brener, associados ao uso de reads de Sanger com baixa cobertura para os dias atuais (~14x *genome coverage*) e à ausência de uso de qualquer mapa físico ou etapas de *closure* durante o projeto, justificam a fragmentação do genoma de CL Brener atualmente disponível. Outras cepas de *T. cruzi* tiveram seus genomas sequenciados por metodologias de short-reads, como G, Sylvio, Y e 231 (BAPTISTA *et al.*, 2018; BRADWELL *et al.*, 2018; CALLEJAS-HERNÁNDEZ *et al.*, 2018; FRANZÉN *et al.*, 2011). Apesar de serem cepas não híbridas, todos estes trabalhos contribuíram para o entendimento da complexidade da espécie e diversidade entre as diferentes cepas de *T. cruzi*. Entretanto, devido ao fato de colapsar regiões altamente repetitivas do genoma, onde em sua maioria são maiores que o tamanho das reads sequenciadas, fez-se necessário o uso de long-reads (AMARASINGHE *et al.*, 2020; PEARMAN; FREED; SILANDER, 2020; POLLARD *et al.*, 2018).

O uso de metodologias de sequenciamento e geração de reads longas permite cobrir uma grande região com uma única leitura, o que pode incluir regiões repetitivas flanqueadas por regiões únicas do genoma, permitindo a localização correta da read sem ambiguidade (AMARASINGHE *et al.*, 2020; PEARMAN; FREED; SILANDER,

2020). No sequenciamento por reads longas, as regiões no genoma com variado teor de GC (guanina e citosina) são menos impactadas pela esta metodologia (RHOADS; AU, 2015), facilitando o sequenciamento de regiões ricas em AT (Adenina e Timina), como as teloméricas (THAM *et al.*, 2023). Para montagem de genomas como os *T. cruzi*, o uso de long-reads se fez necessário, uma vez que desse modo é possível uma maior contiguidade do genoma montado. Algumas cepas de *T. cruzi* já foram montadas com essa abordagem, tais como: Berenice, BrazilA4, Dm28c, TCC e YC6 (BERNÁ *et al.*, 2018a; DÍAZ-VIRAQUÉ *et al.*, 2019; WANG, Wei *et al.*, 2021a). Dentre estes, alguns fizeram uma montagem com metodologias híbridas, que consiste na utilização da metodologia por short e long-reads. Essa montagem mista pode ser utilizada para auxiliar a ligação de contigs gerados pela long-reads através do suporte de cobertura de regiões compartilhadas entre eles, como utilizado para a montagem de Dm28c e TCC (BERNÁ *et al.*, 2018a). Além disso, as short-reads podem ser usadas para a correção de erro nas bases das long reads, que tendem a ter uma maior taxa de erro (AMARASINGHE *et al.*, 2020; PEARMAN; FREED; SILANDER, 2020; ROBERTS; CARNEIRO; SCHATZ, 2013; WATSON; WARR, 2019), como utilizado nas montagens de BrazilA4 e YC6 (WANG, Wei *et al.*, 2021a).

Como mencionado anteriormente, a reconstrução da sequência completa de um genoma se fundamenta na ordenação e organização das reads de modo a agrupá-las com base na similaridade parcial, onde ocorre a sobreposição de suas extremidades. Quando a montagem é realizada por reads longas, as ferramentas usam para tal processo um algoritmo OLC (*overlap layout consensus*), mas adaptado para a maior taxa de erro presente nesse tipo de sequenciamento (GONZALEZ-GARCIA *et al.*, 2023; LI, Zhenyu *et al.*, 2012). Ferramentas como o CANU (KOREN *et al.*, 2017), FALCON

(CHIN *et al.*, 2016) e o FLYE (KOLMOGOROV *et al.*, 2019) utilizam-se desse conceito. O método avalia a sobreposição (overlap) entre as reads da amostra, agrupa todos os dados (layout) e, por fim, determina a melhor sequência consenso (consensus) (MILLER; KOREN; SUTTON, 2010), resultando na obtenção das sequências dos contigs.

Os contigs gerados representam o primeiro estágio na montagem de um genoma. Devido a variações existentes nas coberturas, repetições no genoma e as variações decorrentes de erros do próprio sequenciamento, as sequências geradas são apenas uma representação da montagem (HUSON; REINERT; MYERS, 2002). Um segundo estágio para a construção de um genoma é a etapa de scaffolding. Essa etapa consiste na ligação de contigs utilizando as reads, normalmente paired-ends, e pela cobertura dessa ligação determinar as melhores combinações. Uma etapa anterior ao scaffolding de filtragem de contigs menores e correção de gaps pode ser realizada para otimizar o processo (HUNT *et al.*, 2014; HUSON; REINERT; MYERS, 2002). Montadores como Opera (GAO; SUNG; NAGARAJAN, 2011), SOPRA (DAYARIAN; MICHAEL; SENGUPTA, 2010) e SSPACE (BOETZER *et al.*, 2011) são exemplos de ferramentas que fazem uso de reads paired-ends para ligação e determinação da distância entre os contigs. Com as long-reads, a etapa de scaffolds permitiu sequências ainda maiores que apenas os contigs gerados pelo uso de short-reads (BROWN *et al.*, 2021; MILLER *et al.*, 2017). Algumas ferramentas também foram desenvolvidas para para utilizar as reads longas para o scaffolding, dentre esses tem-se o SAMBA (ZIMIN; SALZBERG, 2022), SLR (LUO *et al.*, 2019) e o LRScaf (QIN *et al.*, 2019), o que permite a ligação de regiões mais distantes (GHURYE *et al.*, 2017). O processo de scaffolding gera gaps de tamanhos estimados entre os contigs. Algumas ferramentas foram

desenvolvidas para resolução desses gaps, com GapFiller (NADALIN; VEZZI; POLICRITI, 2012) e o IMAGE (TSAI; OTTO; BERRIMAN, 2010), que permitem essa correção utilizando reads curtas paired-end.

Após a obtenção da sequência do genoma, é necessário avaliar a qualidade da montagem podendo para isso usar diversas métricas. N50 é uma das mais usadas para determinar a qualidade de uma montagem com base nas sequências geradas, significando o menor comprimento de uma dada sequência, de modo decrescente, que corresponde à metade (50%) da soma de todas a sequências montadas. L50, outra métrica comumente utilizada, se refere ao número de sequências que possuem valores de comprimento maiores ao N50 (VAN DIJK *et al.*, 2014). auN é outra métrica recentemente proposta, que tem como objetivo avaliar a área gerada sob a curva quando se avalia linearmente o tamanho dos contigs, como uma medida de contiguidade (LI, Heng, 2020).

Além dessas formas de avaliar a montagem de um genoma, utilizando as próprias sequências como métricas, há também métodos que se baseiam nas sequências anotadas. Essas sequências são cruzadas quanto um banco construído baseando em sequência de cópia única ortólogas entre cepas de uma mesma espécie ou até mesmo espécies ou gêneros em todo um clado. O BUSCO (Benchmarking Universal SingleCopy Orthologs) (SIMÃO *et al.*, 2015), OrthoFinder (EMMS; KELLY, 2015) e o OrthoMCL (LI, Li; STOECKERT; ROOS, 2003) são ferramentas utilizadas para esse tipo de avaliação. O BUSCO entre elas apresenta um vasto banco já pré construído além da integração com outros softwares que facilitam a análise e detecção de genes, podendo avaliar além de completude a redundância da montagem (MANNI *et al.*, 2021).

Variação do número de cópias cromossômicas

A variação no número de cópias cromossômicas representa um mecanismo comum na manutenção da vida de diversos organismos (COMAI, 2005; FREEMAN *et al.*, 2006). Esse mecanismo pode ocorrer em todo o conjunto de cromossomos de um organismo, denominado poliploidia, ou em um ou alguns cromossomos, gerando as aneuploidias (CHUNDURI; STORCHOVÁ, 2019). Embora a aneuploidia possa ser deletéria na maioria dos organismos pluricelulares, estando presentes em células cancerosas por exemplo (OLTMANN *et al.*, 2018; ORR; GODEK; COMPTON, 2015), em eucariotos unicelulares, pode ser uma estratégia de promover adaptações a mudanças ambientais (COMAI, 2005; IANTORNO *et al.*, 2017; REIS-CUNHA *et al.*, 2018). Alguns organismos multicelulares como as plantas apresentam maior tolerância a existência de aneuploidias, a capacidade de duplicar todo genoma (autoploidia) e a habilidade de geração de híbridos (aloploidia), permitindo a geração de variabilidade e formação de novos fenótipos (KYRIAKIDOU *et al.*, 2018). Por outro lado, o processo de autoploidia, pode estar associado à redução de fertilidade (CROW, 2006).

Em *Leishmania* sp. já foi observado a presença de variações na somia em alguns cromossomos (DUMETZ *et al.*, 2017; MANNAERT *et al.*, 2012; ROGERS *et al.*, 2011; STERKERS *et al.*, 2011). Entretanto, variações no número de cópias cromossômicas parecem ser mais raras em *T. brucei* (ALMEIDA *et al.*, 2018). Para *T. cruzi* alguns trabalhos descrevem a existência de variações no número de cópias cromossômicas entre cepas referências e isolados de campo (MINNING *et al.*, 2011; REIS-CUNHA *et al.*, 2015, 2018). Estudos recentes têm investigado o papel funcional das aneuploidias em *Leishmania*. Pietro Barja e colaboradores postulam que os

isolados de campo de *Leishmania* possuem variações cariotípicas intrínsecas permitindo a seleção de subpopulações com perfis de somias cromossômicas adaptadas a diferentes condições de infecção (PRIETO BARJA *et al.*, 2017). Em *Leishmania*, foi demonstrado para a maioria dos cromossomos uma correlação positiva entre variações no número de cópias cromossômicas e o nível de expressão gênica (DUMETZ *et al.*, 2017). Como a mudança no número de cópias cromossômicas ocorre em uma única rodada de duplicação celular, os mecanismos de aneuploidia tornam-se um mecanismo rápido para alterar a dosagem de diversos genes e, possivelmente, seus níveis de expressão (FEHRMANN *et al.*, 2015; TORRES *et al.*, 2010).

Devido à importância da aneuploidia e poliploidia para eucariotos, várias metodologias de biologia molecular e bioinformática foram desenvolvidas para estimar alterações de número de cópias cromossômicas. Usualmente, os métodos de citometria de fluxo são uma alternativa para estimar os níveis de ploidia, principalmente em estudos biomédicos e de plantas (BLANCO *et al.*, 2013; DOLEZEL; GREILHUBER; SUDA, 2007; HEDLEY *et al.*, 1983; PFAU; AMON, 2015). No entanto, algumas destas estratégias apresentam limitações como a dificuldade de isolamento de material nuclear suficiente sem danos e a baixa detecção por fluorescência (DOLEZEL; GREILHUBER; SUDA, 2007; NATH; MALLICK; JHA, 2014). Recentemente várias ferramentas baseadas em dados de sequenciamento de nova geração (NGS) foram desenvolvidas para estimar ploidia e somia. Essa abordagem tem sido muito utilizada devido à facilidade de sequenciamento, ao baixo custo e disponibilidade de muitas bibliotecas em bancos de dados públicos (VAN DIJK *et al.*, 2014). Ferramentas como ASCAT, AbsCN-seq e CLImAT usam dados NGS para análise e são exemplos de ferramentas destinadas à análise de amostras de câncer (BAO; PU; MESSER, 2014; VAN LOO *et*

al., 2010; YU *et al.*, 2014). O ConPADE foi desenvolvido para estudos de ploidia em plantas, medindo variações através da frequência de alelos heterozigóticos em contigs gerados por sequências de *reads* curta (MARGARIDO; HECKERMAN, 2015). Recentemente, o ploidyNGS foi apresentado como uma ferramenta para estimar variações de ploidia com base em variações alélicas em polimorfismos de nucleotídeo único (SNPs) (SANTOS *et al.*, 2017), e o nQuire foi desenvolvido como um modelo estatístico para distinguir diferentes somia em dados de sequenciamento (WEIS *et al.*, 2018), entretanto de modo semi-automatizada, exigindo que o usuário insira várias linhas de comando durante a análise.

Justificativa

Em estudos genômicos em *T. cruzi*, a cepa CL Brener é utilizada muitas vezes como referência para novas montagens ou anotação de novos genomas de outras cepas (BERNÁ *et al.*, 2018a; CALLEJAS-HERNÁNDEZ *et al.*, 2018; DÍAZ-VIRAQUÉ *et al.*, 2019; WANG, Wei *et al.*, 2021a). Por outro lado, os novos métodos propiciaram melhorar o sequenciamento de organismos com genomas altamente repetitivos (VAN DIJK *et al.*, 2014), como o *T. cruzi*. Apesar das novas montagens para diferentes cepas de *T. cruzi*, o genoma da cepa CL Brener ainda não foi montado com reads longas, de terceira geração. Isso faz com que ainda exista grande parte de genes em contigs pequenos não incorporados aos cromossomos montados, o que dificulta análises a nível cromossômico (REIS-CUNHA; BARTHOLOMEU, 2019). As montagens mais recentes ainda apresentam limitações, principalmente no que tange às famílias multigênicas, que são na sua maioria proteínas que desempenham funções cruciais na sobrevivência do parasito como aquelas presentes nas superfícies relacionadas com adesão/invasão de células e evasão do sistema imune no hospedeiro (BARTHOLOMEU *et al.*, 2009; BUSCAGLIA *et al.*, 2006; DE PABLOS; OSUNA, 2012). Uma melhor montagem pode permitir o estudo de distâncias gênicas e sintenia, além de poder auxiliar na detecção de possíveis alvos para desenvolvimento de vacinas e diagnóstico, como membros da família das trans-sialidase (BONTEMPI *et al.*, 2017; REIS-CUNHA *et al.*, 2022; WANG, Wei *et al.*, 2021a). A montagem de um genoma complexo como o da cepa CL Brener pode auxiliar no desenvolvimento de estratégias também aplicáveis a outros genomas que apresentam características similares, como: genoma altamente repetitivo, de natureza híbrida e com a presença de cromossomos supranumerários.

Além disso, uma nova montagem também possibilita a avaliação do processo de hibridização, visto que o genoma de linhagens de DTUs parentais já estão disponíveis (BAPTISTA *et al.*, 2018; CALLEJAS-HERNÁNDEZ *et al.*, 2018). Esta avaliação pode auxiliar no desenvolvimento de modelos de recombinação, identificando *hotspots* de recombinação que podem ser importantes para a compreensão do processo evolutivo do parasito.

Frente ao exposto, nosso trabalho teve como objetivo montar *de novo* e anotar o genoma de *T. cruzi* cepa CL Brener utilizando de reads longas (PacBio) associado a reads curtas (Illumina). Além disso, propomos reavaliar a estrutura genômica e analisar o número de cópias cromossômicas através de um programa desenvolvido para tal, o CADIn. Essa ferramenta é gratuita, simples e automatizada que estima aneuploidías com base na análise de frequência alélica e variações de profundidade de reads (read depth - RD). Cadin pode ser usados para estudos de variação de ploidia em organismos unicelulares e multicelulares, cujos genomas foram sequenciados por WGS.

Objetivo

Objetivo Geral

Avaliar a estrutura do genoma da cepa CL Brener de *Trypanosoma cruzi* a partir da montagem *de novo* do genoma usando reads longas de SMRT, curtas de Illumina e reads de BACs e Fosmídeos de SANGER, visando identificação de *hotspots* de recombinação e avaliação de somias cromossômicas.

Objetivos específicos

- Gerar uma montagem do genoma *T. cruzi* cepa CL Brener utilizando reads longas de SMRT, reads curtas de Illumina, sequências de Sanger de BAC e fosmídeos por diferentes abordagens;
- Comparar a montagem gerada com outras já existentes;
- Anotar genes codificadores de proteínas e RNAs estruturais, além das regiões repetitivas no genoma, tais quais: regiões teloméricas e satélite;
- Identificar pontos de recombinação entre as linhagens parentais no genoma híbrido de CL Brener;
- Desenvolver e validar um método automatizado para avaliação de ploidias por profundidade de reads (RD) e frequência alélica;

CAPÍTULO 1

NOVA MONTAGEM PARA O GENOMA DE *TRYPANOSOMA CRUZI* CEPA CL BRENER UTILIZANDO DE READS LONGAS DE SMRT ASSOCIADAS A READS CURTAS DE ILLUMINA E DADOS DE SANGER

Metodologia

Sequenciamento e genotipagem

Epimastigotas de CL Brener foram genotipados (BURGOS *et al.*, 2007; DE FREITAS, Jorge M. *et al.*, 2006) e em seguida $1x10^8$ parasitos foram usados para extração de DNA. O material genético foi sequenciado em *reads* curtas pela Macrogen (Seul, Coreia do Sul) utilizando um sequenciador Illumina Hiseq2000, gerando reads pareadas de 100pb com intervalo de 350pb. Já o sequenciamento em reads longas SMRT foi realizado no Broad Institute (Boston, Massachusetts, Estados Unidos).

Montagem *de novo*

O processo de montagem do genoma foi dividido em duas etapas: obtenção dos contigs e *scaffolding*. A primeira etapa consistiu em montar contigs utilizando apenas *long-reads* e os softwares CANU (KOREN *et al.*, 2017) ou HGAP (*Hierarchical Genome Assembly Process*) versão 3 (CHIN *et al.*, 2013), com intuito de avaliar qual apresentaria o melhor resultado quanto às métricas de montagem. Ambos os programas foram rodados nas configurações padrão de cada software, variando apenas o comprimento esperado do genoma a ser montado. As métricas foram analisadas com Quast (GUREVICH *et al.*, 2013) e pelo GAEMR (*Genome Assembly Evaluation Metrics and Reporting*) (BROAD INSTITUTE, [*s. d.*]) para avaliação da melhor montagem.

Após a montagem, os contigs passaram por correção com Pilon (WALKER *et al.*, 2014), usando as *reads* curtas e pareadas de Illumina previamente trimadas utilizando o Trimmomatic (BOLGER; LOHSE; USADEL, 2014), eliminando as leituras com uma

qualidade inferior a 30 (Q30) e um comprimento inferior a 50bp, bem como as sequências dos adaptadores. O objetivo desta etapa foi corrigir possíveis erros de sequenciamento das reads longas que vieram a ser incorporadas nos contigs montados. No fim da etapa inicial os contigs foram avaliados pela ferramenta IPA (OTTO, 2017) para remover pequenos contigs, redundância de montagem ou regiões sobrepostas.

Na etapa seguinte, realizou-se o processo de *scaffolding* pela combinação e ordenamento sequencial de contigs filtrados obtidos na etapa anterior. Essa etapa foi subdividida em 4 passos: (1) scaffolding utilizando SMRT long-reads; (2) uso de BACs e fosmídeos; (3) ligação e extensão com short-reads da Illumina; e (4) o fechamento dos gaps. Todas essas etapas foram realizadas de modo iterativo respeitando a classificação de sequências ambíguas no primeiro passo, finalizando assim que nenhuma sequência fosse classificada como ambígua.

O SLR (LUO *et al.*, 2019) foi selecionado por ser uma ferramenta de scaffolding. SLR ordena os contigs atrás de reads longas e classifica as sequências geradas ao fim da análise em únicas ou ambíguas baseadas em informações de alinhamento de *reads* longas aos contigs. Além disso, o SLR usa apenas contigs únicos nas etapas iniciais de scaffolding e, posteriormente, incorpora contigs ambíguos gerando as scaffolds finais, o que auxilia na montagem de genomas repetitivos.

Na etapa seguinte, usou-se o SSpace (BOETZER *et al.*, 2011) com *reads* de fosmídeos sequenciadas pelo método de Sanger (espaçamento de *paired-end* 34 kb) e BACs (espaçamento de *paired-end* de 90 kb) (EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005). Nessa fase, os valores mínimos do suporte de reads necessárias para conectar duas regiões foram reduzidos devido a baixa cobertura desses tipo de dado,

mas a diferença entres possíveis sequências ligantes, nos casos da sequência “A” poder se ligar a “B” ou “C”, foi aumentada. Isso permitiu que, mesmo com menor suporte de reads na região, quando duas sequências foram combinadas, aquela a ser conectada tenda a ser única.

O SSpace também foi utilizado para geração dos scaffolds. Por usarmos reads curtas, o suporte necessário para ligar as sequências foi aumentado, ao contrário do anterior, mas mantendo a diferença da cobertura quando houvesse mais de um ligante. Ou seja, quando a região “A” tiver 100 reads pareadas e dessas 60 ligam-se a “B” e as demais 40 em “C” a razão C/B é de 0.77, no nosso caso para confirmar que A se liga a B o valor da razão deve ser inferior a 0.8 e o valor mínimo de read pareadas ligando A e B de 100. Adicionalmente foi realizada a extensão das sequências montadas. Assim, foram usados *reads* curtas para combinar (*paired-end*) e estender (*single-ends*) os contigs/scaffolds gerados. Em cada iteração o suporte foi reduzido em 10, sendo iniciado em 100 paired-end ligantes até chegar em 10 paired-ends. Após cada iteração, foi feita a correção dos *gaps* usando o GapFiller (NADALIN; VEZZI; POLICRITI, 2012) usando as reads de Illumina paired-end para reduzir o número de N’s nas sequências. Por fim, após a etapa de scaffolding, o Pilon foi novamente utilizado para correção de algum possível erro gerado, utilizando as reads curtas paired-end.

Anotação

Para detecção de possíveis regiões gênicas, a anotação foi particionada para os diferentes classes de genes, como os codificadores de proteínas (CDS), os RNAs (incluindo os não codificantes, transportadores e ribossomais), os elementos transponíveis, regiões de satélite (SAT) e telômeros.

A anotação das regiões codificantes foi realizada usando o Augustus (KELLER *et al.*, 2011), sendo selecionado por permitir a anotação *de novo* de sequências codificadoras de proteínas. O treinamento foi realizado com três diferentes conjuntos de dados de anotação, sendo um construído com apenas informações de proteínas da cepa CL Brener, outro utilizando dados de cepas de *T. cruzi*, dentre elas: BrazilA4, Dm28c, G, marinkeleiB7, Sylvio, TCC e YC6, denominado *T. cruzi*, e terceiro banco com dados de *T. brucei*. Todas as informações sobre as cepas foram recuperadas do banco de dados TritrypDB na versão 53 (ASLETT *et al.*, 2010). Todas as sequências utilizadas no treinamento foram filtradas quanto à presença da metionina inicial, ausência de códon de parada interno, presença do mesmo no fim da sequência, e comprimento superior a 30 nucleotídeos. Essas etapas foram realizadas com o objetivo de remover possíveis sequências truncadas do treinamento. Após o treinamento, a predição foi realizada buscando genes em ambos os sentidos e completos, gerando as sequências preditas para cada conjunto de dados treinados As sequências foram comparadas com o respectivo conjunto de dados pelo BLASTp (ISMAIL, 2022), e foram classificadas em idênticas (valores 100% de identidade e cobertura), similares (>= 80 % de identidade, mas não idêntica) e preditas (as demais), as duas primeiras foram renomeadas com o nome dos genes do conjunto de treinamento. Após esse processo os dados foram comparados entre os conjuntos de CL Brener, *T. cruzi* e *T. brucei*, e nos casos em que houve sobreposição da anotação manteve-se o melhor valor de predição gerado pelo Augustus, e quando uma região anotada possuía nomenclatura de anotação diferente e similar escore, manteve-se os dados da cepa CL Brener, seguido do *T. cruzi* e finalmente *T. brucei*, ficando: CL Brener > *T. cruzi* > *T. brucei*. Os genes preditos em CL

Brener e que não teve correspondência entre os outros conjuntos de dados foram anotados como proteínas preditas.

As famílias multigênicas DGF-1, MASP, Mucina, RHS e Trans-sialidase foram reanalisadas usando informações dos genes membros dessas famílias na anotação CL Brener. Para cada família foi utilizado uma pipeline desenhada para buscar similaridade, através do Blastn (ISMAIL, 2022), feita a correção da fase de leitura, e nos casos em que o códon final não fosse encontrado, a sequência era estendida para a posição de stop códon mais próxima à porção 3' da região de similaridade. Posteriormente, a região foi traduzida e comparada com a sequência da proteína com o Blastp. As sequências com identidade e cobertura superior a 85% foram as aprovadas no teste de similaridade, e aquelas com valores abaixo desse foram descartadas. Para as trans-sialidases, os motivos SAPA, VTV, FRIP foram também avaliados. As sequências detectadas pelo Augustus e pelo pipeline foram comparadas entre si. Os genes anotados como proteína predita (sequências *de novo* para os dados treinados por CL Brener, mas sem similaridade quando testadas) ou hipotética foram reanotados como membros da respectiva família multigênica, e nos casos em que não houve predição, o gene foi adicionado aos dados de anotação.

Para as regiões não codificadoras de proteínas o Blast foi usado para avaliar similaridade entre as regiões gênicas na montagem de 2005 e a montagem realizada. Foram usadas informações recuperas do TritrypDB (versão 53), realizadas a fim de pesquisar tRNA, rRNA, slRNA, snoRNA, snRNA, sRNA e elementos transponíveis utilizando filtro de 90% de identidade e cobertura. Já as de SAT (satélites) o filtro foi de 70% de identidade. Para estudar a parte telomérica, buscamos a repetição TTAGGG já descrita para essa região (KIM *et al.*, 2005), usando o blastn-short, filtro de tamanho de

20 nucleotídeos em tandem. O Aragorn (LASLETT; CANBACK, 2004) foi utilizado para anotação de novo de tRNA diferentes do já detectado e de tmRNA se caso houve.

Comparação da montagem

A qualidade da montagem final foi comparada àquela das cepas Brazil, Bug2148, Dm28, TCC e YC6, todas sequências oriundas de *reads* longas por Pacbio (BERNÁ *et al.*, 2018a; CALLEJAS-HERNÁNDEZ *et al.*, 2018; WANG, Wei *et al.*, 2021a). Além desses, foi adicionado a cepa Berenice, uma montagem realizada por Nanopore (DÍAZ-VIRAQUÉ *et al.*, 2019). Todas as sequências dos genomas foram recuperadas do TrytripDB e avaliadas com o software Quast para observar as métricas de montagens. Outra análise também realizada foi da completude através do BUSCO (SIMÃO *et al.*, 2015) versão 5.4.6 utilizando o metaeuk como preditor gênico e no modo euk_genome_met para avaliação contra o banco de genes ortólogos para eucariotos. Os *scaffolds* finais também foram comparados quanto à similaridade contra a montagem dos cromossomos da cepa CL Brener (WEATHERLY; BOEHLKE; TARLETON, 2009) e contra si mesmo e contra a cepa TCC, utilizando o programa Nucmer (MARÇAIS *et al.*, 2018).

Estrutura do genoma montado

Os pacotes chromplot (ORÓSTICA; VERDUGO, 2016) e tidyverse (WICKHAM, H.; WICKHAM, [*s. d.*]) do R (R CORE TEAM, 2022) foram usados para construir mapas das scaffolds contendo genes das famílias multigênicas, elementos transponíveis e regiões satélites e regiões teloméricas.

O conteúdo GC (Guanina e Citosina) foi calculado para detectar possíveis bias ao longo do genoma anotado. Para isso os scaffolds foram subdivididos em subregiões de 10kb usando a função “makewindows” do bedtools (QUINLAN; HALL, 2010) e então o conteúdo GC dessas subregiões foi calculado com a função “nuc”. O tamanho de 10kb foi selecionado de modo tentar contemplar mais de um gene na mesma região, uma vez que o objetivo não seria avaliar o GC de genes específicos e sim da região anotada.

Para avaliação do número de cópias cromossômicas na montagem de CL Brener foi utilizado o CADIn, ferramenta descrita no capítulo 2 desta tese.

Busca de regiões de recombinação

Reads curtas das cepas Y e 231, representantes dos grupos parentais, DTUs TcII e TcIII, respectivamente, foram mapeadas no genoma montado de CL Brener (DTU TcVI), a fim de buscar possíveis sítios de recombinação que possam ter ocorrido na formação do genoma híbrido. Para tal, as bibliotecas de reads foram mapeadas com o BWA (LI, Heng; DURBIN, 2009) utilizando a função mem, e o mapeamento processado com o SAMTools (LI, Heng *et al.*, 2009), função depth com a flag -a para computar posições de 0 reads, para a avaliação da profundidade de cobertura para cada uma das linhagens dos DTUs parentais ao longo do genoma de CL Brener. O programa R (R CORE TEAM, 2022) foi utilizado para cálculos e visualização dos dados com o pacote ggplot2 (WICKHAM, Hadley, 2011). Por fim, as *reads* longas de PacBio foram alinhadas contra a montagem final para observar se as possíveis regiões de recombinação eram verdadeiras ou apenas erros de montagem, uma vez que o mapeamento de *reads*

longas na região confirma a ocorrência da recombinação. O alinhamento das reads de PacBio foi realizado utilizando o Minimap2 (LI, Heng, 2018), e o IGV (Integrative Genomics Viewer) (ROBINSON; ZEMO JTEL, 2017) para a visualização dos dados. O alinhamento dos parentais também foi realizado no genoma de TCC, cepa híbrida e do mesmo DTU de CL Brener, para a avaliação desses mesmos possíveis pontos de recombinação. Para isso foi utilizado os dados gerados pelo teste de similaridade realizado pelo Nucmer entre CL Brener e TCC (descrito no tópico anterior). As regiões de TCC descritas (BERNÁ *et al.*, 2018b) tiveram a sua cobertura avaliada e a região com maior similaridade em CL Brener também foi analisada quanto à variação da cobertura dos parentais contra seu genoma.

Resultados

Montagem híbrida do genoma de CL Brener

Um total de 2.366.431 *reads* longas *single-end* foram gerados com tamanho variando entre 100 e 110.848 bases para as reads brutas. Essas *reads* foram usadas no início da montagem dos *contigs*. Todas as etapas da montagem foram avaliadas e as métricas podem ser vistas na Tabela 1-1. As *reads* longas permitiram a montagem de 2.697 *contigs* totalizando 88,7 Mb utilizando o HGAP como montador, e 1621 contigs para o CANU. Apesar da menor contagem de contigs pelo CANU, ele mostrou menores valores de N50 e tamanho total de contigs quando comparado à montagem do HGAP (Tabela 1-1), e assim não foram usados nas etapas seguintes. Ao mapear as reads longas na montagem foi alcançado um valor de mediana de cobertura de 31. Ao ser avaliada a montagem dos contigs contra o genoma de CL Brener de 2005 é possível observar a presença de grande parte dos cromossomos contemplados (Figura 1-1). Os contigs obtidos com HGAP foram submetidos ao Pilon para correção das bases nas sequências usando reads curtas sequenciadas, e após correção foi avaliado pelo IPA para remoção de redundâncias e pequenos contigs, resultando em 1.883 contigs, os quais foram usados na etapa de scaffolding.

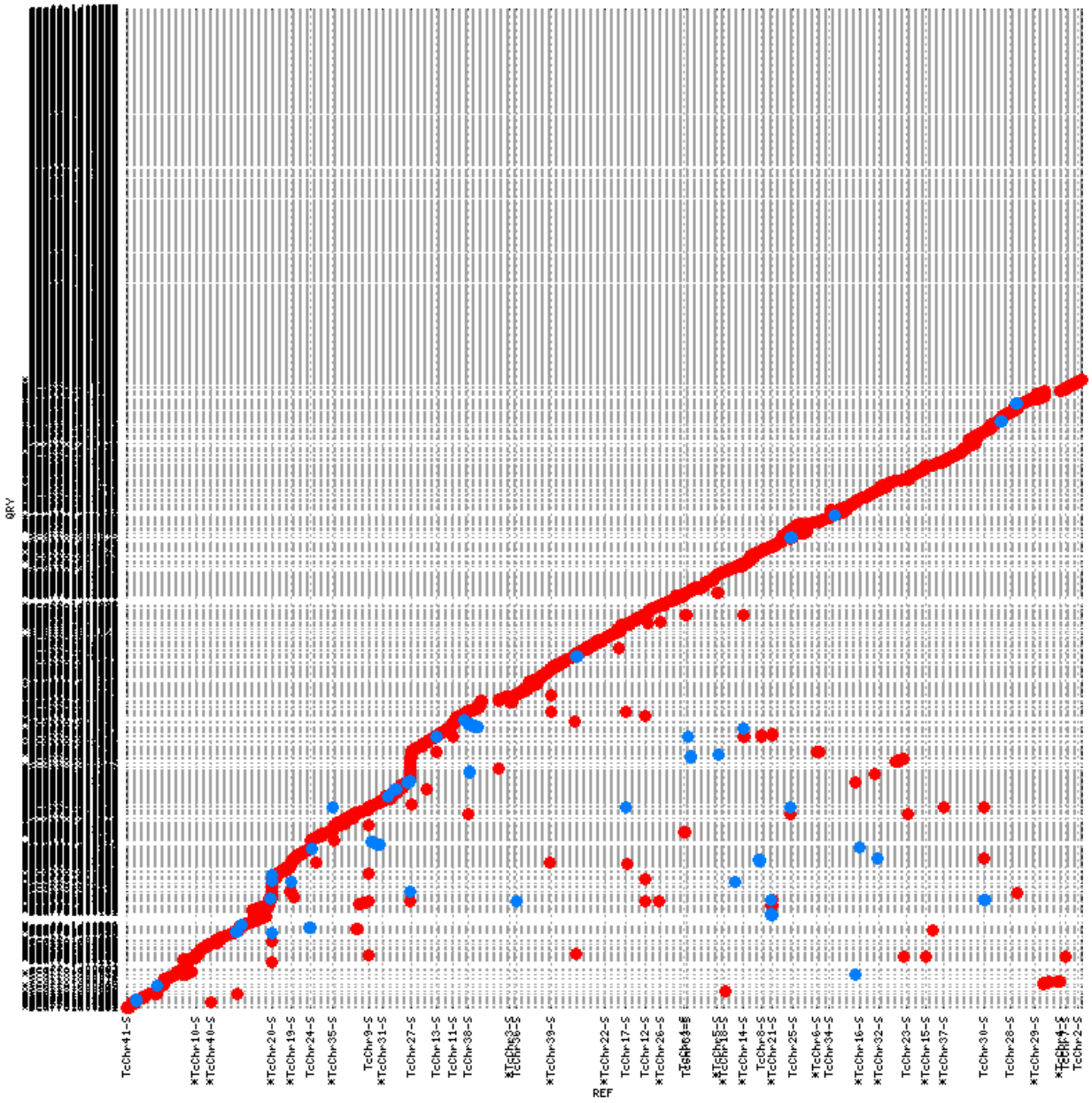


Figura 1-1 Alinhamento entre os contigs montados pelo HGAP pré-correção (eixo y) e as sequências CL Brener de 2005 (eixo x). Cada ponto representa similaridade de sequência, sendo que os em vermelho se alinham na mesma direção e os em azul, na direção complementar. Imagem gerada pelo pacote GAEMR.

O processo de scaffolding foi realizado em 10 iterações, sendo que no início da décima primeira, o SLR não classificou em única ou ambígua as sequências, não realizando assim o scaffolding. Nenhuma das iterações alterou consideravelmente a porcentagem de GC, mas como esperado houve uma redução gradual no número de sequências e aumento considerável no valor do N50, de 377 Mb para 548 Mb. Isso demonstra a melhoria da montagem à medida que as interações aconteceram, o que também é refletido no aumento do valor de auN de 464 Mb para 656 Mb, demonstrando que as sequências apresentam uma melhora na contiguidade. Em relação ao número de N's (*gaps*) foi maior na iteração final quando comparado ao início do *scaffolding* (Tabela 1-1). No entanto, tentamos minimizar o número de N's através das *reads* curtas incorporadas pelo GapFiller. Como pode ser visto na Tabela 1-2, há uma redução no número de N's ao fim do scaffolding de cada iteração, ou seja, após correções de *gaps* com o GapFiller.

Tabela 1-1 Avaliação da qualidade da montagem a partir dos resultados de contiguidade e completude do genoma da cepa CL Brener de *T. cruzi* antes e após cada interação. Após as etapas iniciais de montagem usando HGAP/CANU, o script IPA e 10 iterações de SLR usando *reads* longas de PacBio, SSPACE com *reads* de Sanger, SSPACE com *reads* curtas *paired-end* e *single-end* da Illumina e GapFiller com *reads paired-ends* da Illumina.

Montagem	Contagem sequências	Tamanho total	Maior sequência	GC (%)	N50	N90	auN	L50	L90	N's/100 kpb
CANU	1621	65772805	571796	51,44	48206	19785	92225,8	319	1208	0
HGAP	2697	88773354	1254158	51,71	83077	12236	164376	234	1460	0
Pilon 1	2697	88366197	1237483	51,7	82610	12204	163401,6	234	1462	0
IPA	1883	81109134	1378337	51,55	122436	16637	193858,1	169	927	0
Iteração 1	1118	84334387	1565086	51,54	377819	28682	426438,3	70	356	3889,73
Iteração 2	987	84722880	1696269	51,54	414844	35873	485355,8	63	299	4307,56
Iteração 3	934	84871895	1696415	51,54	430259	38045	497422,1	61	282	4463,87
Iteração 4	886	84933260	1696415	51,54	435468	41717	501626,4	61	270	4518,96
Iteração 5	853	84947706	1696415	51,54	440805	45273	503286,6	61	261	4519,85
Iteração 6	802	84957964	1696415	51,54	442886	47904	508609,8	60	253	4516,7
Iteração 7	716	84973130	1696415	51,54	458411	53190	513836	60	240	4515,33
Iteração 8	604	84980836	1696415	51,54	468125	65932	535892,5	58	221	4512,96
Iteração 9	473	85013075	1831339	51,54	527296	97090	603102,5	52	187	4534,32
Iteração 10	446	85029489	2407301	51,54	548585	107867	656462,7	49	174	4535,85
Final (Pilon 2)	446	84908214	2407335	51,53	548600	107874	656226,4	49	174	4531,04

Tabela 1-2 Impacto do fechamento de *Gaps* na redução dos valores de N`s/100kbs antes e depois da implementação do GapFiller.

Iteração	N`s/100kbs **Antes do GapFiller**	N`s/100kbs **Depois do GapFiller**	N`s/100kbs **Diferença**
Iteração 1	3933,21	3889,73	-43,48
Iteração 2	4320,66	4307,56	-13,1
Iteração 3	4472,48	4463,87	-8,61
Iteração 4	4523,88	4518,96	-4,92
Iteração 5	4524,34	4519,85	-4,49
Iteração 6	4520,91	4516,7	-4,21
Iteração 7	4520,79	4515,33	-5,46
Iteração 8	4515,88	4512,96	-2,92
Iteração 9	4537,4	4534,32	-3,08
Iteração 10	4539,6	4535,85	-3,75

Por fim, o mapeamento de *reads paired-end* de Illumina na montagem mostrou que 96,36% das *reads* foram usadas e estão presentes nos *scaffolds* (Tabela 1-3). Além disso, verificamos que 4,19 milhões de reads pareadas (7,52%) a mais quando comparados aos “cromossomos” do genoma anterior de 2005, e 4,83 milhões se forem consideradas as reads paired-end e single-end. Quando analisamos as reads de alta qualidade de mapeamento (>=30 – MQ30) 100% delas foram mapeadas. Esses dados demonstram que a metodologia de montagem conseguiu aproveitar grande parte daquilo que foi sequenciado, ainda mais se tratando do genoma repetitivo da cepa CL Brener.

Tabela 1-3 A proporção de *reads* mapeada na montagem final da cepa de *T. cruzi* CL Brener e a mesma cepa na montagem anterior (2005).

	CL Brener 2023	**CL Brener 2005**
Filtros		

	Nº total de *reads* mapeadas	Proporção de reads mapeadas	Nº total de *reads* mapeadas	Proporção de reads mapeadas
Paired-end	54.297.344	96,36%	50.111.157	88,84%
Unpaired	8.630.299	96,05%	7.981.190	88,67%
Merge	62.927.643	96,34%	58.092.347	88,82%
MQ 30	34.435.874	100%	35,259,304	100%

Anotação

O programa Augustus gerou 29.649, 29.064 e 33.464 sequências preditas para o software treinado com CL Brener, outras cepas de *T. cruzi* e dados de *T. brucei* respectivamente. Muitas destas sequências não apresentaram similaridade aos genes já anotados. A cepa CL Brener apresentou 20.892 genes com similaridade, sendo 11.173 genes idênticos às proteínas dos dados de treinamento do programa (Tabela 1-4). O treinamento com os dados de *T. brucei* foi o que retornou maior número de sequências preditas, mas como esperado com mais baixa similaridade, e nenhuma idêntica. No final, depois de combinar as proteínas preditas de diferentes conjuntos de dados e checagem dos genes sobrepostos, entre idênticos e similares e entre os grupos treinados, foram contabilizados 29.759 genes preditos pelo Augustus.

Tabela 1-4 O número de genes *ab initio* preditos com Augustus para os diferentes conjuntos de dados de treinamento. As colunas referem-se ao número total, aquelas que apresentam alguma similaridade quando comparadas com BlastP versus o proteoma do respectivo conjunto de dados, o número de proteínas idênticas anotadas e o valor final recuperado pela remoção de sequências encontradas em dois ou mais datasets.

Dataset	Total de genes preditos	Genes idênticos	Genes Similares*	Idênticos + Similares	Anotados**	Final***

CL Brener 2005	29.649	11.173	9.719	20.892	12.436	21.607
T. cruzi	29.064	16.321	15.571	31.892	8.135	8.135
T. brucei	33.464	0	637	637	17	17

**>=80% de identidade e sequência não idêntica, **sem sobreposição, ***para CL Brener são computados genes preditos sem similaridade.*

Membros das famílias multigênicas DGF, MASP, Mucinas, RHS e Trans-sialidases foram analisados quanto à similaridade dos genes na montagem de CL Brener de 2005 e os scaffolds montados. Um total de 1392 sequências de trans-sialidases e 1095 MASPs foram montadas. O pipeline permitiu uma maior detecção de genes em relação à predição pelo Augustus (Tabela 1-5).

Tabela 1-5 Total de genes anotados nos scaffolds montados de CL Brener para cinco diferentes famílias multigênicas de *T. cruzi*.

Genes	**Augustus**	**Total**
DGF	162	265
MASP	948	1.095
Mucina	804	926
RHS	683	795
Trans-sialidase	691	1.392

A predição de sequências não codificadoras de proteínas resultou em 7.391 elementos (Tabela 1-6). Um total de 127 genes de tRNA, distribuídos em 43 dos *scaffolds,* foram identificados com maior presença de tRNA para Arginina, Glicina e Leucina, e menor número de tRNA de Tirosina e Triptofano (Tabela 1-7). Um total de 115 tRNAs foram anotados na montagem de 2005, sinalizando para 12 possíveis novas

cópias de tRNAs, sendo essas preditas *de novo* pela ferramenta Aragorn. Além disso, tRNAs de selenocisteína foram preditos, não sendo observados anteriormente (Tabela 1-7).

Tabela 1-6 Número de regiões anotadas para cada classe e o número de scaffolds em que sua presença foi observada.

Anoltação	SAT	TE*	snoRNA	slRNA	tRNA	rRNA	snRNA	sRNA	Total
Quantidade	2.399	4394	189	112	127	152	16	2	7.391
Contagem de *Scaffolds*	30	308	114	3	43	13	14	2	330

**Elementos transponíveis.*

As outras classes de sequências foram identificadas por similaridade (Tabela 1-6). As regiões satélites (SAT), que correspondem a sequências de 195 pb que se repetem *in tandem*, foi uma das repetições mais abundantes da montagem. Em nossa anotação, observamos 2.399 unidades das sequências SAT, em 30 scaffolds, incluindo sequências completas e truncadas, algumas por elementos transponíveis como SIRE e VIPER. Um total de 4.394 elementos transponíveis foram identificados. Eles foram observados próximos ou truncando genes codificadores de proteínas, além das sequências SAT. Entre as outras classes de RNA, a mais comum foi o snoRNA, com 189 sequências, depois o rRNA (152), o slRNA (112), o snRNA (16) e o sRNA (2). Mais da metade dos *scaffolds* montados (330) contêm sequências correspondentes a RNAs estruturais, repetições SAT e elementos transponíveis (Tabela 1-6).

Tabela 1-7 – Número de tRNAs preditos com Aragorn.

tRNA	CLBrener 2023	CLBrener 2005
Ala	6	5
Arg	12	12
Asn	4	4
Asp	5	2
Cys	3	2
Gln	6	6
Glu	6	5
Gly	10	8
His	4	4
Ile	6	7
Leu	12	12
Lys	6	6
Met	6	6
Phe	4	4
Pro	6	6
SeC	7	0
Ser	8	8
Thr	7	6
Trp	2	2
Tyr	1	2
Val	6	8

Regiões teloméricas e subteloméricas

Foram encontradas 24 repetições teloméricas distribuídas em diferentes *scaffolds* com tamanhos variando de 32 a 4521 nucleotídeos (Tabela 1-8) e, como esperado, apenas nas extremidades de *scaffolds* (Figura 1-S1). Adjacentes às repetições teloméricas (regiões subteloméricas), genes de algumas famílias multigênicas foram observados em comum entre os *scaffolds* (Figura 1-2 A), em um padrão bem distinto quando comparado a porções mais distantes aos telômeros

(Figura 1-2 B). Em regiões subteloméricas, foram encontrados membros das famílias multigênicas RHS e Trans-sialidases, bem como elementos repetitivos como DIRE, L1Tc, NARTc, SIRE e VIPER foram observados com mais frequência (Figura 1-3). Por outro lado, membros das famílias MASP e Mucinas são encontrados nas regiões internas dos cromossomos, em um número menor de scaffolds (Figura 1-2).

Tabela 1-8 *Scaffolds* contendo repetições teloméricas. Informações dos *scaffolds*, posição inicial e final, orientação e tamanho das regiões dos telômeros encontradas na montagem final do genoma de CL Brener *T. cruzi*.

Scaffold	**Coordenada Inícial**	**Coordenada Final**	**Orientação**	**Comprimento**
TcBrS003	1	106	-	106
TcBrS007	1	92	-	92
TcBrS008	1	4521	-	4521
TcBrS028	689240	689325	+	86
TcBrS029	1	52	-	52
TcBrS032	1	1326	-	1326
TcBrS033	641507	642352	+	846
TcBrS044	1	1861	-	1861
TcBrS045	589690	590667	+	978
TcBrS048	549766	549797	+	32
TcBrS052	1	32	-	32
TcBrS055	520017	520048	+	32
TcBrS064	436668	436699	+	32
TcBrS078	1	340	-	340
TcBrS086	351757	351788	+	32
TcBrS087	349186	349217	+	32
TcBrS093	305505	305536	+	32
TcBrS124	1	32	-	32
TcBrS129	190543	190678	+	136
TcBrS172	1	175	-	175

TcBrS173	110880	110945	+	66
TcBrS185	85380	86955	+	1576
TcBrS226	48041	48072	+	32
TcBrS272	30647	32650	+	2004

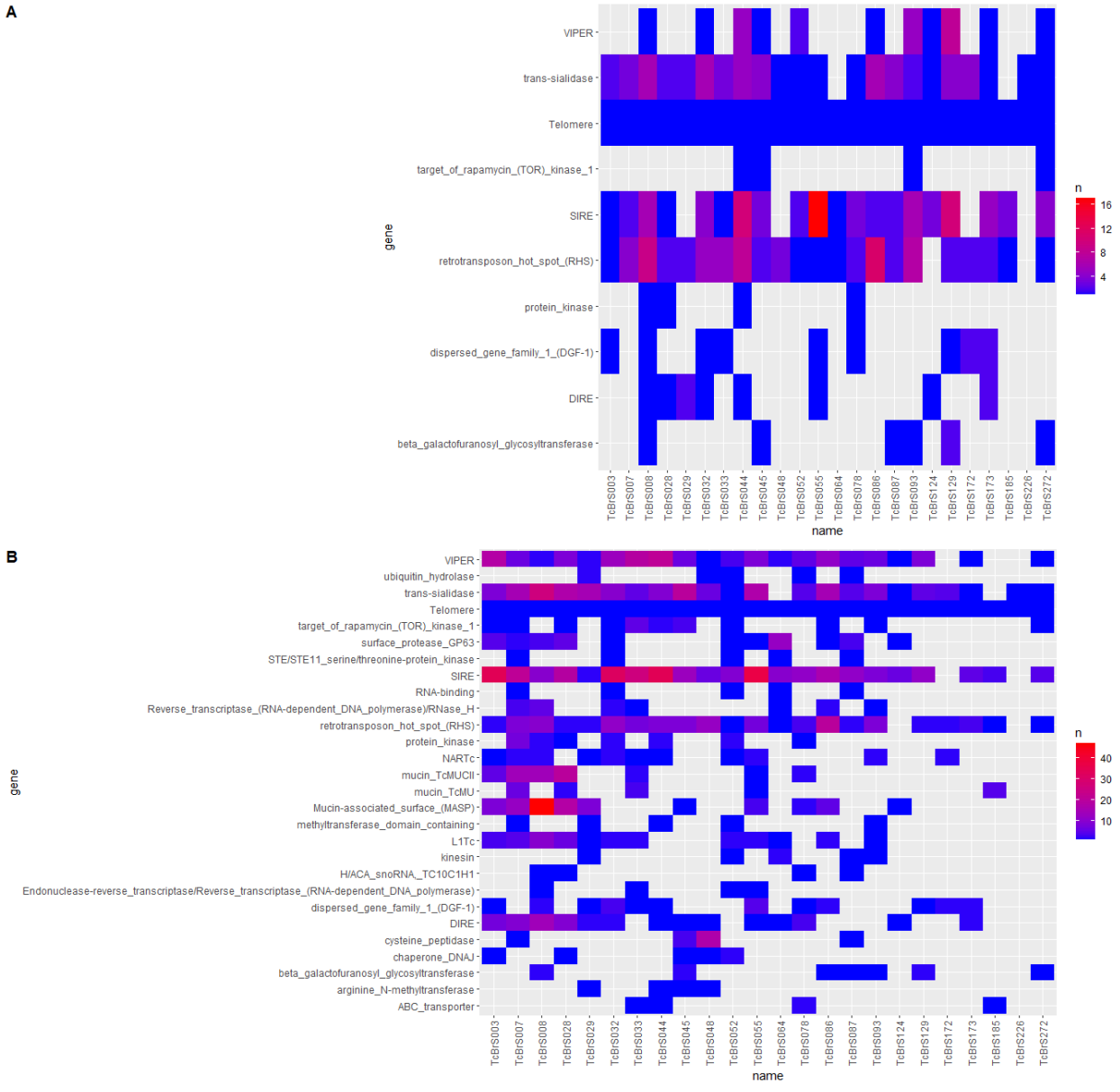


Figura 1-2 Heatmap de genes presentes nas (A) regiões subteloméricas (100kb após o telômero) e (B) 1Mb após o fim do telômero na montagem de CL Brener. O número de cópias

está determinado pela intensidade de calor, sendo azul baixo número de cópias e vermelho, alto número de cópias.

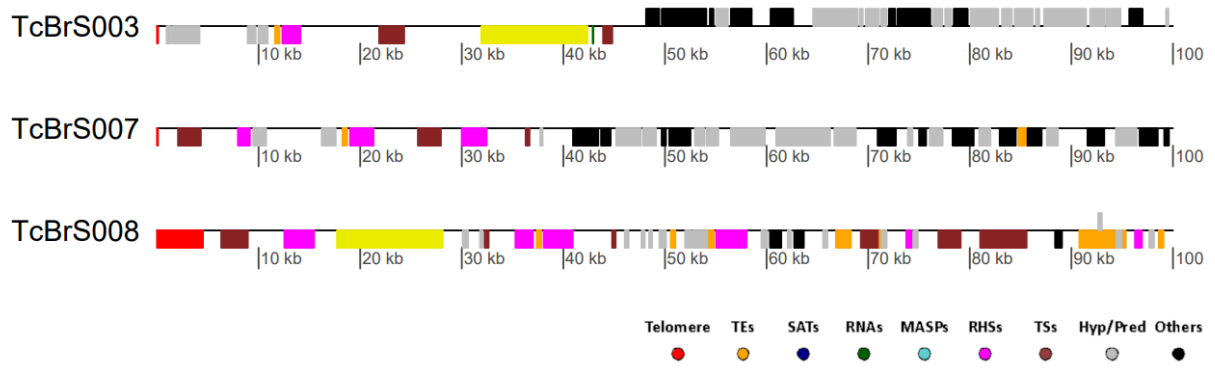


Figura 1-3 Representação gráfica da porção próxima aos telômeros de alguns *scaffolds* na montagem final do *T. cruzi* cepa CL Brener. As caixas representam diferentes genes/elementos anotados no cromossomo, com destaque para as caixas vermelhas nas extremidades, referentes às regiões dos telômeros.

Completude da montagem

A análise do BUSCO resultou da comparação de 130 genes cópia única ortólogos entre os eucariotos (Tabela 1-9). A montagem resultou em 97% desses genes, sendo a sua maioria duplicados. A maior quantidade de genes duplicados foi observado para a montagem anterior de CL Brener e TCC, todos os genomas híbridos.

Cepa	Completos *	Único	Duplicados	Fragmentados	Ausentes
CL Brener 2023	127	26	101	3	0
CL Brener	129	80	49	1	0
Berenice	128	128	0	1	1
BrazilA4	129	129	0	1	0
Dm28c	122	112	10	8	0

TCC	129	21	108	1	0
Yc6	128	1	1	1	1

**Único+Duplicados*

Comparação entre montagens

Ao avaliar a qualidade da montagem da cepa CL Brener em relação aos genomas de outras cepas já sequenciadas, verificou-se que a montagem obtida apresenta bons parâmetros, tanto em termos de número de sequências genômicas montadas quanto em relação aos valores de N50 e N75 (Tabela 1-10). Dentre todas as montagens analisadas ela possui um dos menores número de sequências montadas. Ao comparar as métricas entre as cepas hibridas, nossa montagem chegou a valores superiores que a TCC. Vale ressaltar que a montagem do genoma de CL Brener apresentou inclusive métricas melhores que algumas genomas de cepas não híbridas, com exceção apenas de BrazilA4 e Yc6 que utilizam de metodologias de sequenciamento mais recentes (WANG, Wei *et al.*, 2021b).

Tabela 1-10 Dados finais de qualidade de montagem da cepa CL Brener de *T. cruzi* e outras da mesma espécie, todas montadas com base em *reads* longas.

Cepa	Quantidade	Maior sequência	Tamanho	GC%	N50	N90	L50	L90	auN	Referência
Brener2023	446	2407335	84908214	51,53	548600	107874	49	174	656226,4	-
Berenice	923	926516	40801262	51,2	156193	16124	61	430	236304,2	(DÍAZ-VIRAQUÉ *et al.*, 2019)
BrazilA4	402	2738928	45556784	51,58	914771	27220	17	128	974049,5	(WANG, Wei *et al.*, 2021a)
Bug2148	929	1305792	55157397	51,27	200364	21923	64	489	310383,8	(CALLEJAS-HERNÁNDEZ

										et al., 2018)
Dm28c	636	1645565	53271887	51,56	317638	30658	47	308	399616,2	(BERNÁ *et al.*, 2018a)
TCC	1236	1305230	87058484	51,72	264196	23494	92	536	341227,3	(BERNÁ *et al.*, 2018a)
YC6	266	2951016	47218089	51,58	889019	52667	18	98	970236,4	(WANG, Wei *et al.*, 2021a)

Os resultados da comparação de montagem de CL Brener do presente trabalho e o genoma já publicado (WEATHERLY; BOEHLKE; TARLETON, 2009) demonstraram que muitos dos “cromossomos” obtidos anteriormente estão representados (Figura 1-4), corroborando também com os dados de long-reads nos contigs (Figura 1-1). Em alguns casos, quando avaliada separadamente, a similaridade entre scaffolds e cromossomos, é observado a presença de *scaffolds* montados apresentando semelhanças com mais de uma sequência da antiga montagem (Figura 1-5).

O *scaffold* TcBrS07 (Figura 1-5) obteve similaridade com dois cromossomos da montagem de 2005 (TcChr33 e TcChr11), bem como o TcBrS010 com similaridade aos TcChr36-P e TcChr36-S. Outros *scaffolds* tinham um padrão semelhante. Ao avaliar a anotação nestas regiões, observamos que regiões repetidas *in tandem* dificultam o alinhamento, como observado para TcBrS001 (Figura 1-6). Esse foi o maior scaffold montado, com mais de 2.4Mb de comprimento. Ele apresentou alta similaridade ao cromossomo TcChr41-P, maior dentre os montados anteriormente onde cerca de 1,5Mb são altamente correspondentes, possuindo uma grande seção, entre 1,5Mb e 2Mb, não apresentando correspondência com alta exatidão, coincidindo com uma região de satélite (Figura 1-6 B). Outro padrão observado foi a ocorrência de uma mesma região presente em diferentes cromossomos, como regiões contendo elementos

transponíveis, observado em TcBrS059 (Figura 1-7), que pode ser encontrado em diferentes orientações nos diferentes cromossomos da montagem anterior.

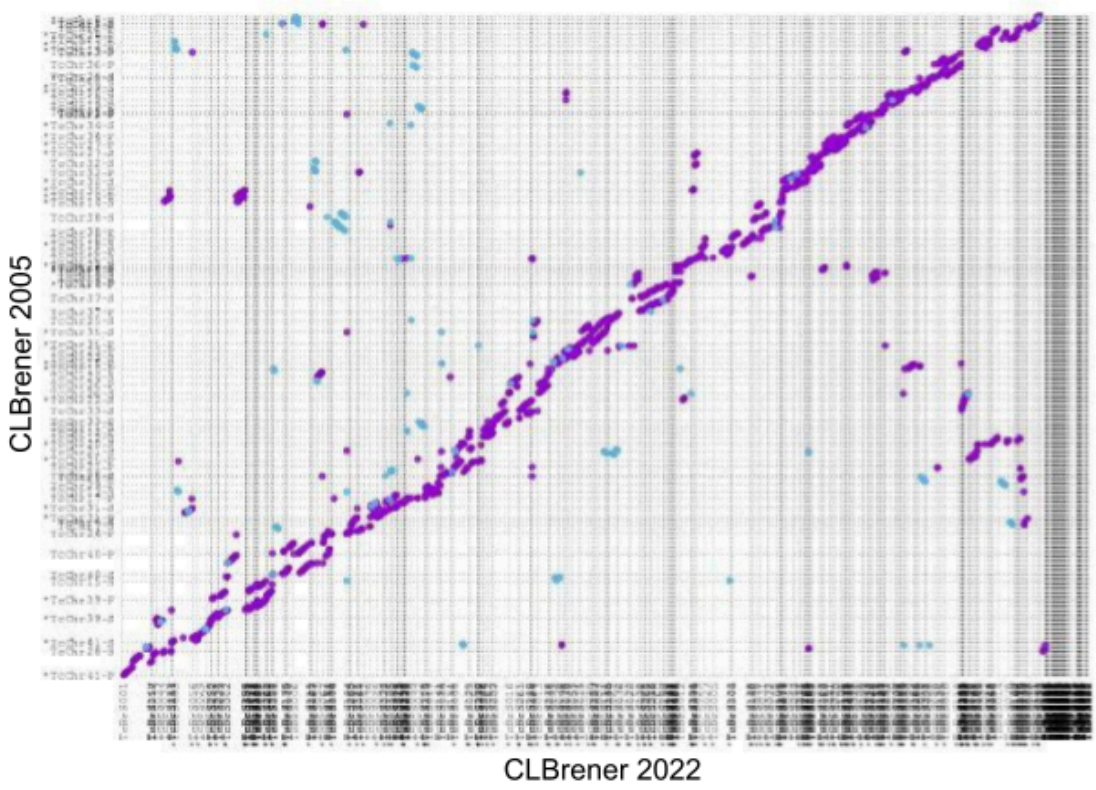


Figura 1-4 Representação gráfica do alinhamento entre nossa montagem final de *T. cruzi* cepa CL Brener, contra a montagem antiga da mesma cepa. A nova montagem é representada no eixo x e o eixo y estão o cromossomo da montagem anterior. Cada ponto representa similaridade de sequência, sendo que os em roxo se alinham na mesma direção e os em azul claro, na direção complementar.

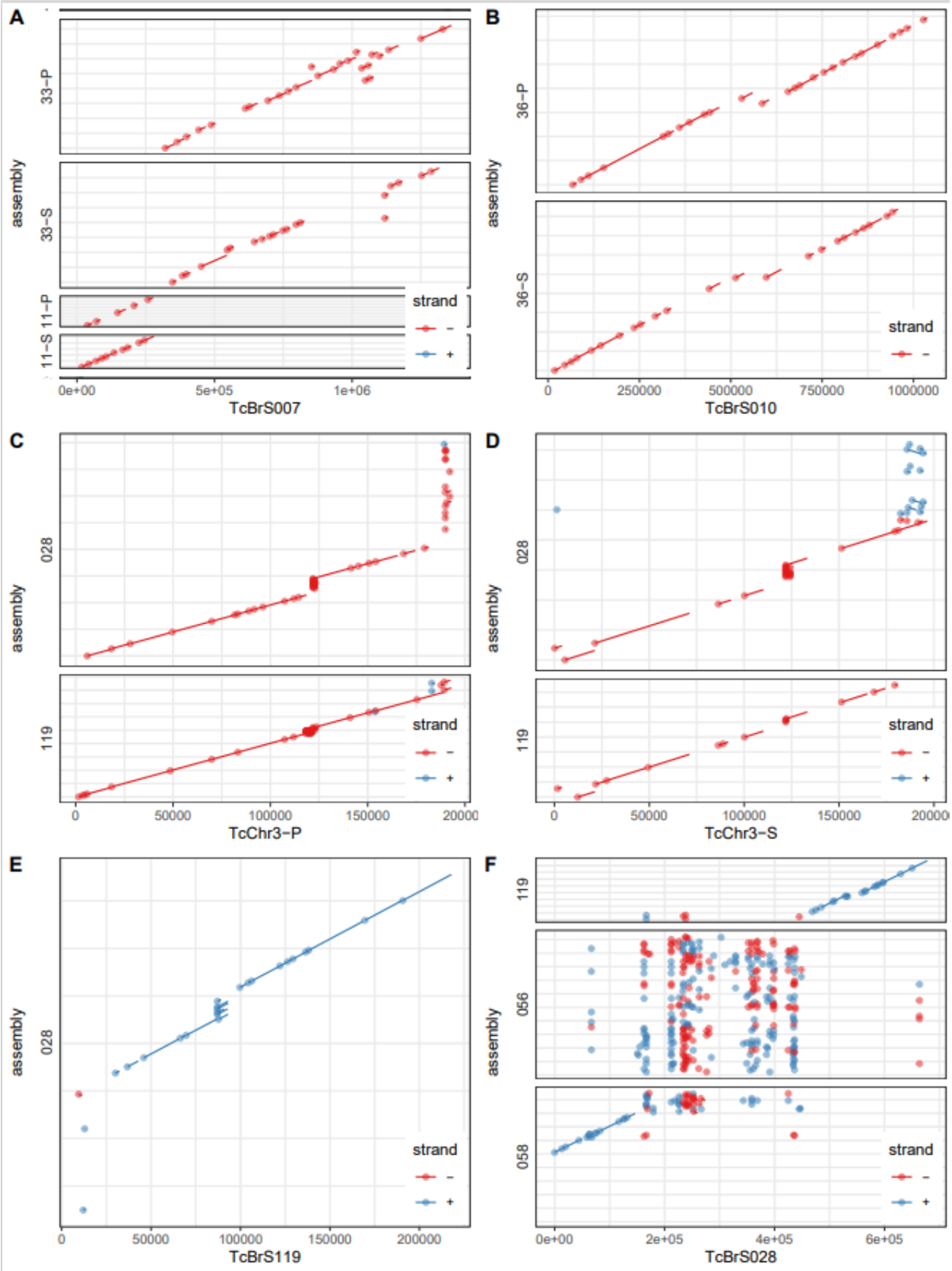

A
assembly
33−P
33−S
11−P
11−S
strand
−
+
0e+00
5e+05
1e+06
TcBrS007
B
assembly
36−P
36−S
strand
−
0
250000
500000
750000
1000000
TcBrS010
C
assembly
028
119
strand
−
+
0
50000
100000
150000
20000
TcChr3−P
D
assembly
028
119
strand
−
+
0
50000
100000
150000
20000
TcChr3−S
E
assembly
028
strand
−
+
0
50000
100000
150000
200000
TcBrS119
F
assembly
119
056
058
strand
−
+
0e+00
2e+05
4e+05
6e+05
TcBrS028

Figura 1-5 Representação gráfica do alinhamento entre o *scaffold* TcBrS007 **(A)** e TcBrS010 **(B)** da nossa montagem final da cepa CL Brener de *T. cruzi*, contra os cromossomos antigos montados. Os *scaffolds* são representados no eixo x e o eixo y são os cromossomos da montagem anterior. Cada ponto representa similaridade de sequência, sendo que os em roxo se alinham na mesma direção e os em azul claro, na direção complementar. (C) e **(D)** representam o alinhamento dos cromossomos TcChr3-P e TcChr3-S na nova montagem. Sendo representados os scaffolds no eixo y e os cromossomos no eixo x. Os scaffolds TcBrS028 e TcBrS119 foram alinhados entre si, tendo TcBrS119 **(E)** como referência e posteriormente TcBrS028 **(F)**.

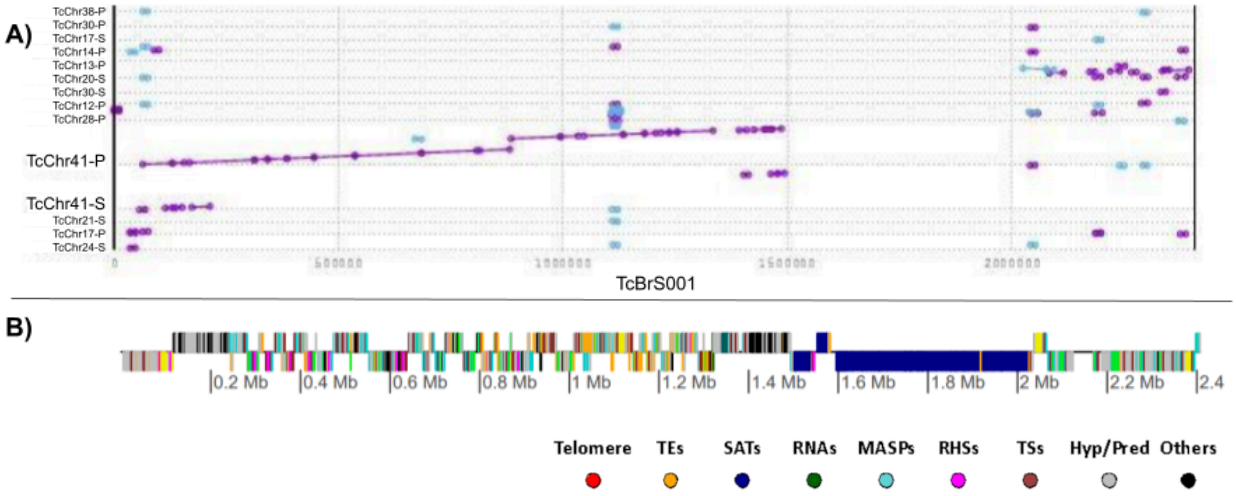


Figura 1-6 (A) Representação gráfica do alinhamento entre o scaffold TcBrS001 contra os pseudo cromossomos de CL Brener de 2005. A scaffolds TcBrS001 da nova montagem é representada no eixo x e no eixo y são os cromossomos da montagem anterior. Cada ponto representa similaridade de sequência, sendo que os em roxo se alinham na mesma direção e os em azul claro, na direção complementar. **(B)** Representação cromossômica do scaffold TcBrS001, nas caixas os diferentes genes/elementos anotados no cromossomo. Em azul um destaque as regiões de satélite repetitivas.

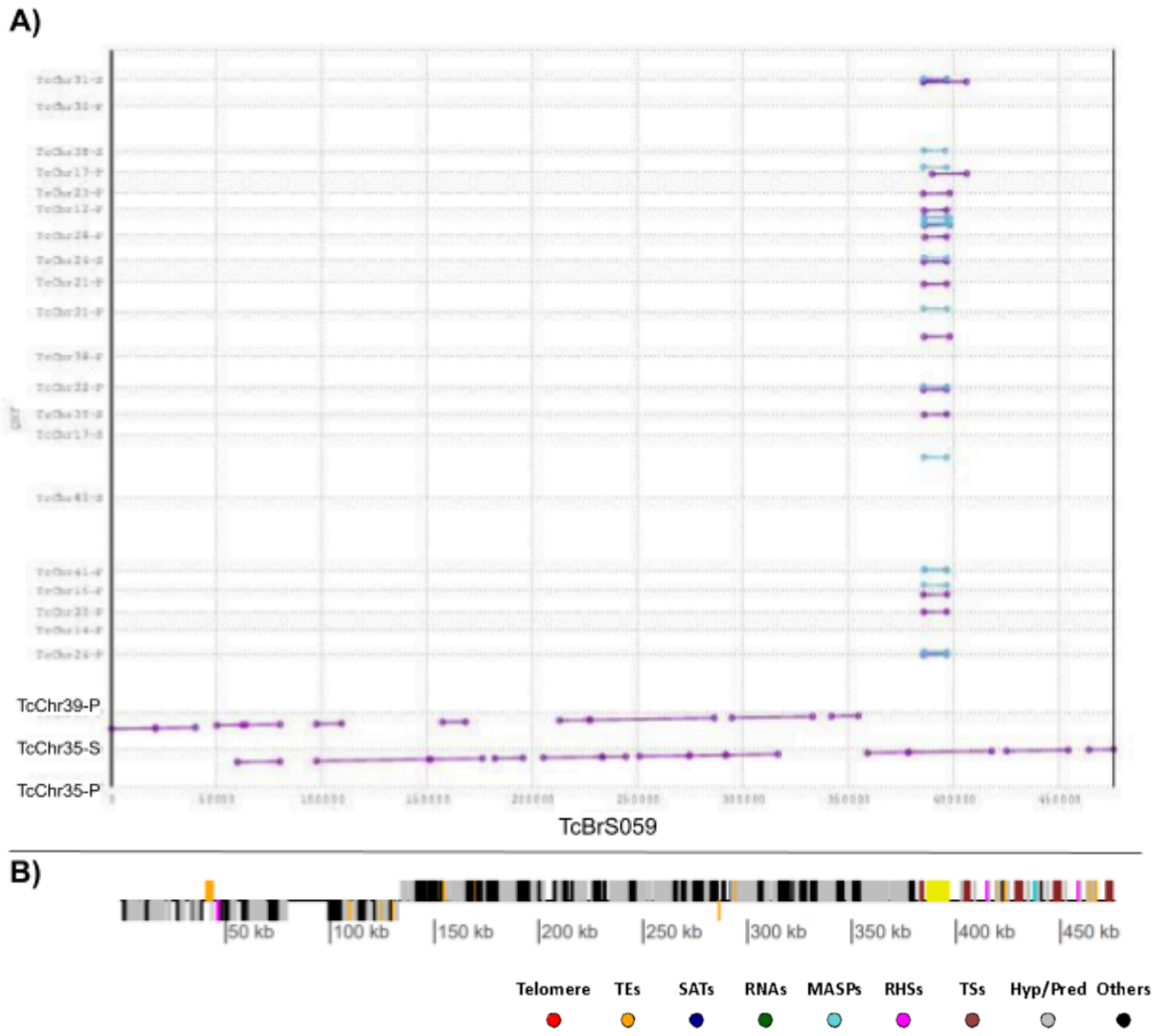


Figure 1-7 (A) Representação gráfica do alinhamento entre o scaffold TcBrS059 contra os pseudo cromossomos de CL Brener de 2005. A scaffold TcBrS059 é representada no eixo x e no eixo y são os cromossomos da montagem anterior. Cada ponto representa similaridade de sequência, sendo que os em roxo se alinham na mesma direção e os em azul claro, na direção complementar. (B) Representação da anotação do scaffold TcBrS059, nas caixas os diferentes genes/elementos anotados no cromossomo. Em amarelo um destaque para as regiões de elementos transponíveis.

Detecção de possíveis sítios de recombinação

Para identificar possíveis sítios de recombinação entre os genomas parentais (TcII e TcIII) da cepa CL Brener (TcVI), foi realizado o mapeamento das *reads* Illumina das cepas Y (TcII) e 231 (TcIII) na montagem do genoma CL Brener gerada no presente trabalho. Um total de 91 *scaffolds* tiveram algum tipo de inversão na profundidade entre reads derivadas das cepas representantes das DTUs TcII e TcIII (Figura 1-S2). Ou seja, essas regiões apresentam alta cobertura de reads derivadas de uma cepa de um dos haplótipos parentais, bem como baixa ou nenhuma cobertura *reads* do outro haplótipo, seguida de uma inversão de cobertura, onde a primeira passa a apresentar pouca ou nenhuma cobertura e a outra passa a ter cobertura aumentada na mesma região (Figura 1-8). A cobertura de reads de PacBio nestas regiões de inversão (Figura 1-9), corroboram com a noção de que são de fato pontos de recombinação que ocorreram durante a evolução do genoma híbrido de CL Brener.

Figura 1-8 Representação gráfica da profundidade de reads de bibliotecas de um representante dos DTUs TcII e TcIII, além das leituras longas e leituras curtas usadas na montagem, nos diferentes scaffolds montados.

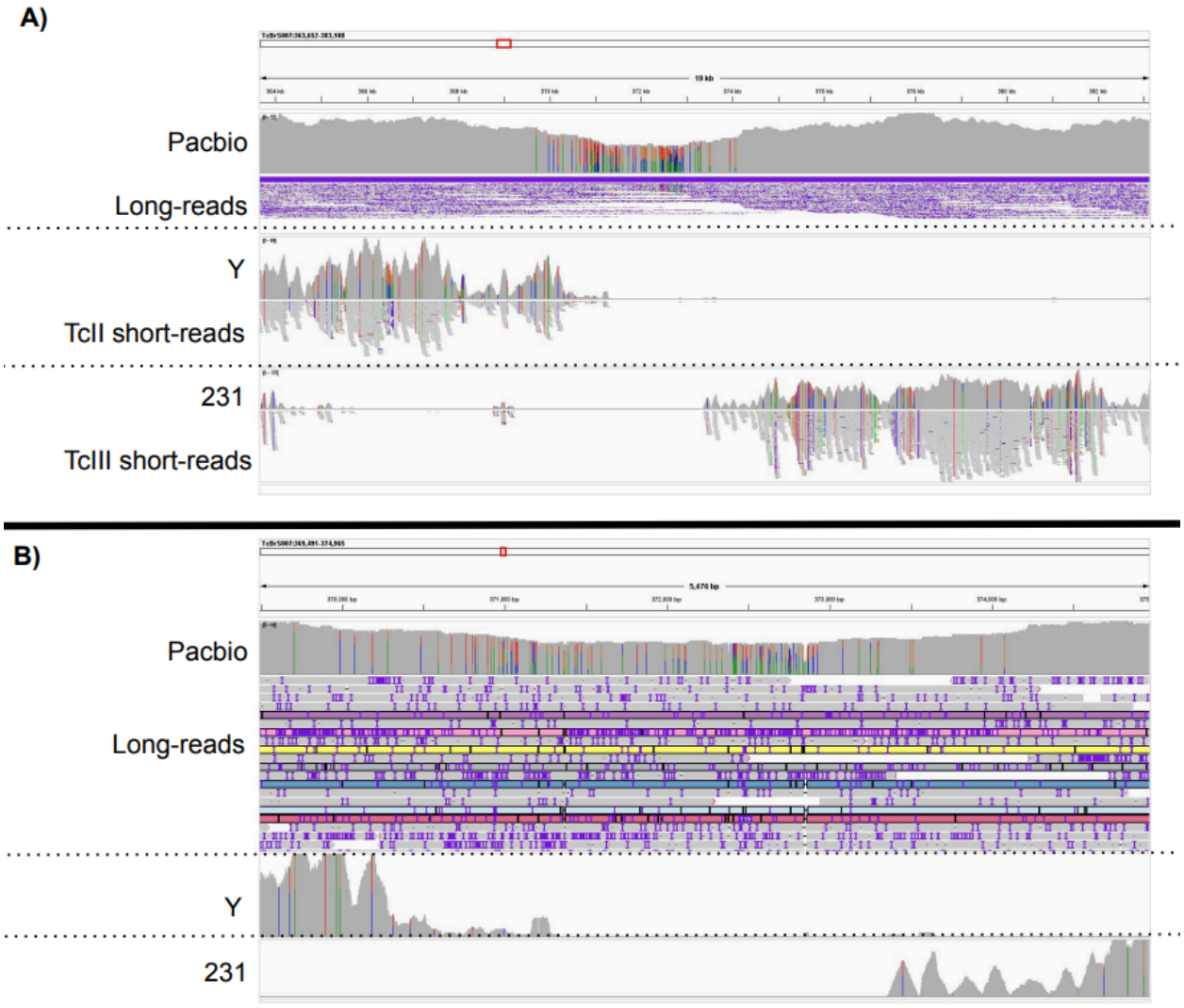


Figura 1-9 Imagem do IGV da região no scaffolds TcBrS007 onde há mudança de reads alinhada de Y para 231 (A). Abaixo (B) a mesma região mas dando enfoque nas reads longas que vão desde a região com alinhamento apenas em Y até a região com alinhamento apenas em 231. As linhas verticais coloridas representam variantes pontuais ou indels e o comprimento representa a frequência da variante quando comparado ao genoma de referência. Já no box Long-reads as linhas horizontais representam as reads alinhadas, em roxo são pontos de indels. As linhas em destaque em Long-reads em B (roxa, lilás,

amarela, anil, azul e vermelha) são aquelas que contemplam toda a extensão da região.

Ao avaliar o ponto de recombinação encontrado na sequência 133 e 64 da cepa TCC (BERNÁ *et al.*, 2018b) por similaridade contra a montagem, nós observamos o mesmo padrão, sinalizando como um possível ponto de recombinação compartilhado entre as cepas híbridas de *T. cruzi*. A análise usando as cepas 231 e Y contra TCC e nossa montagem de CL Brener permitiu avaliar os mesmos dados de sequenciamento dos DTUs parentais. Esse ponto de inversão no mapeamento entre os diferentes DTUs em dois scaffolds tanto na nova montagem quanto em TCC sugere que ambas as cepas podem ter se originado a partir de um mesmo evento de hibridação ou pode representar um hotspot de recombinação característico das cepas híbridas.

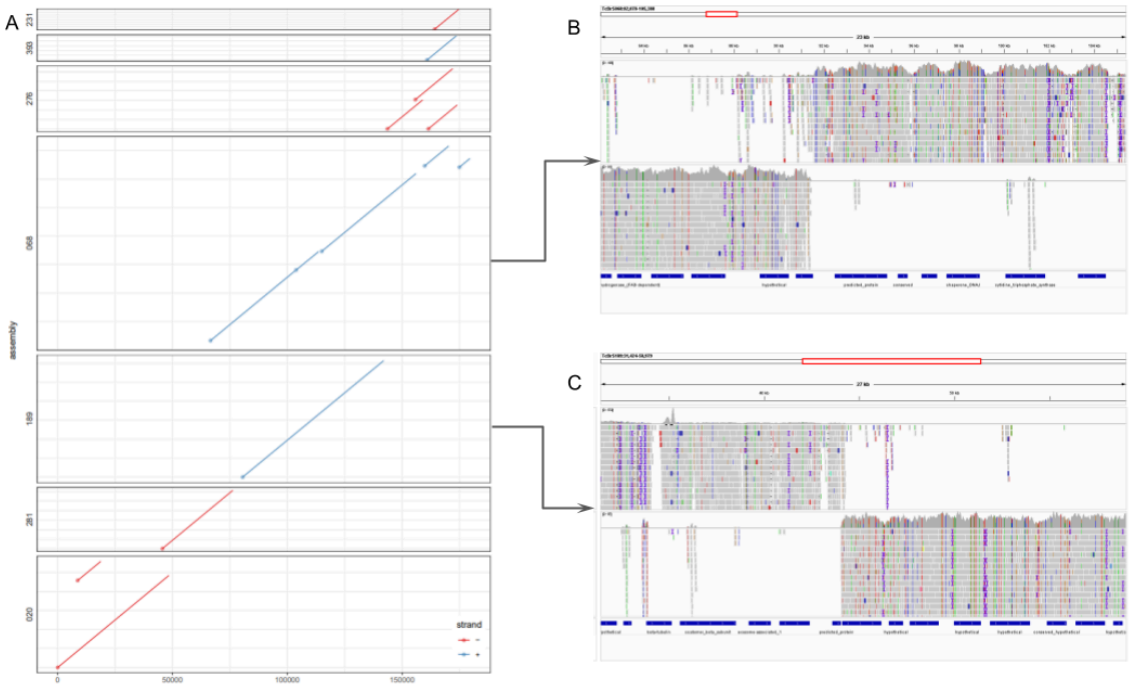


Figura 1-10 Representação gráfica do alinhamento entre o scaffold 133 da cepa TCC de *T. cruzi*, e os scaffolds da nova montagem de CL Brener. Ao lado, tem a imagem do IGV da região de possível troca nos scaffolds TcBrS068 (B) e TcBrS189 (C) da nova montagem alinhados com reads de Y (acima) e 231 (abaixo).

Perfil de CG na composição do genoma

Embora o genoma contenha aproximadamente 51% de conteúdo CG, existem diferenças entre porções do genoma, dependendo da estrutura e do conteúdo gênico (Figura 1-11). Isso é observado nos diferentes valores observados entre os membros das principais famílias multigênicas anotadas (Tabela 1-11), como DGF1, com valores superiores a 64% de conteúdo de GC, seguidas das mucinas com 57%. Quanto aos demais genes, os valores são mais próximos do genoma total, com 52%. Por outro lado, notou-se uma variação entre os genes preditos pelo Augustus, no que se refere aos genes que não apresentaram similaridade superior ao corte na anotação geral dos genes codificantes de proteínas, mas também não foram anotados com avaliação específica dos membros das famílias multigênicas. Isso pode significar que entre esses membros preditos possa haver ainda sequências como de DGF-1 e mucinas, devido ao valor de GC.

Tabela 1-11 Porcentagem de CG nas regiões da montagem de CL Brener de *T. cruzi* para os respectivos genes.

Genes	GC%
DGF	64,1
MASP	51,7
Mucina	57,4
RHS	50,9
TS	53,4

Proteínas preditas	54,4
Outros genes	52,5

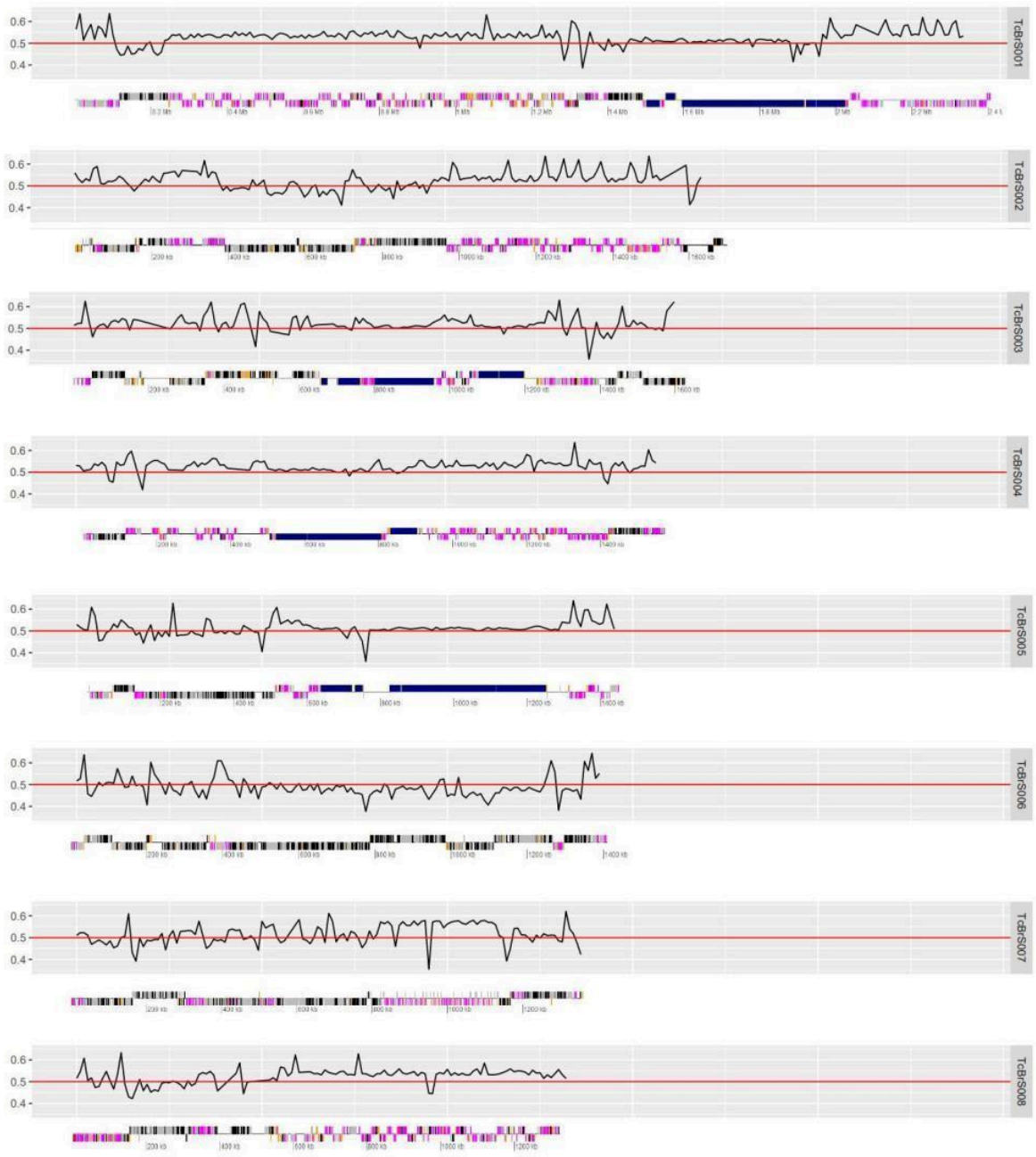


Figure 1-11 Valores do conteúdo CG (eixo y) em janelas de tamanho de 10 Kpb ao longo dos *scaffolds* (eixo x) montados da cepa CL Brener de *T. cruzi*. A linha em preto representa a

porcentagem GC na janela utilizada, em vermelho o valor de 0,5 (50%) e abaixo de cada gráfico de linha, a representação da anotação de cada *scaffold* em questão destacando em magenta os genes membros das principais famílias multigênicas (DGF-1, MASP, Mucinas, RHS e Trans-sialidases), em azul, regiões satélite e em amarelo, elementos transponíveis.

Avaliação de aneuploidias em CL Brener

Para avaliar a ocorrência de aneuploidias no genoma de CL Brener, foi utilizado o CADIn, desenvolvido nesta tese (ver capítulo 2). A análise foi feita nas 40 scaffolds de maior tamanho. (Figura 1-12). Nas análises de frequência alélica (Figura 1-12 A), para muitas sequências analisadas, foi verificada a ocorrência de uma baixa quantidade de variantes bialélicas (eixo y). Entretanto, um número considerável de variantes (>100) foi detectado para scaffolds TcBrS006, TcBrS010, TcBrS019, TcrS026, TcBrS030, TcBr038 e TcBrS039, sugerindo fortemente um padrão dissômico. O mesmo foi observado para TcBrS021 e TcBrS038, que apresentaram um padrão trissômico. Como a maioria muitas scaffolds apresentaram baixa contagem de variantes, dificultando a robustês da análise de somia por frequência alélica, foi realizada a predição de somia por profundidade de reads (Figura 1-12 B). Nesta abordagem, a grande maioria das sequências foram classificadas como dissômicas. A predição de trissomia sugerida pela análise de frequência alélica para os scaffolds TcBrS021 e TcBrS038 foi confirmada pela análise de profundidade de reads, que também sugere a ocorrência de outros scaffolds com padrão supranumerário, como TcBrS001, TcBrS013, TcBrS035, os apresentaram uma baixa contagem de variantes. Em resumo, as análises sugerem uma prevalência de cromossomos dissômicos e alguns supranumerários, confirmando a natureza aneuplóide do genoma de CL Brener.

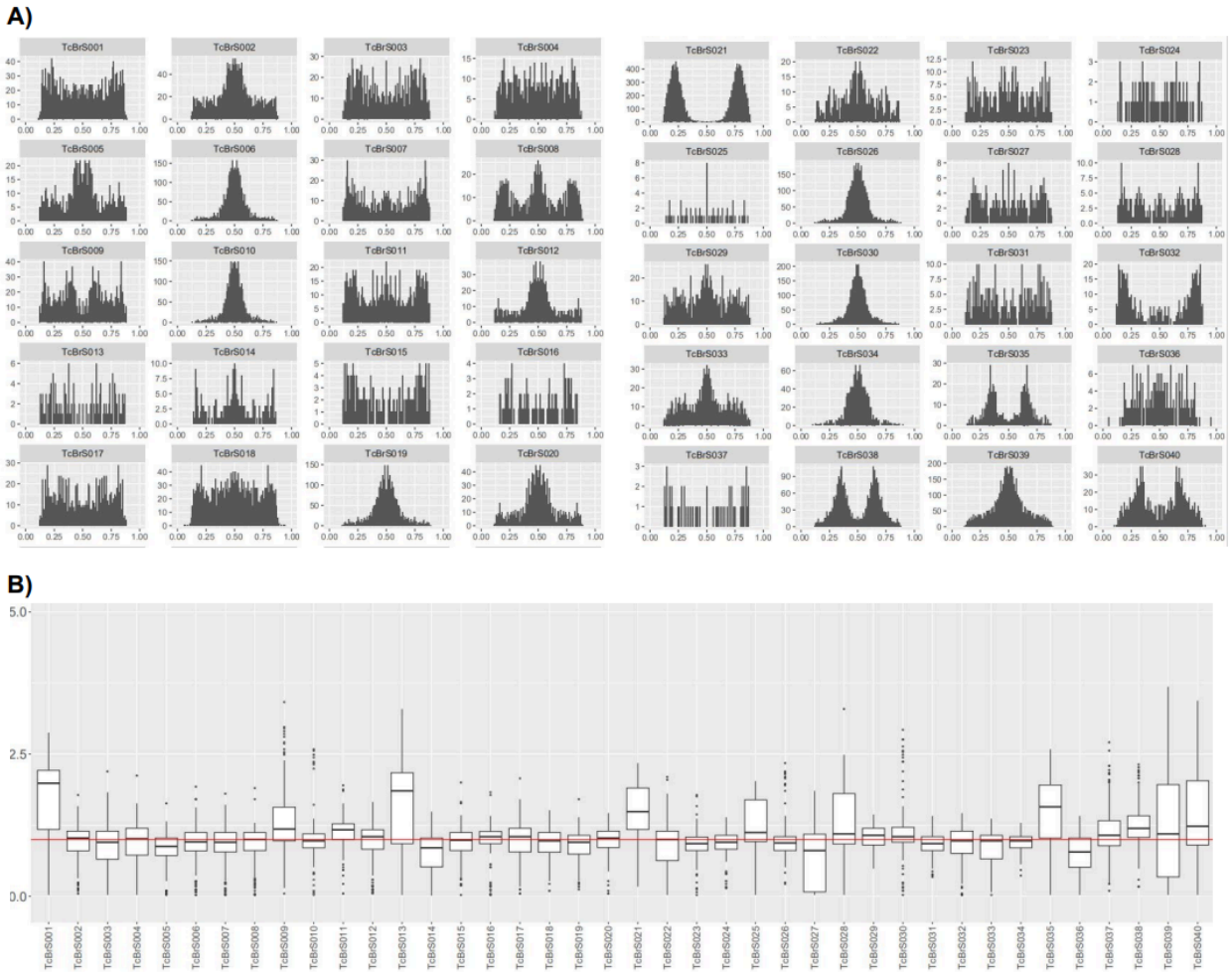


Figura 1-12. Estimativas de somia pelo CADIn para scaffolds de CL Brener com base na frequência alélica de SNPs heterozigóticos (A) e profundidade de reads (B). Em A, o eixo x contém valores de frequência de SNPs heterozigóticos e o eixo y corresponde ao número de posições cromossômicas com uma determinada frequência alélica. Em B, cada boxplot refere-se a um cromossomo e a linha vermelha (valor 1) corresponde ao número de cópias cromossômicas por genoma haplóide. Os valores foram calculados com base na cobertura mediana de todos os genes em um cromossomo normalizados pela cobertura do genoma.

CAPÍTULO 2

CADIn - CHROMOSOMAL AMPLIFICATION AND DELETION INFERENCE TOOL

CADIn, ferramenta desenvolvida para avaliação de variações de somias cromossômicas baseado em *read depth coverage* e análises de frequência alélica.

Metodologia

Informações iniciais

O CADIn é um pipeline desenvolvido para avaliar ploidia e somática por análise de variantes e cobertura de profundidade de leitura (*reads*), usando apenas um comando via terminal. Como entrada, requer um ou mais arquivos no formato BAM contendo *reads* de NGS, um arquivo de referência, usado para mapeamento, em formato FASTA e um arquivo de anotação no formato GFF3. O arquivo BAM pode ser originado do mapeamento das *reads* usando ferramentas como BWA (LI, Heng; DURBIN, 2009) e Bowtie (LANGMEAD *et al.*, 2009). Os dados de entrada não precisam ser indexados, filtrados ou classificados, a própria ferramenta faz a análise quando solicitado. Usando parâmetros padrão, o arquivo será filtrado pela qualidade de mapeamento de 30 (-q), e indexado automaticamente. O arquivo de referência pode ser filtrado, após o mapeamento, pelas regiões/cromossomos de interesse, para análises de ploidia ou somia. Os dados de anotação devem ter o nome dos cromossomos condizentes com a referência e precisam ser filtrados por região anotada (gene, CDS, mRNA). O CADIn usa genes como padrão, mas isso pode ser modificado (-f). Quando necessário, a conversão do GFF3 para BED, por exemplo, a ferramenta realiza automaticamente. Os cálculos de ploidia genômica e somia cromossômica são baseados em duas análises independentes: frequência de SNPs heterozigóticos e pela profundidade de *reads* cobertas, como descrito abaixo.

Análise de SNPs heterozigóticos

Esta análise consiste em avaliar a frequência alélica de variantes de nucleotídeos pontuais (SNPs) encontrada nas *reads* mapeadas no genoma de referência. Somente as variantes heterozigotas com duas variantes são consideradas para o cálculo (-v). Todas as variantes usadas precisam ter uma qualidade superior a 10 (-k) e pelo menos cinco leituras de suporte (-d). Posteriormente, é feita a relação entre a contagem de cada variante e a profundidade total de reads na posição. Depois, os valores calculados para todas as variantes são agrupados por cromossomo e plotados em um gráfico de frequência, ou os dados de todos os cromossomos são usados para inferir a ploidia genômica. Ao observar o padrão de frequência dos SNPs heterozigóticos é possível determinar a somia cromossômica ou ploidia do genoma. O maior número de variantes em 0,5 caracteriza o conjunto avaliado como dissômico/diplóide, pois cada variante corresponde a 50% da contagem total. Valores de 0,33 e/ou 0,77 infere-se trissomia/triploidia, e 0,25, 0,5 e/ou 0,75, tetrassomia/tetraploidia.

Profundidade de *reads* cobertas

Outra metodologia usada para avaliar a somia cromossômica é usar a profundidade de reads mapeadas nas regiões anotadas. Computando cada posição dos genes (região anotada), a profundidade do gene é calculada com base na mediana (-m). Os genes que apresentarem menos de 50% (-l) de sua extensão coberta por *reads* serão descartados, e os demais normalizados pela profundidade do genoma, que

pode ser calculado com base na profundidade de todos os genes ou na profundidade total do genoma mapeado (-p).

Após a normalização, o teste Grubs é executado iterativamente para cada cromossomos para remover regiões com outliers. Esse teste permite remover regiões genômicas altamente repetitivas, como aquelas ricas em famílias multigênicas. Através deste novo conjunto de dados, *boxplots* e *heatmaps* são gerados. A avaliação desses gráficos precisa ser feita em uma base comparativa devido às discrepâncias e variações entre os cromossomos de uma amostra. Um valor de 1 é considerado a linha de base do genoma; no entanto, valores acima/abaixo correspondem a um aumento ou diminuição do número de cópias do cromossomo em relação ao genoma total. O teste de classificação de Mann-Whitney-Wilcoxon é realizado para avaliar estatisticamente se valores elevados ou abaixo da linha de base são significativamente válidos (p-valor <0,05). Para isso, avalia-se os valores encontrados e observados e se a são inferiores a 0,5 e 1 e acima de 1, 1,5, 2, 2,5 e 3. Um teste de correlação (pairwise test) também é realizado comparando os cromossomos da amostra, ajustados por correção de Bonferroni, e permitindo um teste assintótico. Todas as comparações estão em um gráfico do tipo heatmap destacando valores-p < 0,01, 0,05 e 0,10, destacando as diferenças quando houver.

Validação da pipeline

Usando amostras de banco de dados

Para validações, cinco amostras clínicas de *Saccharomyces cerevisiae* e duas espécies de *Leishmania* (*L. infantum* e *L. major*) tiveram dados de *reads* baixadas no NCBI SRA (LEINONEN *et al.*, 2010) (Tabela 2-1). Eles foram avaliados por qualidade e posteriormente trimados pelo Fastqc (versão 0.11.8) (ANDREWS, 2010) e Trimmomatic (versão 0.33) (BOLGER; LOHSE; USADEL, 2014), respectivamente. Concomitantemente, os arquivos de genoma (formato FASTA) e anotação (formato GFF) de *S. cerevisiae* S288C (GCA_000146045.2) obtidos do NCBI GenBank (O'LEARY *et al.*, 2016) e os de *Leishmania major* cepa Friedlin e *Leishmania infantum* JPCM5 recuperados do TritrypDB (versão 46) (ASLETT *et al.*, 2010). Em seguida, todas as bibliotecas foram mapeadas para o respectivo genoma de referência com BWA (versão 0.7.12) (LI, Heng; DURBIN, 2009). Depois que o CADIn foi executado com os parâmetros padrão, os arquivos BAM de cada espécie foram executados separadamente usando os genomas de referência e as respectivas anotações. As sequências de genoma mitocondrial foram removidas antes das estimativas.

Tabela 2-1 - Todos os códigos utilizados para análise e respectivos códigos SRA e plataforma de sequenciamento em que obra as bibliotecas foram sequenciadas, bem como o código do estudo e o projeto em que estão cadastradas.

Codigo	SRA	Plataforma	Model	SRAStudy	BioProject
CBS2919	SRX1648794	ILLUMINA	Illumina HiSeq 2000	SRP072079	PRJNA315044
CBS7837	SRX1648801	ILLUMINA	Illumina HiSeq 2000	SRP072079	PRJNA315044

CBS9564	SRX1648806	ILLUMINA	Illumina HiSeq 2000	SRP072079	PRJNA315044
YJM1098	ERX2050129	ILLUMINA	Illumina HiSeq 3000	ERP023217	PRJEB20998
YJM466	ERX2050130	ILLUMINA	Illumina HiSeq 3000	ERP023217	PRJEB20998
LinJ	ERX005632	ILLUMINA	Illumina Genome Analyzer II	ERP000169	PRJEB2115
LmjF	ERX005636	ILLUMINA	Illumina Genome Analyzer II	ERP000169	PRJEB2115

Comparação com ferramenta existente

O nQuire (WEIS *et al.*, 2018) também foi utilizado para avaliar as bibliotecas de *S. cerevisiae*, e *Leishmania sp*. para detectar aneuploidias em organismos com poucas variantes. Os arquivos de mapeamento para essa análise foram classificados e filtrados por qualidade de mapeamento 30, e um formato BED foi gerado com cada uma das coordenadas dos cromossomos.

Bibliotecas simuladas

Bibliotecas simuladas com diferentes níveis de cobertura foram geradas para avaliar a estratégia de profundidade de cobertura usada pelo CADin para estimativa de somia. Bibliotecas de *reads* pareadas de *S. cerevisiae* S288C foram simuladas usando o software ART (HUANG *et al.*, 2012). As bibliotecas foram criadas com cobertura 50x. Os cromossomos 2, 3, 4 e 5 tiveram a cobertura aumentada para 75x, 100x, 125x e 150x. Essas simulações foram realizadas usando HiSeq 1000 (100bp), HiSeq 2000 (100bp), HiSeq 2500 (125bp e 150) e HiSeqX TruSeq (150bp), para verificar diferentes

plataformas e tamanhos de *reads* variáveis. Posteriormente, as *reads* foram mapeadas na referência, e o CADIn realizou a análise, conforme descrito acima.

Implementação

Todas as análises foram realizadas em um laptop com 8 gigabases de ram, 256 gigabases de espaço em disco e um processador i5 com 2 núcleos. Foi testado para sistemas baseados em Debian e Redhat e macOS. O pipeline fez a análise usando 1 núcleo de processamento, não sendo paralelizado. O tempo gasto para cada análise foi relatado, sendo que o tempo para as bibliotecas simuladas foi registrado separadamente para verificar a influência do método de sequenciamento na execução.

Resultados

Conjuntos de dados reais

O CADin também foi usado para avaliar a ploidia do genoma e a somia cromossômica em conjuntos de dados de genomas reais de *S. cerevisiae* e *Leishmania sp*.. Inicialmente, para avaliar a ploidia global do genoma dos isolados de *S. cerevisiae*, análises de frequência alélica foram realizadas considerando os SNPs heterozigotos de todos os cromossomos. Foi detectado em *S. cerevisiae* diferentes valores de somia entre e dentro das amostras. Os dados de frequência alélica de todo o genoma indicaram que: CBS7837 e YJM1098 são diplóides, CBS2919 e YJM466 triplóides e CBS9564 tetraplóide (Figura 2-1). Ao avaliar cada cromossomo separadamente por abordagem de frequência alélica (Figura 2-2) e pelos resultados das análises de profundidade (Figura 2-3), variações na somia foram observadas para isolados os CBS2919 (cromossomo 1), CBS9564 (cromossomos 9 e 13), YJM1098 (cromossomo 12) e YJM466 (cromossomos 6 e 9), enquanto CBS7837 não exibiu aneuploidias.

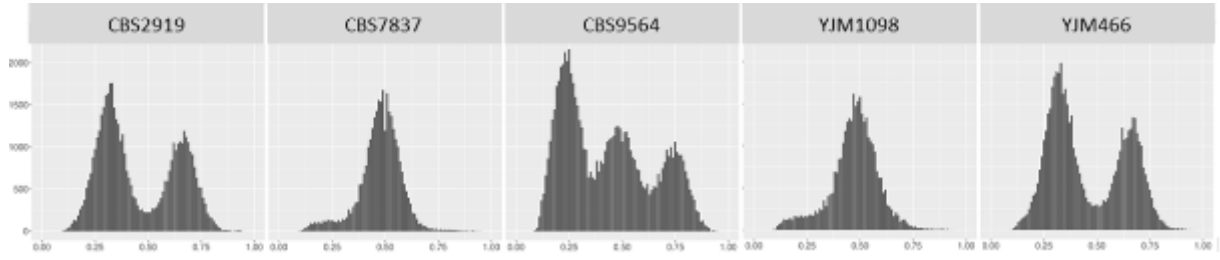


Figura 2-1 - Estimativas de ploidia pelo CADIn de diferentes amostras de *Saccharomyces cerevisiae* com base na frequência alélica de SNPs heterozigotos de todos os cromossomos. O eixo x contém valores de frequência de SNPs heterozigóticos e o eixo y corresponde ao número de posições cromossômicas com uma determinada frequência alélica.

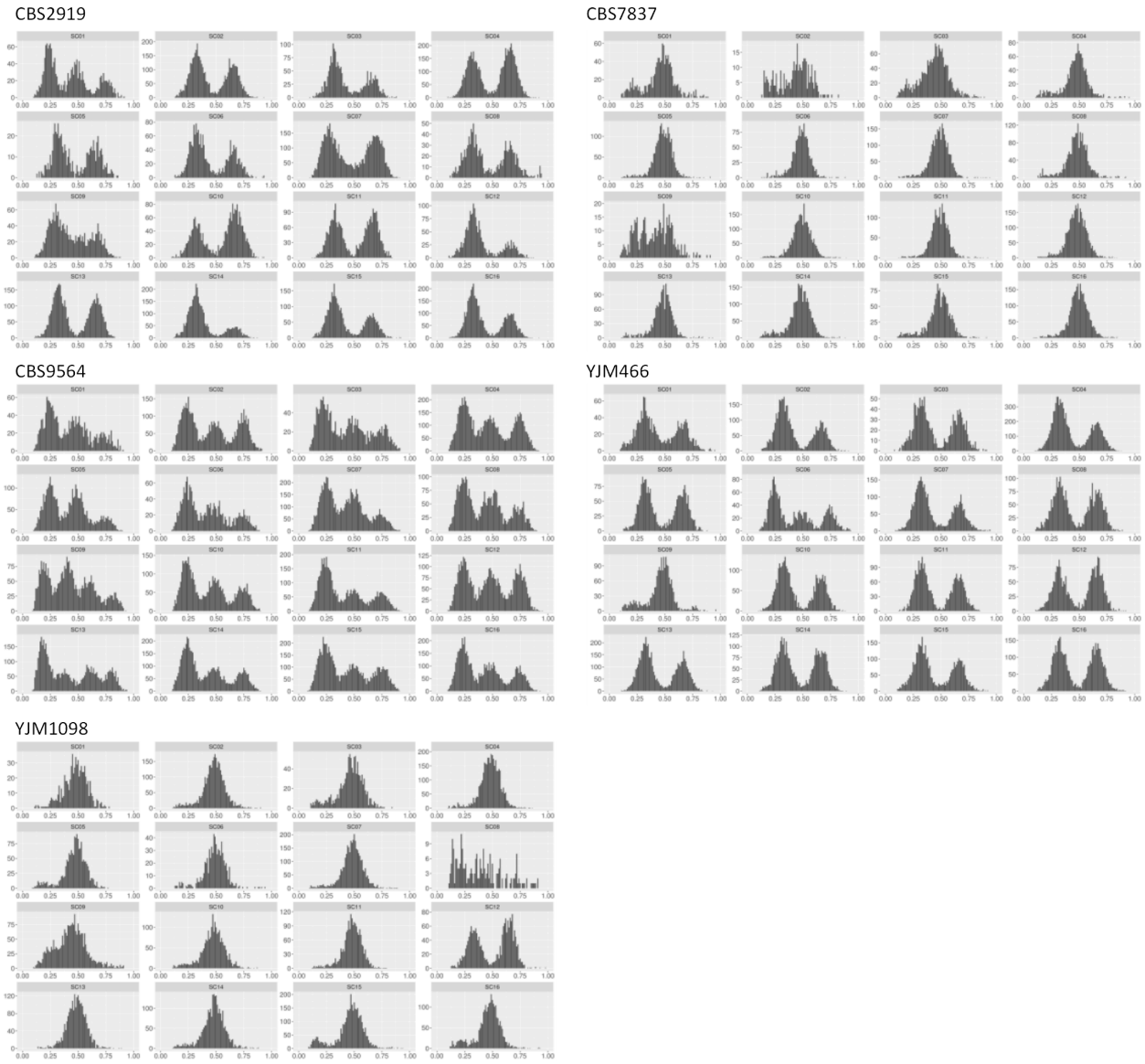


Figura 2-2 - Análise da frequência alélica de SNPs heterozigóticos calculada com o CADIn para cada cromossomo para amostras de *S. cerevisiae*. O eixo X contém valores de frequência alélica e o eixo Y corresponde ao número de ocorrências com essa frequência.

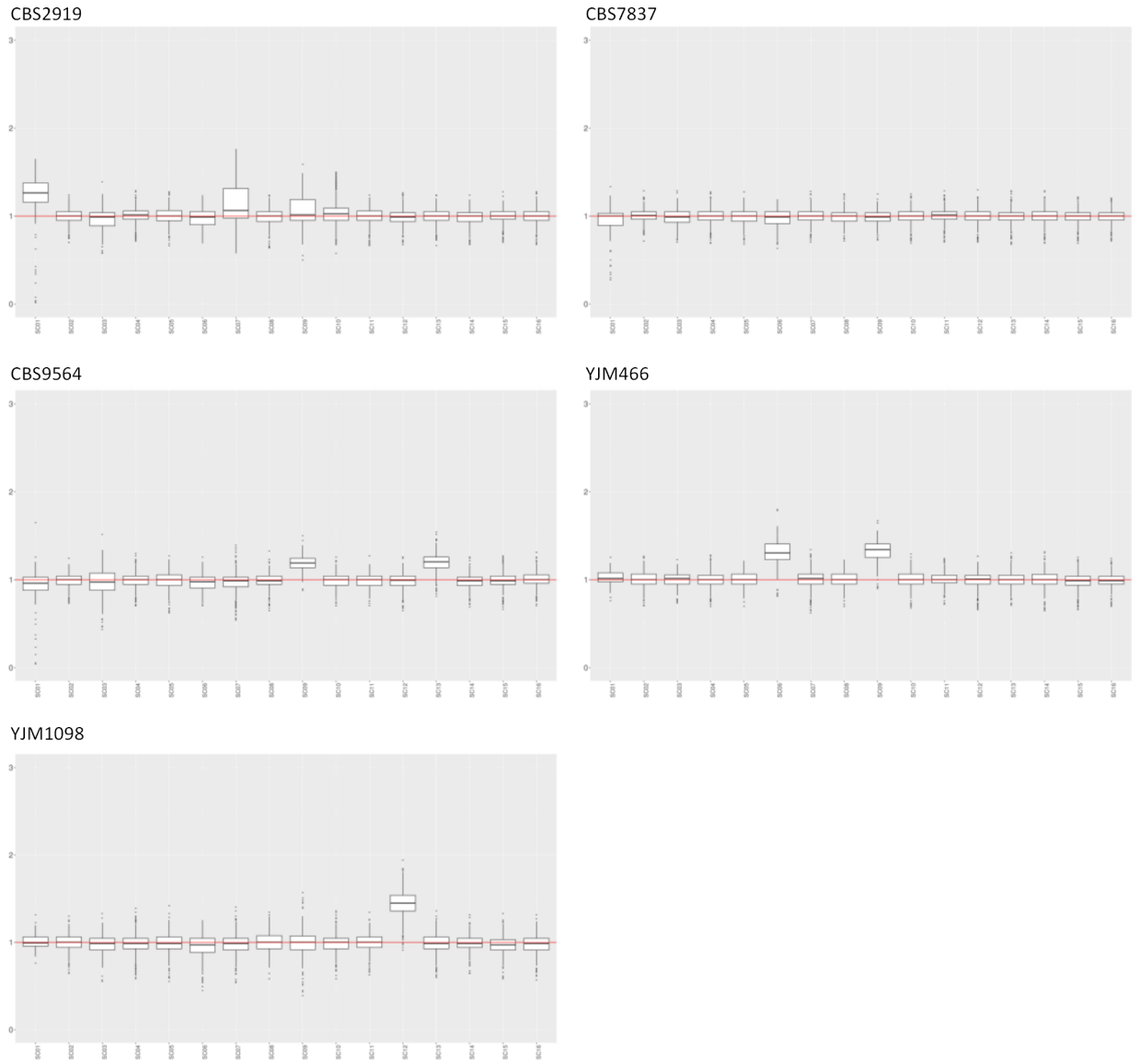


Figura 2-3 - Estimativas de soma para cada cromossomo de amostras de *S. cerevisiae* usando análises de profundidade de *reads* cobertas calculada pelo CADIn. Os valores foram calculados com base na cobertura mediana de todos os genes em um cromossomo normalizado pela cobertura do genoma. Cada boxplot refere-se a um cromossomo e a linha vermelha (valor 1) corresponde ao número de cópias cromossômicas por genoma haplóide.

Para *S. cerevisiae* foi possível observar as diferenças através do valor médio da profundidade dos genes naquela sequência (Figura 2-4). O mesmo padrão foi visto

pelo teste de correlação (Figuras 2-5) quando avaliado a profundidade dos cromossomos, apontando estatisticamente as diferenças e semelhanças das somia estimadas.

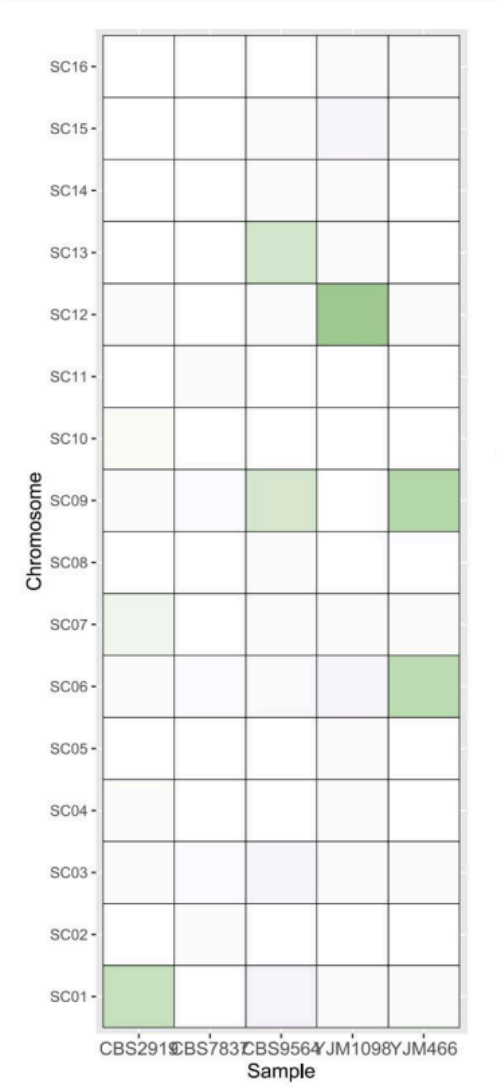


Figura 2-4 - Heatmap com a mediana dos valores de profundidade normalizados dos genes presentes em cada cromossomo (eixo y) para as diferentes amostras (eixo x). As cores representam o aumento ou redução do número de cópias do cromossomo em comparação com a linha de base de 1 (branco).

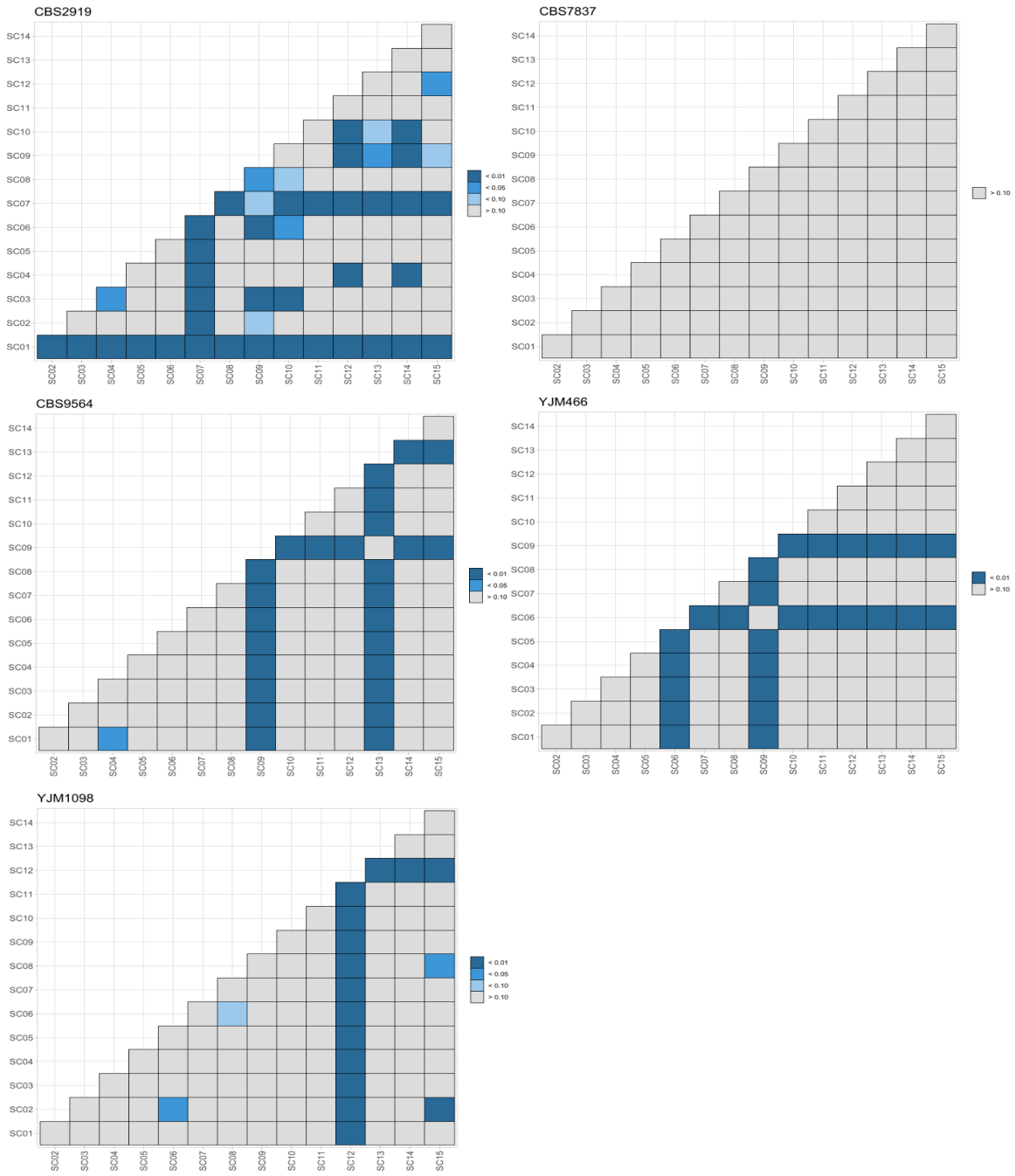


Figure 2-5 - Heatmap dos valores de significância obtidos pelo teste de Wilcox pairwise para amostras de *S. cerevisiae*, avaliadas entre cromossomos da mesma espécie. Em cinza, os que não obtiveram diferença estatística, e em azul, gradativamente, aumentou a confiança estatística.

Os resultados da frequência alélica de SNPs heterozigóticos para as duas espécies de *Leishmania* analisadas mostraram um número extremamente reduzido de variantes, não só em relação aos cromossomos individuais (Figura 2-6), mas também em relação ao genoma total (Figura 2-7), comprometendo a robustez das predições. As análises de profundidade de reads, por outro lado, foram muito mais informativas, sugerindo um padrão aneuplóide para ambos isolados de cada espécie analisada (Figura 2-8). Em *L. major*, apenas o cromossomo 31 apresentou-se como supranumerário, enquanto que *L. infantum*, um número muito maior de cromossomos apresentou este padrão (cromossomos 6, 8, 9, 17, 22, 25, 31, 33 e 35). Esses resultados foram estatisticamente confirmados (Figura 2-9).

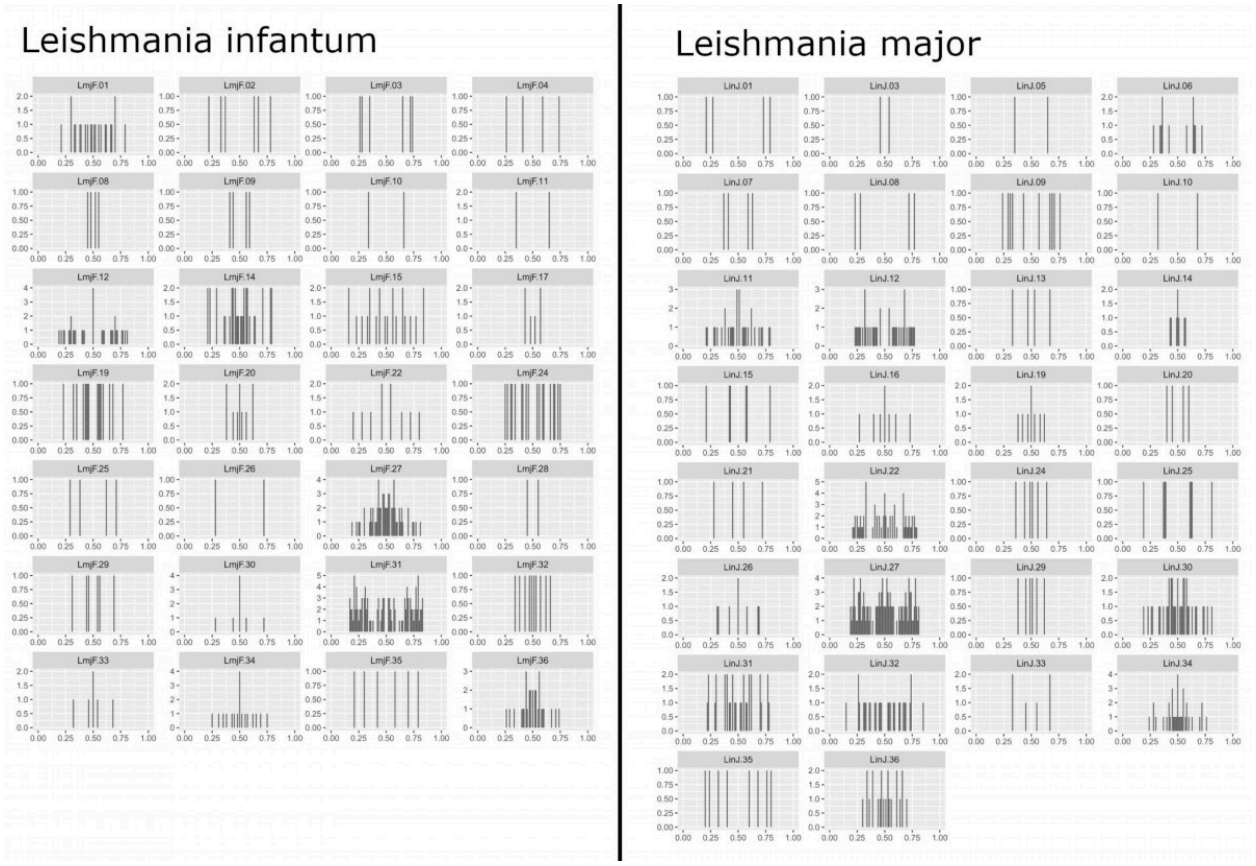


Figura 2-6 - Análise de frequência alélica para 36 cromossomos de *L. infantum* e *L. major*. O eixo X contém valores de frequência alélica de SNPs heterozigóticos e o eixo Y corresponde ao número de ocorrências com essa frequência. Cada caixa corresponde a um cromossomo. Os cromossomos ausentes são os que não apresentaram SNPs heterozigotos.

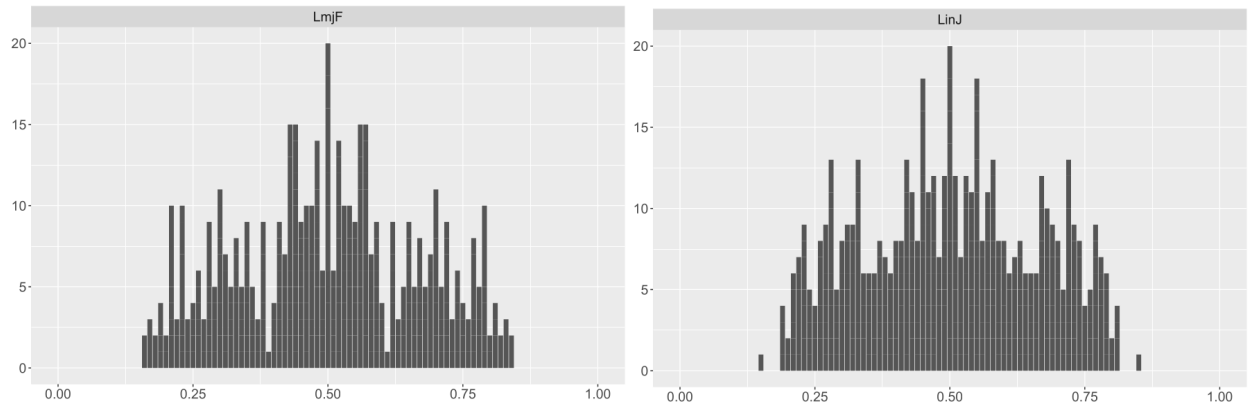


Figure 2-7 - Estimativas de ploidia genômica para diferentes espécies de *Leishmania* sp por frequência alélica de SNPs heterozigotos de todos os cromossomos. Painel à esquerda, *L. major*; painel à direita, *L. infantum*. O eixo x contém valores de frequência de SNPs heterozigóticos e o eixo y corresponde ao número de posições cromossômicas com uma determinada frequência alélica.

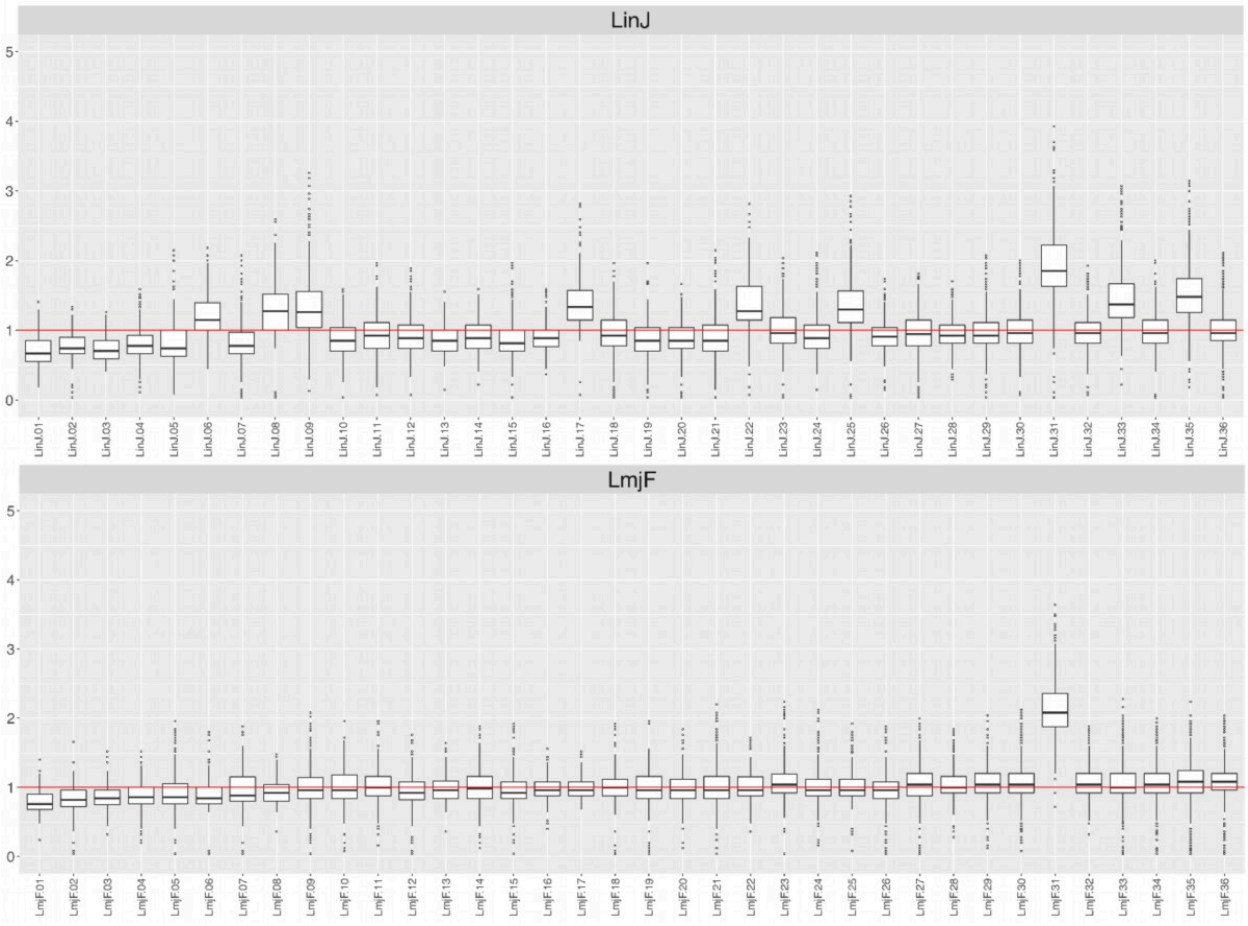


Figura 2-8 - Estimativas de somia de cada cromossomo de amostras de *L. infantum* e *L. major* com base na análise de profundidade. As estimativas foram calculadas com base na cobertura mediana de todos os genes no respectivo cromossomo normalizado pela cobertura do genoma. O eixo x contém os nomes dos cromossomos e o eixo y corresponde à soma cromossômica calculada. A linha vermelha é o valor base de 1, os valores correspondem ao número de cópias cromossômicas por genoma haplóide.

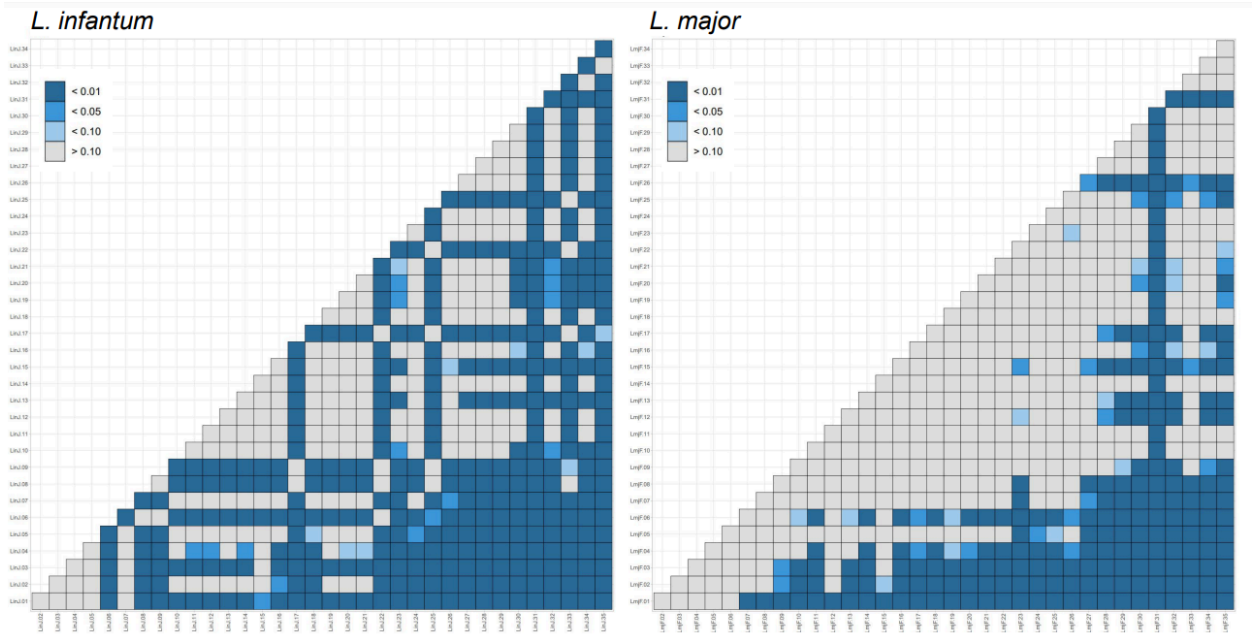


Figura 2-9 - Heatmap dos valores de significância obtidos pelo teste de Wilcox pairwise para amostras de *Leishmania*, avaliadas entre cromossomos da mesma espécie. Em cinza, os que não obtiveram diferença estatística, e em azul, gradativamente, aumentou a confiança estatística.

Ao contrário dos resultados obtidos por análise de profundidade por CADIn para amostras de *Leishmania*, aqueles obtidos por nQuire sugeriram um padrão de tetrassomia para todos os cromossomos analisados de *L. infantum* (Figura 2-10), algo nunca observados em outros estudos .

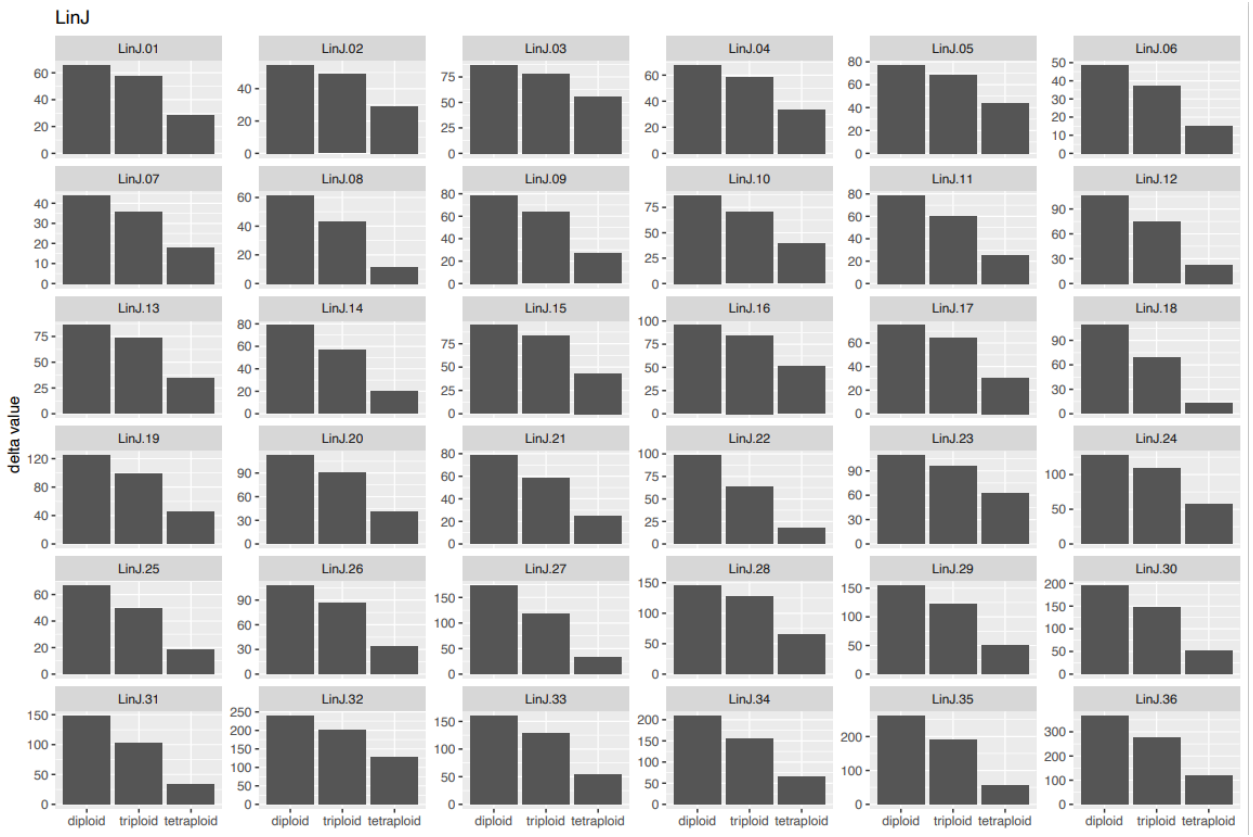


Figura 2-10 - Resultados do nQuire representados em um gráfico com valores delta para amostra de *L. infantum*. Esses valores foram obtidos após análise de *Gaussian Mixture Model* (eixo y), os valores correspondem ao teste de diploidia, triploidia e tetraploidia (eixo x). Entre as três avaliações, aquela com os valores mais baixos terá a melhor confiança.

Dados simulados

O desempenho do CADIn foi também avaliado na detecção de variações de ploidia e somia usando dados de cobertura de *reads* simuladas para o genoma de *Saccharomyces cerevisiae*. Para este fim, as coberturas de profundidade de *reads* para os cromossomos 2, 3, 4 e 5 foram aumentadas artificialmente para 75x, 100x, 125x e 150x (ver material e métodos). Para todas as bibliotecas, foram observadas as

variações de somia esperadas para os cromossomos 2, 3, 4 e 5, sem alteração na somia dos outros cromossomos (Figura 2-11).

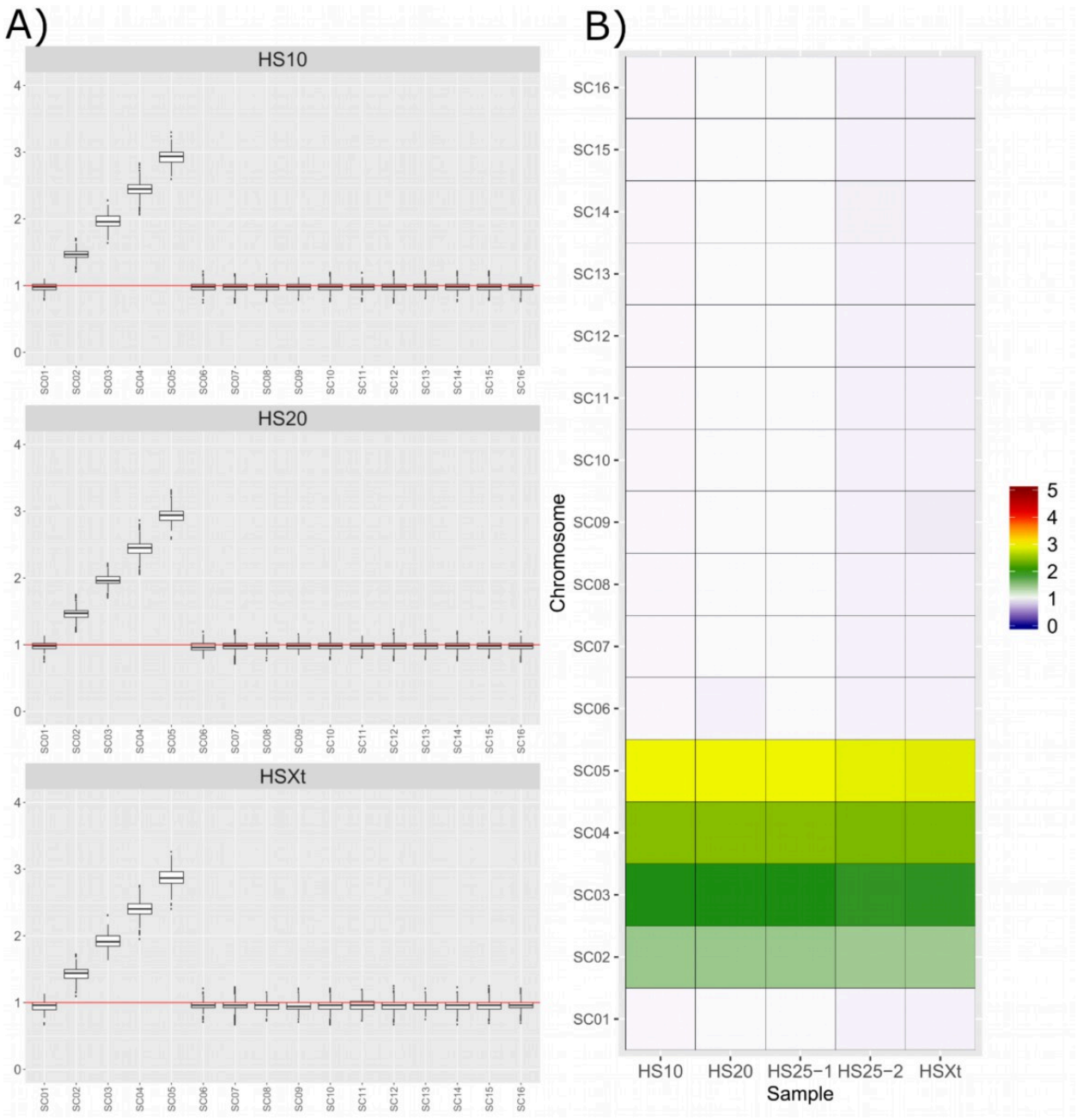


Figura 2-11 - Estimativas da somia de cada cromossomo das amostras simuladas de *S. cerevisiae*. As bibliotecas de *reads* simuladas foram geradas utilizando a ferramenta ART. A) As bibliotecas foram criadas com cobertura de 50x, exceto os cromossomos 2, 3, 4 e 5, criados

com cobertura de 75, 100, 125 e 150 x, respectivamente. As estimativas foram baseadas na cobertura mediana de todos os genes em um cromossomo normalizado pela cobertura do genoma. O eixo x contém os nomes dos cromossomos e o eixo y corresponde à somia predita do cromossomo. Cada boxplot refere-se a um cromossomo e a linha vermelha (valor 1) corresponde ao número de cópias cromossômicas por genoma haploide. B) Heatmap com a mediana dos valores de profundidade normalizados dos genes presentes em cada cromossomo (eixo y) para as diferentes amostras (eixo x). As cores da caixa representam o aumento ou redução do número de cópias do cromossomo em comparação com a linha de base de 1 (branco).

DISCUSSÃO

O ano de 2005 representa um marco no estudo da biologia dos tripanossomatídeos, quando os genomas de *Trypanosoma cruzi*, *Trypanosoma brucei* e *Leishmania major*, importantes parasitos causadores de doenças negligenciadas, foram publicados conjuntamente na revista Science (BERRIMAN *et al.*, 2005; EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005; EL-SAYED; MYLER; BLANDIN; *et al.*, 2005; IVENS *et al.*, 2005; TØRRESEN *et al.*, 2019). Os três genomas apresentam características e níveis de complexidade bastante distintos, tendo sido empregadas, em cada projeto, estratégias de sequenciamento que resultam em diferentes níveis de completude (BARTHOLOMEU; TEIXEIRA; CRUZ, 2021). Como resultado, o genoma de CL Brener, cepa referência do projeto genoma de *T. cruzi*, apresentou-se como o mais fragmentado.

Características intrínsecas do genoma de CL Brener como alto conteúdo repetitivo, presença de cromossomos supranumerários e a natureza híbrida, composto por genomas parentais divergentes (TcII e TcIII), representam por si só enormes desafios para a reconstrução de cromossomos completos durante o processo de montagem. Além disso, a estratégia de sequenciamento escolhida na época do projeto genoma foi a *whole genome shotgun* (WGS), que é atualmente a técnica usada para reconstruir genomas. Mas na época do projeto genoma de *T. cruzi*, WGS havia sido aplicada em poucos projetos e a metodologia de sequenciamento disponível (Sanger) resultava em níveis de cobertura muito menores do que se observa hoje com

next-generation sequencing. Além disso, nenhuma informação de mapa físico do genoma e nenhuma etapa de closure, para fechar gaps gerados durante a montagem, foram incorporados no projeto. Todos estes aspectos contribuíram para a fragmentação do genoma de CL Brener disponível em banco de dados públicos. No presente trabalho, ressaltamos o desafio da montagem do genoma de CL Brener usando metodologias de sequenciamento e computacionais atualmente disponíveis.

Na busca por um genoma referência mais completo para *T. cruzi,* vários grupos tem usado tecnologias de sequenciamento e estratégias computacionais aplicadas no sequenciamento de genomas de diferentes cepas do parasito (BAPTISTA *et al.*, 2018; BERNÁ *et al.*, 2018a; BRADWELL *et al.*, 2018; CALLEJAS-HERNÁNDEZ *et al.*, 2018; DÍAZ-VIRAQUÉ *et al.*, 2019; EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005; WANG, Wei *et al.*, 2021a). No presente trabalho, a escolha do HGAP como montador para os *contigs* ao invés do CANU se mostrou mais eficaz para o genoma de CL Brener, uma vez que resultou em melhores métricas de montagem. Apesar de ambos utilizarem de uma metodologia baseada em OLC e adaptada para reads longas devido a alta taxa de erro (aproximadamente 15%) (CHIN *et al.*, 2013; GONZALEZ-GARCIA *et al.*, 2023; KOREN *et al.*, 2017; LI, Zhenyu *et al.*, 2012), foi possível verificar uma melhor performance do HGAP. Esses valores divergentes quanto às métricas de montagem podem ser devido ao impacto da cobertura vertical que foi obtido no processo de sequenciamento para as *reads* longas. O CANU não apresenta bons resultados para cobertura menor que 50x (KOREN *et al.*, 2017), sendo recomendada cobertura superior para métodos de montagem hierárquica (CHAKRABORTY *et al.*, 2016). Essa diferença presente no CANU em relação ao HGAP está refletida nos valores alcançados no N50

e N90. Comparando esses valores entre CANU e HGAP é possível observar que mesmo com cerca de 1000 sequências a menos montadas pelo primeiro os valores de N50 são quase 90% maiores para o HGAP. Isso demonstra um padrão inverso do desejado, no que diz respeito à montagem, decorrentes talvez uma possível sobreposição na montagem, o que é esperado uma vez que o genoma de *T .cruzi* CL Brener é cepa híbrida (EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005; WEATHERLY; BOEHLKE; TARLETON, 2009; ZINGALES *et al.*, 2012). Por conta disso, os contigs gerados pelo HGAP foram escolhidos para as próximas etapas de montagem, algo também realizado em TCC, outra cepa híbrida (BERNÁ *et al.*, 2018a).

A metodologia de scaffolding usando etapas iterativas ajudou na montagem, pois o alto número de regiões repetitivas no genoma de CL Brener dificulta o processo (EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005; TØRRESEN *et al.*, 2019). A estratégia de juntar os contigs baseado no decréscimo gradual da cobertura em cada iteração por reads curta se mostrou promissor. Além disso, o uso de *reads* longas e curtas no processo de *scaffolding*, bem como *reads* de Sanger de bibliotecas de diferentes tamanhos de inserto auxiliaram o processo. Essa estratégia já havia sido utilizada para outros organismos usando outras ferramentas e metodologias (IANTORNO *et al.*, 2017; PAGE *et al.*, 2016; WANG, Anqi *et al.*, 2018). Além disso, a correção de montagem por *reads* curtas provou ser eficaz, conforme já descrito em outros trabalhos (CHAKRABORTY *et al.*, 2016; MAHMOUD *et al.*, 2019; ZHANG; JAIN; ALURU, 2020).

Com relação ao conteúdo gênico, ao avaliar a completude da montagem maior de 97% das sequências ortólogas e de cópias únicas foram encontradas. Ao comparar

com montagens anteriores, a mesma proporção foi encontrada quando avaliadas no BUSCO versão 5. O genoma como o de Berenice (DÍAZ-VIRAQUÉ *et al.*, 2019) já havia sido avaliado entretanto com a versão 3 do software que utiliza outro fluxo no processo de análise e outras sequências no seu banco de dados (MANNI *et al.*, 2021). A montagem de BrazilA4 e Yc6 também se utilizou do BUSCO para comparar não só as cepas montadas como Dm28c e TCC, entretanto na mesma versão que Berenice (WANG, Wei *et al.*, 2021a). Ao avaliar essas amostras destaca-se o número de sequências duplicadas para todas as cepas híbridas sinalizando para uma possível característica da espécie em manter cópias duplicadas de genes codificantes de proteínas mais basais, entretanto se faz necessário observar se esse número de cópias está relacionado em uma maior expressão.

Em relação à organização genômica, detectamos 24 scaffolds com as repetições teloméricas em uma das extremidades. Nenhum scaffold representa cromossomos completos, já que não foram gerados scaffolds com repetições teloméricas em ambas as extremidades, o que reforça a dificuldade de se resolver essas regiões durante o processo de montagem como verificado em outros trabalhos (EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005; WEATHERLY; BOEHLKE; TARLETON, 2009; ZINGALES *et al.*, 2012).

Como descrito no projeto genoma de 2005 ((EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005; TØRRESEN *et al.*, 2019) e em (BERNÁ *et al.*, 2018a; HERREROS-CABELLO *et al.*, 2020), observamos a ocorrência de clusters gênicos de proteínas de superfície, como MASPs, mucinas, trans-sialidases, e GP63 em regiões internas dos cromossomos, onde também se encontram elementos transponíveis e

outros genes como DGF-1, RHS. Essas regiões foram denominadas compartimento disruptivo (BERNÁ *et al.*, 2018a; HERREROS-CABELLO *et al.*, 2020), as quais se alternam com regiões denominadas compartimento core, onde se encontram genes que codificam proteínas estruturais e do metabolismo basal e que são sintênicas com regiões dos genomas de *T. brucei* e *Leishmania* (EL-SAYED; MYLER; BLANDIN; *et al.*, 2005). As regiões subteloméricas são enriquecidas com elementos transponíveis e genes de trans-sialidase, RHS e DGF-1, uma composição gênica distinta daquela do compartimento disruptivo.

Em relação aos elementos transponíveis, SINE (*short interspersed nuclear element*) e L1Tc são normalmente flanqueados por RHSs, conforme já relatado para esta espécie (DÍAZ-VIRAQUÉ *et al.*, 2019; THOMAS *et al.*, 2010). A família dos RHSs já foi descrita como responsável/facilitadora pela inserção principalmente de retrotransposons (THOMAS *et al.*, 2010). Em *T. bruce*i, a presença desses genes foi relacionada a processos evolutivos nas regiões subteloméricas (BRINGAUD *et al.*, 2002; NAGULESWARAN *et al.*, 2021). Já em outras cepas da DTU TcI de *T. cruzi,* o gene foi relacionado à aneuploidia segmentar (CRUZ-SAAVEDRA *et al.*, 2022). Os retrotransposons non-LTR, por outro lado, são descritos como localizados perto das extremidades dos cromossomos em regiões correspondentes à inversão e associados à ocorrência de recombinação (GHEDIN *et al.*, 2004; MACÍAS *et al.*, 2018). Dessa forma, a participação de elementos transponíveis na geração de variabilidade é algo factível, como sugerido previamente (TALAVERA-LÓPEZ *et al.*, 2021).

Berná e colaboradores (BERNÁ *et al.*, 2018a; HERREROS-CABELLO *et al.*, 2020) haviam demonstrado que o compartimento disruptivo apresenta maior conteúdo

GC que as demais regiões do genoma. Aqui verificamos que as regiões subteloméricas também apresentam esse padrão. Verificamos que os genes de DGF-1 e mucinas, comumente encontrados nessas regiões, apresentam maior conteúdo GC que os demais genes destas áreas, o que pode explicar o padrão verificado por (BERNÁ *et al.*, 2018a; HERREROS-CABELLO *et al.*, 2020). Genes que codificam proteínas preditas pelo nosso pipeline de anotação também apresentam maior conteúdo GC, podendo ser genes divergentes ou fragmentos gênicos destas famílias.

Conteúdos mais altos de GC já foram observados em genes de *Saccharomyces cerevisiae (KIKTEV et al., 2018)*, mamíferos (GALTIER, 2003; KUDLA; HELWAK; LIPINSKI, 2004) e tripanossomatídeos (EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005; HORN; BARRY, 2005). Em outros trabalhos, já foi relatado que o aumento do conteúdo de GC está relacionado com os níveis de expressão dos genes (CHÁVEZ *et al.*, 2017; KUDLA *et al.*, 2006). No scaffold TcBrS001, observou-se maior quantidade de guanina e citosina na extremidade do *scaffold*, além do número de genes pertencentes às famílias multigênicas. Possivelmente, no *scaffold* TcBrS008, onde também se observa a presença de sequências teloméricas no início do scaffold, as regiões subteloméricas podem estar associadas a uma maior taxa de mutações, bem como de expressão uma vez que há aumento de GC nestas regiões.

Dentre as famílias multigênicas, trans-sialidases apresentam um maior número de genes preditos. Para alguns destes membros, foi possível identificar o resíduo Tyr342, e motivos catalíticos e estruturais centrais, como Asp-box, VTV e SAPA, elementos característicos do grupo I da trans-sialidase (BURLE-CALDAS *et al.*, 2022; FREITAS, Leandro M. *et al.*, 2011; SCHENKMAN *et al.*, 1994). Esse é o único grupo da

família da trans-sialidase que apresenta função catalítica associada à transferência de ácido siálico a glicoconjugados presentes nas proteínas mucinas na superfície do parasito (OPPEZZO *et al.*, 2011). As mucinas foram anotadas e os subgrupos com previamente descritos (BUSCAGLIA *et al.*, 2006), identificados, exceto alguns genes que podem ser quimeras com genes de MASP, como previamente descrito (BARTHOLOMEU *et al.*, 2009; EL-SAYED; MYLER; BARTHOLOMEU; *et al.*, 2005).

Sítios de recombinação foram detectados no genoma híbrido de CL Brener, alguns compartilhados com a cepa TCC, também um híbrido derivado da mesma DTU de CL Brener, TcVI. Esses dados sugerem que um único evento de hibridização pode ter originado essa DTU ou que esses sítios sejam hot-spots de recombinação. Uma análise mais exaustiva destes pontos de recombinação precisa ser realizada para que conclusões mais assertivas possam ser feitas. Seria também interessante incluir nestas análises representantes da DTU TcV, uma outra linhagem híbrida, a fim de se ampliar o estudo da origem e evolução dos genomas híbridos de *T. cruzi.* Adicionalmente, seria interessante avaliar a contribuição de cada genoma parental nos híbridos para avaliar o processo de erosão genômica.

As aneuploidias, uma condição comum em diferentes isolados de *T. cruzi* (REIS-CUNHA *et al.*, 2018) e *Leishmania (ROGERS et al., 2011)*, mas raro no clado *T. brucei* (ALMEIDA *et al.*, 2018), adicionam uma outra camada de complexidade a estes genomas. Ao avaliar o número de cópias dos maiores scaffolds da montagem, foi possível observar sequências claramente dissômicas, enquanto outras apresentam mais de duas cópias, evidenciando a ocorrência de aneuploidias também em CL

Brener. Essas análises foram realizadas com a ferramenta CADIn, desenvolvida neste trabalho é discutida a seguir.

Em resumo, apesar da montagem de CL Brener obtida neste trabalho não apresentar melhores métricas quando comparadas com as cepas não híbridas BrazilA4 e Yc6, ela representa uma considerável evolução em relação à montagem anterior de CL Brener e melhores métricas daquelas obtidas na montagem de TCC, a outra cepa híbrida já sequenciada (BERNÁ *et al.*, 2018a; HERREROS-CABELLO *et al.*, 2020). O trabalho evidenciou a natureza aneuplóide do genoma de CL Brener e corroborou o entendimento da estrutura genômica do parasito, formado por dois compartimentos, core e disruptivo, que apresentam diferenças em relação ao conteúdo GC, repertório gênico e padrões de diversidade. Foi ainda possível identificar a ocorrência de sítios de recombinação entre os genomas parentais de CL Brener e TCC, contribuindo para um melhor entendimento da evolução dos genomas híbridos do parasito.

Neste trabalho também desenvolvemos CADIn, uma ferramenta automatizada que executa análises de frequências alélicas de SNPs heterozigóticos e cobertura de profundidade de reads para detectar ploidia genômica e variações de somia cromossômica usando dados NGS.

Para validar o CADIn, as variações de ploidia e somia cromossômica em linhagens de *S. cerevisiae* foram avaliadas. Esta espécie foi selecionada, uma vez que foi observado que mais de 40% dos isolados avaliados apresentaram variações no número de cópias cromossômicas (STROPE *et al.*, 2015; ZHU; SHERLOCK; PETROV, 2016). A ferramenta CADIn provou ser muito eficaz na detecção de ploidia por frequência de SNPs heterozigóticos e por cálculo de profundidade, obtendo-se alta

concordância na predição pelos dois métodos, e concordantes com os resultados encontrados anteriormente (SANTOS *et al.*, 2017; ZHU; SHERLOCK; PETROV, 2016). Para as amostras CBS2919, CBS7837 e CBS9564, os mesmos resultados foram obtidos por outra ferramenta (WEIS *et al.*, 2018). No entanto, o CADIn difere das análises anteriores por fornecer suporte estatístico.

A frequência alélica é uma forma eficaz para analisar variações de somia e já é bem utilizada em outras metodologias (SANTOS *et al.*, 2017; WEIS *et al.*, 2018; YU *et al.*, 2014). Em genomas euplóides, esta abordagem é a recomendada uma vez que a análise de profundidade de reads normaliza a somia cromossômica pela cobertura total do genoma. Entretanto, não se aplica para organismos que apresentam baixas taxas de variantes. Nas espécies de *L. braziliensis*, *L. donovani*, *L. infantum*, *L. major* e *L. mexicana*, que são exemplos de organismos com essa característica (DUMETZ *et al.*, 2017; ROGERS *et al.*, 2011), essa abordagem não se mostrou informativa tanto quando usamos CADIn quanto nQuire. Desta forma, a análise por profundidade de reads deve ser, neste caso, o método de escolha.

Análise de profundidade de reads é também muito informativa para detectar aumento ou diminuição de somia em comparação com a ploidia geral do genoma e é particularmente útil para identificar cromossomos monossômicos onde SNPs heterozigotos não são esperados. Genomas altamente repetitivos como *T. cruzi*, por outro lado, podem ter as estimativas somia cromossômica baseadas em profundidade de reads comprometidas. Nestes casos, a remoção de genes que se comportam como outliers se faz necessária. CADIn proporciona esta funcionalidade, excluindo automaticamente esses genes, o que aumenta a acurácia do método. Além disso, o

usuário pode excluir da análise sequências/regiões genômicas sabidamente repetitivas, o que também contribui para aumentar a confiabilidade das predições por profundidade de reads.

Em resumo, ao combinar as análises de frequência alélica e profundidade de reads, CADIn se mostrou capaz de avaliar variações do número de cópias cromossômicas com grande precisão em genomas com diferentes níveis de complexidade. O suporte estatístico do CADIn garante a confiabilidade de estimativas de ploidia e somia. A ferramenta é robusta, simples de instalar e usar, sendo executada em um único comando, e será útil para pesquisas que visam estudar variações de ploidia e somia em qualquer genoma sequenciado usando dados NGS.

REFERÊNCIAS

ALMEIDA, L. V.; COQUEIRO-DOS-SANTOS, A.; RODRIGUEZ-LUIZ, G. F.; MCCULLOCH, R.; BARTHOLOMEU, D. C.; REIS-CUNHA, J. L. Chromosomal copy number variation analysis by next generation sequencing confirms ploidy stability in Trypanosoma brucei subspecies. **Microbial genomics**, vol. 4, no. 10, Oct. 2018. DOI 10.1099/mgen.0.000223. Available at: http://dx.doi.org/10.1099/mgen.0.000223.

AMARASINGHE, S. L.; SU, S.; DONG, X.; ZAPPIA, L.; RITCHIE, M. E.; GOUIL, Q. Opportunities and challenges in long-read sequencing data analysis. **Genome biology**, vol. 21, no. 1, p. 30, 7 Feb. 2020. .

ANDREWS, S. FastQC: a quality control tool for high throughput sequence data. 2010. .

ASLETT, M.; AURRECOECHEA, C.; BERRIMAN, M.; BRESTELLI, J.; BRUNK, B. P.; CARRINGTON, M.; DEPLEDGE, D. P.; FISCHER, S.; GAJRIA, B.; GAO, X.; GARDNER, M. J.; GINGLE, A.; GRANT, G.; HARB, O. S.; HEIGES, M.; HERTZ-FOWLER, C.; HOUSTON, R.; INNAMORATO, F.; IODICE, J.; … WANG, H. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic acids research**, vol. 38, no. suppl_1, p. D457–D462, Jan. 2010. .

BAO, L.; PU, M.; MESSER, K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. **Bioinformatics** , vol. 30, no. 8, p. 1056–1063, 15 Apr. 2014. .

BAPTISTA, R. P.; REIS-CUNHA, J. L.; DEBARRY, J. D.; CHIARI, E.; KISSINGER, J. C.; BARTHOLOMEU, D. C.; MACEDO, A. M. Assembly of highly repetitive genomes using short reads: the genome of discrete typing unit III Trypanosoma cruzi strain 231. **Microbial genomics**, vol. 4, no. 4, Apr. 2018. DOI 10.1099/mgen.0.000156. Available at: http://dx.doi.org/10.1099/mgen.0.000156.

BARTHOLOMEU, D. C.; CERQUEIRA, G. C.; LEÃO, A. C. A.; DAROCHA, W. D.; PAIS, F. S.; MACEDO, C.; DJIKENG, A.; TEIXEIRA, S. M. R.; EL-SAYED, N. M. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen Trypanosoma cruzi. **Nucleic acids research**, vol. 37, no. 10, p. 3407–3417, Jun. 2009. .

BARTHOLOMEU, D. C.; TEIXEIRA, S. M. R.; CRUZ, A. K. Genomics and functional genomics in Leishmania and Trypanosoma cruzi: statuses, challenges and perspectives. **Memorias do Instituto Oswaldo Cruz**, vol. 116, p. e200634, 29 Mar. 2021. .

BERNÁ, L.; RODRIGUEZ, M.; CHIRIBAO, M. L.; PARODI-TALICE, A.; PITA, S.; RIJO, G.; ALVAREZ-VALIN, F.; ROBELLO, C. Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. **Microbial genomics**, vol. 4, no. 5, 1 May 2018a. DOI 10.1099/mgen.0.000177. Available at: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000177.

BERNÁ, L.; RODRIGUEZ, M.; CHIRIBAO, M. L.; PARODI-TALICE, A.; PITA, S.; RIJO, G.; ALVAREZ-VALIN, F.; ROBELLO, C. Expanding an expanded genome: long-read sequencing of

Trypanosoma cruzi. **Microbial genomics**, vol. 4, no. 5, 1 May 2018b. DOI 10.1099/mgen.0.000177. Available at: http://dx.doi.org/10.1099/mgen.0.000177.

BERRIMAN, M.; GHEDIN, E.; HERTZ-FOWLER, C.; BLANDIN, G.; RENAULD, H.; BARTHOLOMEU, D. C.; LENNARD, N. J.; CALER, E.; HAMLIN, N. E.; HAAS, B.; BÖHME, U.; HANNICK, L.; ASLETT, M. A.; SHALLOM, J.; MARCELLO, L.; HOU, L.; WICKSTEAD, B.; ALSMARK, U. C. M.; ARROWSMITH, C.; … EL-SAYED, N. M. The genome of the African trypanosome Trypanosoma brucei. **Science**, vol. 309, no. 5733, p. 416–422, 15 Jul. 2005. .

BLANCO, R.; RENGIFO, C. E.; CEDEÑO, M.; FRÓMETA, M.; RENGIFO, E. Flow Cytometric Measurement of Aneuploid DNA Content Correlates with High S-Phase Fraction and Poor Prognosis in Patients with Non-Small-Cell Lung Cancer. **International Scholarly Research Notices**, vol. 2013, 7 Aug. 2013. DOI 10.1155/2013/354123. Available at: https://www.hindawi.com/journals/isrn/2013/354123/. Accessed on: 13 Jan. 2023.

BOETZER, M.; HENKEL, C. V.; JANSEN, H. J.; BUTLER, D.; PIROVANO, W. Scaffolding pre-assembled contigs using SSPACE. **Bioinformatics** , vol. 27, no. 4, p. 578–579, 15 Feb. 2011. .

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics (Oxford, England)**, vol. 30, no. 15, p. 2114–2120, 1 Aug. 2014. .

BONTEMPI, I.; FLEITAS, P.; POATO, A.; VICCO, M.; RODELES, L.; PROCHETTO, E.; CABRERA, G.; BELUZZO, B.; ARIAS, D.; RACCA, A.; GUERRERO, S.; MARCIPAR, I. Trans-sialidase overcomes many antigens to be used as a vaccine candidate against *Trypanosoma cruzi*. **Immunotherapy**, vol. 9, no. 7, p. 555–565, 2017. DOI 10.2217/imt-2017-0009. Available at: http://dx.doi.org/10.2217/imt-2017-0009.

BRADWELL, K. R.; KOPARDE, V. N.; MATVEYEV, A. V.; SERRANO, M. G.; ALVES, J. M. P.; PARIKH, H.; HUANG, B.; LEE, V.; ESPINOSA-ALVAREZ, O.; ORTIZ, P. A.; COSTA-MARTINS, A. G.; TEIXEIRA, M. M. G.; BUCK, G. A. Genomic comparison of Trypanosoma conorhini and Trypanosoma rangeli to Trypanosoma cruzi strains of high and low virulence. **BMC genomics**, vol. 19, no. 1, p. 770, 24 Oct. 2018. .

BRINGAUD, F.; BITEAU, N.; MELVILLE, S. E.; HEZ, S.; EL-SAYED, N. M.; LEECH, V.; BERRIMAN, M.; HALL, N.; DONELSON, J. E.; BALTZ, T. A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of Trypanosoma brucei. **Eukaryotic cell**, vol. 1, no. 1, p. 137–151, Feb. 2002. .

BROAD INSTITUTE. GAEMR. [*s. d.*]. **GAEMR - Genome Assembly Evaluation, Metrics and Reporting**. Available at: https://software.broadinstitute.org/software/gaemr/. Accessed on: 10 Feb. 2023.

BROWN, C. L.; KEENUM, I. M.; DAI, D.; ZHANG, L.; VIKESLAND, P. J.; PRUDEN, A. Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. **Scientific reports**, vol. 11, no. 1, p. 3753, 12 Feb. 2021. .

BURGOS, J. M.; ALTCHEH, J.; BISIO, M.; DUFFY, T.; VALADARES, H. M. S.; SEIDENSTEIN, M. E.; PICCINALI, R.; FREITAS, J. M.; LEVIN, M. J.; MACCHI, L.; MACEDO, A. M.; FREILIJ,

H.; SCHIJMAN, A. G. Direct molecular profiling of minicircle signatures and lineages of Trypanosoma cruzi bloodstream populations causing congenital Chagas disease. **International journal for parasitology**, vol. 37, no. 12, p. 1319–1327, Oct. 2007. .

BURLE-CALDAS, G. de A.; DOS SANTOS, N. S. A.; DE CASTRO, J. T.; MUGGE, F. L. B.; GRAZIELLE-SILVA, V.; OLIVEIRA, A. E. R.; PEREIRA, M. C. A.; REIS-CUNHA, J. L.; DOS SANTOS, A. C.; GOMES, D. A.; BARTHOLOMEU, D. C.; MORETTI, N. S.; SCHENKMAN, S.; GAZZINELLI, R. T.; TEIXEIRA, S. M. R. Disruption of Active Trans-Sialidase Genes Impairs Egress from Mammalian Host Cells and Generates Highly Attenuated Trypanosoma cruzi Parasites. **mBio**, vol. 13, no. 1, p. e0347821, 25 Jan. 2022. .

BUSCAGLIA, C. A.; CAMPO, V. A.; FRASCH, A. C. C.; DI NOIA, J. M. Trypanosoma cruzi surface mucins: host-dependent coat diversity. **Nature reviews. Microbiology**, vol. 4, no. 3, p. 229–236, Mar. 2006. .

CALLEJAS-HERNÁNDEZ, F.; RASTROJO, A.; POVEDA, C.; GIRONÈS, N.; FRESNO, M. Genomic assemblies of newly sequenced Trypanosoma cruzi strains reveal new genomic expansion and greater complexity. **Scientific reports**, vol. 8, no. 1, p. 14631, 2 Oct. 2018. .

CHAKRABORTY, M.; BALDWIN-BROWN, J. G.; LONG, A. D.; EMERSON, J. J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. **Nucleic acids research**, vol. 44, no. 19, p. e147, 2 Nov. 2016. .

CHÁVEZ, S.; EASTMAN, G.; SMIRCICH, P.; BECCO, L. L.; OLIVEIRA-RIZZO, C.; FORT, R.; POTENZA, M.; GARAT, B.; SOTELO-SILVEIRA, J. R.; DUHAGON, M. A. Transcriptome-wide analysis of the Trypanosoma cruzi proliferative cycle identifies the periodically expressed mRNAs and their multiple levels of control. **PloS one**, vol. 12, no. 11, p. e0188441, 28 Nov. 2017. .

CHIN, C.-S.; ALEXANDER, D. H.; MARKS, P.; KLAMMER, A. A.; DRAKE, J.; HEINER, C.; CLUM, A.; COPELAND, A.; HUDDLESTON, J.; EICHLER, E. E.; TURNER, S. W.; KORLACH, J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. **Nature methods**, vol. 10, no. 6, p. 563–569, Jun. 2013. .

CHIN, C.-S.; PELUSO, P.; SEDLAZECK, F. J.; NATTESTAD, M.; CONCEPCION, G. T.; CLUM, A.; DUNN, C.; O'MALLEY, R.; FIGUEROA-BALDERAS, R.; MORALES-CRUZ, A.; CRAMER, G. R.; DELLEDONNE, M.; LUO, C.; ECKER, J. R.; CANTU, D.; RANK, D. R.; SCHATZ, M. C. Phased diploid genome assembly with single-molecule real-time sequencing. **Nature methods**, vol. 13, no. 12, p. 1050–1054, Dec. 2016. .

CHUNDURI, N. K.; STORCHOVÁ, Z. The diverse consequences of aneuploidy. **Nature cell biology**, vol. 21, no. 1, p. 54–62, Jan. 2019. .

COMAI, L. The advantages and disadvantages of being polyploid. **Nature reviews. Genetics**, vol. 6, no. 11, p. 836–846, Nov. 2005. .

CROW, K. D. What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? **Molecular Biology and Evolution**, vol. 23, no. 5, p. 887–892, 2006. DOI 10.1093/molbev/msj083. Available at: http://dx.doi.org/10.1093/molbev/msj083.

CRUZ-SAAVEDRA, L.; SCHWABL, P.; VALLEJO, G. A.; CARRANZA, J. C.; MUÑOZ, M.; PATINO, L. H.; PANIZ-MONDOLFI, A.; LLEWELLYN, M. S.; RAMÍREZ, J. D. Genome plasticity

driven by aneuploidy and loss of heterozygosity in Trypanosoma cruzi. **Microbial genomics**, vol. 8, no. 6, Jun. 2022. DOI 10.1099/mgen.0.000843. Available at: http://dx.doi.org/10.1099/mgen.0.000843.

DAYARIAN, A.; MICHAEL, T. P.; SENGUPTA, A. M. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. **BMC bioinformatics**, vol. 11, p. 345, 24 Jun. 2010. .

DE FREITAS, J. M.; AUGUSTO-PINTO, L.; PIMENTA, J. R.; BASTOS-RODRIGUES, L.; GONÇALVES, V. F.; TEIXEIRA, S. M. R.; CHIARI, E.; JUNQUEIRA, A. C. V.; FERNANDES, O.; MACEDO, A. M.; MACHADO, C. R.; PENA, S. D. J. Ancestral genomes, sex, and the population structure of Trypanosoma cruzi. **PLoS pathogens**, vol. 2, no. 3, p. e24, Mar. 2006. .

DE PABLOS, L. M.; OSUNA, A. Multigene families in Trypanosoma cruzi and their role in infectivity. **Infection and immunity**, vol. 80, no. 7, p. 2258–2264, Jul. 2012. .

DÍAZ-VIRAQUÉ, F.; PITA, S.; GREIF, G.; DE SOUZA, R. de C. M.; IRAOLA, G.; ROBELLO, C. Nanopore Sequencing Significantly Improves Genome Assembly of the Protozoan Parasite Trypanosoma cruzi. **Genome biology and evolution**, vol. 11, no. 7, p. 1952–1957, 1 Jul. 2019. .

DOLEZEL, J.; GREILHUBER, J.; SUDA, J. Estimation of nuclear DNA content in plants using flow cytometry. **Nature protocols**, vol. 2, no. 9, p. 2233–2244, 2007. .

DUMETZ, F.; IMAMURA, H.; SANDERS, M.; SEBLOVA, V.; MYSKOVA, J.; PESCHER, P.; VANAERSCHOT, M.; MEEHAN, C. J.; CUYPERS, B.; DE MUYLDER, G.; SPÄTH, G. F.; BUSSOTTI, G.; VERMEESCH, J. R.; BERRIMAN, M.; COTTON, J. A.; VOLF, P.; DUJARDIN, J. C.; DOMAGALSKA, M. A. Modulation of Aneuploidy in Leishmania donovani during Adaptation to Different In Vitro and In Vivo Environments and Its Impact on Gene Expression. **mBio**, vol. 8, no. 3, 23 May 2017. DOI 10.1128/mBio.00599-17. Available at: http://dx.doi.org/10.1128/mBio.00599-17.

EL-SAYED, N. M.; MYLER, P. J.; BARTHOLOMEU, D. C.; NILSSON, D.; AGGARWAL, G.; TRAN, A.-N.; GHEDIN, E.; WORTHEY, E. A.; DELCHER, A. L.; BLANDIN, G.; WESTENBERGER, S. J.; CALER, E.; CERQUEIRA, G. C.; BRANCHE, C.; HAAS, B.; ANUPAMA, A.; ARNER, E.; ASLUND, L.; ATTIPOE, P.; … ANDERSSON, B. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. **Science**, vol. 309, no. 5733, p. 409–415, 15 Jul. 2005. .

EL-SAYED, N. M.; MYLER, P. J.; BLANDIN, G.; BERRIMAN, M.; CRABTREE, J.; AGGARWAL, G.; CALER, E.; RENAULD, H.; WORTHEY, E. A.; HERTZ-FOWLER, C.; GHEDIN, E.; PEACOCK, C.; BARTHOLOMEU, D. C.; HAAS, B. J.; TRAN, A.-N.; WORTMAN, J. R.; ALSMARK, U. C. M.; ANGIUOLI, S.; ANUPAMA, A.; … HALL, N. Comparative genomics of trypanosomatid parasitic protozoa. **Science**, vol. 309, no. 5733, p. 404–409, 15 Jul. 2005. .

EMMS, D. M.; KELLY, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome biology**, vol. 16, no. 1, p. 157, 6 Aug. 2015. .

FEHRMANN, R. S. N.; KARJALAINEN, J. M.; KRAJEWSKA, M.; WESTRA, H.-J.; MALONEY, D.; SIMEONOV, A.; PERS, T. H.; HIRSCHHORN, J. N.; JANSEN, R. C.; SCHULTES, E. A.; VAN HAAGEN, H. H. H. B. M.; DE VRIES, E. G. E.; TE MEERMAN, G. J.; WIJMENGA, C.; VAN

VUGT, M. A. T. M.; FRANKE, L. Gene expression analysis identifies global gene dosage sensitivity in cancer. **Nature genetics**, vol. 47, no. 2, p. 115–125, Feb. 2015. .

FRANZÉN, O.; OCHAYA, S.; SHERWOOD, E.; LEWIS, M. D.; LLEWELLYN, M. S.; MILES, M. A.; ANDERSSON, B. Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1 and comparison with T. cruzi VI CL Brener. **PLoS neglected tropical diseases**, vol. 5, no. 3, p. e984, 8 Mar. 2011. .

FREEMAN, J. L.; PERRY, G. H.; FEUK, L.; REDON, R.; MCCARROLL, S. A.; ALTSHULER, D. M.; ABURATANI, H.; JONES, K. W.; TYLER-SMITH, C.; HURLES, M. E.; CARTER, N. P.; SCHERER, S. W.; LEE, C. Copy number variation: new insights in genome diversity. **Genome research**, vol. 16, no. 8, p. 949–961, Aug. 2006. .

FREITAS, L. M.; DOS SANTOS, S. L.; RODRIGUES-LUIZ, G. F.; MENDES, T. A. O.; RODRIGUES, T. S.; GAZZINELLI, R. T.; TEIXEIRA, S. M. R.; FUJIWARA, R. T.; BARTHOLOMEU, D. C. Genomic analyses, gene expression and antigenic profile of the trans-sialidase superfamily of Trypanosoma cruzi reveal an undetected level of complexity. **PloS one**, vol. 6, no. 10, p. e25914, 19 Oct. 2011. .

GALTIER, N. Gene conversion drives GC content evolution in mammalian histones. **Trends in genetics: TIG**, vol. 19, no. 2, p. 65–68, Feb. 2003. .

GAO, S.; SUNG, W.-K.; NAGARAJAN, N. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. **Journal of computational biology: a journal of computational molecular cell biology**, vol. 18, no. 11, p. 1681–1691, Nov. 2011. .

GHEDIN, E.; BRINGAUD, F.; PETERSON, J.; MYLER, P.; BERRIMAN, M.; IVENS, A.; ANDERSSON, B.; BONTEMPI, E.; EISEN, J.; ANGIUOLI, S.; WANLESS, D.; VON ARX, A.; MURPHY, L.; LENNARD, N.; SALZBERG, S.; ADAMS, M. D.; WHITE, O.; HALL, N.; STUART, K.; … EL-SAYED, N. M. A. Gene synteny and evolution of genome architecture in trypanosomatids. **Molecular and biochemical parasitology**, vol. 134, no. 2, p. 183–191, Apr. 2004. .

GHURYE, J.; POP, M.; KOREN, S.; BICKHART, D.; CHIN, C.-S. Scaffolding of long read assemblies using long range contact information. **BMC genomics**, vol. 18, no. 1, Dec. 2017. DOI 10.1186/s12864-017-3879-z. Available at: http://dx.doi.org/10.1186/s12864-017-3879-z.

GONZALEZ-GARCIA, L.; GUEVARA-BARRIENTOS, D.; LOZANO-ARCE, D.; GIL, J.; DÍAZ-RIAÑO, J.; DUARTE, E.; ANDRADE, G.; BOJACÁ, J. C.; HOYOS-SANCHEZ, M. C.; CHAVARRO, C.; GUAYAZAN, N.; CHICA, L. A.; BUITRAGO ACOSTA, M. C.; BAUTISTA, E.; TRUJILLO, M.; DUITAMA, J. New algorithms for accurate and efficient de novo genome assembly from long DNA sequencing reads. **Life science alliance**, vol. 6, no. 5, May 2023. DOI 10.26508/lsa.202201719. Available at: http://dx.doi.org/10.26508/lsa.202201719.

GUREVICH, A.; SAVELIEV, V.; VYAHHI, N.; TESLER, G. QUAST: quality assessment tool for genome assemblies. **Bioinformatics** , vol. 29, no. 8, p. 1072–1075, 15 Apr. 2013. .

HEDLEY, D. W.; FRIEDLANDER, M. L.; TAYLOR, I. W.; RUGG, C. A.; MUSGROVE, E. A. Method for analysis of cellular DNA content of paraffin-embedded pathological material using flow cytometry. **The journal of histochemistry and cytochemistry: official journal of the Histochemistry Society**, vol. 31, no. 11, p. 1333–1335, Nov. 1983. .

HERREROS-CABELLO, A.; CALLEJAS-HERNÁNDEZ, F.; GIRONÈS, N.; FRESNO, M. Trypanosoma Cruzi Genome: Organization, Multi-Gene Families, Transcription, and Biological Implications. **Genes**, vol. 11, no. 10, 14 Oct. 2020. DOI 10.3390/genes11101196. Available at: http://dx.doi.org/10.3390/genes11101196.

HORN, D.; BARRY, J. D. The central roles of telomeres and subtelomeres in antigenic variation in African trypanosomes. **Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology**, vol. 13, no. 5, p. 525–533, 2005. .

HUANG, W.; LI, L.; MYERS, J. R.; MARTH, G. T. ART: a next-generation sequencing read simulator. **Bioinformatics** , vol. 28, no. 4, p. 593–594, 15 Feb. 2012. .

HUNT, M.; NEWBOLD, C.; BERRIMAN, M.; OTTO, T. D. A comprehensive evaluation of assembly scaffolding tools. **Genome biology**, vol. 15, no. 3, p. R42, 3 Mar. 2014. .

HUSON, D. H.; REINERT, K.; MYERS, E. W. The greedy path-merging algorithm for contig scaffolding. **Journal of the ACM**, vol. 49, no. 5, p. 603–615, Sep. 2002. .

IANTORNO, S. A.; DURRANT, C.; KHAN, A.; SANDERS, M. J.; BEVERLEY, S. M.; WARREN, W. C.; BERRIMAN, M.; SACKS, D. L.; COTTON, J. A.; GRIGG, M. E. Gene Expression in Leishmania Is Regulated Predominantly by Gene Dosage. **mBio**, vol. 8, no. 5, 12 Sep. 2017. DOI 10.1128/mBio.01393-17. Available at: http://dx.doi.org/10.1128/mBio.01393-17.

ISMAIL, H. D. Basic local alignment search tool. **Bioinformatics**. New York: Chapman and Hall/CRC, 2022. p. 407–452.

IVENS, A. C.; PEACOCK, C. S.; WORTHEY, E. A.; MURPHY, L.; AGGARWAL, G.; BERRIMAN, M.; SISK, E.; RAJANDREAM, M.-A.; ADLEM, E.; AERT, R.; ANUPAMA, A.; APOSTOLOU, Z.; ATTIPOE, P.; BASON, N.; BAUSER, C.; BECK, A.; BEVERLEY, S. M.; BIANCHETTIN, G.; BORZYM, K.; … MYLER, P. J. The genome of the kinetoplastid parasite, Leishmania major. **Science**, vol. 309, no. 5733, p. 436–442, 15 Jul. 2005. .

KELLER, O.; KOLLMAR, M.; STANKE, M.; WAACK, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. **Bioinformatics (Oxford, England)**, vol. 27, no. 6, p. 757–763, 15 Mar. 2011. .

KIKTEV, D. A.; SHENG, Z.; LOBACHEV, K. S.; PETES, T. D. GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. **Proceedings of the National Academy of Sciences of the United States of America**, vol. 115, no. 30, p. E7109–E7118, 24 Jul. 2018. .

KIM, D.; CHIURILLO, M. A.; EL-SAYED, N.; JONES, K.; SANTOS, M. R. M.; PORCILE, P. E.; ANDERSSON, B.; MYLER, P.; DA SILVEIRA, J. F.; RAMÍREZ, J. L. Telomere and subtelomere of Trypanosoma cruzi chromosomes are enriched in (pseudo)genes of retrotransposon hot spot and trans-sialidase-like gene families: the origins of T. cruzi telomeres. **Gene**, vol. 346, p. 153–161, 14 Feb. 2005. .

KOLMOGOROV, M.; YUAN, J.; LIN, Y.; PEVZNER, P. A. Assembly of long, error-prone reads using repeat graphs. **Nature biotechnology**, vol. 37, no. 5, p. 540–546, May 2019. .

KOREN, S.; WALENZ, B. P.; BERLIN, K.; MILLER, J. R.; BERGMAN, N. H.; PHILLIPPY, A. M.

Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. **Genome research**, vol. 27, no. 5, p. 722–736, May 2017. .

KUDLA, G.; HELWAK, A.; LIPINSKI, L. Gene conversion and GC-content evolution in mammalian Hsp70. **Molecular biology and evolution**, vol. 21, no. 7, p. 1438–1444, Jul. 2004. .

KUDLA, G.; LIPINSKI, L.; CAFFIN, F.; HELWAK, A.; ZYLICZ, M. High guanine and cytosine content increases mRNA levels in mammalian cells. **PLoS biology**, vol. 4, no. 6, p. e180, Jun. 2006. .

KYRIAKIDOU, M.; TAI, H. H.; ANGLIN, N. L.; ELLIS, D.; STRÖMVIK, M. V. Current Strategies of Polyploid Plant Genome Sequence Assembly. **Frontiers in plant science**, vol. 9, p. 1660, 21 Nov. 2018. .

LANGMEAD, B.; TRAPNELL, C.; POP, M.; SALZBERG, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome biology**, vol. 10, no. 3, p. R25, 4 Mar. 2009. .

LASLETT, D.; CANBACK, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. **Nucleic acids research**, vol. 32, no. 1, p. 11–16, 2 Jan. 2004. .

LEINONEN, R.; SUGAWARA, H.; SHUMWAY, M.; COLLABORATION, I. N. S. D. The sequence read archive. **Nucleic acids research**, vol. 39, no. suppl_1, p. D19–D21, 2010. .

LIDANI, K. C. F.; ANDRADE, F. A.; BAVIA, L.; DAMASCENO, F. S.; BELTRAME, M. H.; MESSIAS-REASON, I. J.; SANDRI, T. L. Chagas Disease: From Discovery to a Worldwide Health Problem. **Frontiers in public health**, vol. 7, p. 166, 2 Jul. 2019. .

LI, H. auN: a new metric to measure assembly contiguity. Apr. 2020. **Heng Li's blog**. Available at: https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity. Accessed on: Jul. 2023.

LI, H. Minimap2: pairwise alignment for nucleotide sequences. **Bioinformatics** , vol. 34, no. 18, p. 3094–3100, 15 Sep. 2018. .

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics** , vol. 25, no. 14, p. 1754–1760, 15 Jul. 2009. .

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R.; 1000 GENOME PROJECT DATA PROCESSING SUBGROUP. The Sequence Alignment/Map format and SAMtools. **Bioinformatics** , vol. 25, no. 16, p. 2078–2079, 15 Aug. 2009. .

LI, L.; STOECKERT, C. J., Jr; ROOS, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. **Genome research**, vol. 13, no. 9, p. 2178–2189, Sep. 2003. .

LI, Z.; CHEN, Y.; MU, D.; YUAN, J.; SHI, Y.; ZHANG, H.; GAN, J.; LI, N.; HU, X.; LIU, B.; YANG, B.; FAN, W. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. **Briefings in functional genomics**, vol. 11, no. 1, p. 25–37, Jan. 2012. .

LUO, J.; LYU, M.; CHEN, R.; ZHANG, X.; LUO, H.; YAN, C. SLR: a scaffolding algorithm based

on long reads and contig classification. **BMC bioinformatics**, vol. 20, no. 1, p. 539, 30 Oct. 2019. .

MACÍAS, F.; AFONSO-LEHMANN, R.; LÓPEZ, M. C.; GÓMEZ, I.; THOMAS, M. C. Biology of Trypanosoma cruzi Retrotransposons: From an Enzymatic to a Structural Point of View. **Current genomics**, vol. 19, no. 2, p. 110–118, Feb. 2018. .

MAHMOUD, M.; ZYWICKI, M.; TWARDOWSKI, T.; KARLOWSKI, W. M. Efficiency of PacBio long read correction by 2nd generation Illumina sequencing. **Genomics**, vol. 111, no. 1, p. 43–49, Jan. 2019. .

MANNAERT, A.; DOWNING, T.; IMAMURA, H.; DUJARDIN, J.-C. Adaptive mechanisms in pathogens: universal aneuploidy in Leishmania. **Trends in parasitology**, vol. 28, no. 9, p. 370–376, Sep. 2012. .

MANNI, M.; BERKELEY, M. R.; SEPPEY, M.; SIMÃO, F. A.; ZDOBNOV, E. M. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. **Molecular biology and evolution**, vol. 38, no. 10, p. 4647–4654, 27 Sep. 2021. .

MARÇAIS, G.; DELCHER, A. L.; PHILLIPPY, A. M.; COSTON, R.; SALZBERG, S. L.; ZIMIN, A. MUMmer4: A fast and versatile genome alignment system. **PLoS computational biology**, vol. 14, no. 1, p. e1005944, Jan. 2018. .

MARGARIDO, G. R. A.; HECKERMAN, D. ConPADE: genome assembly ploidy estimation from next-generation sequencing data. **PLoS computational biology**, vol. 11, no. 4, p. e1004229, Apr. 2015. .

MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. **Genomics**, vol. 95, no. 6, p. 315–327, Jun. 2010. .

MILLER, J. R.; ZHOU, P.; MUDGE, J.; GURTOWSKI, J.; LEE, H.; RAMARAJ, T.; WALENZ, B. P.; LIU, J.; STUPAR, R. M.; DENNY, R.; SONG, L.; SINGH, N.; MARON, L. G.; MCCOUCH, S. R.; MCCOMBIE, W. R.; SCHATZ, M. C.; TIFFIN, P.; YOUNG, N. D.; SILVERSTEIN, K. A. T. Hybrid assembly with long and short reads improves discovery of gene family expansions. **BMC genomics**, vol. 18, no. 1, Dec. 2017. DOI 10.1186/s12864-017-3927-8. Available at: http://dx.doi.org/10.1186/s12864-017-3927-8.

MINNING, T. A.; WEATHERLY, D. B.; FLIBOTTE, S.; TARLETON, R. L. Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in Trypanosoma cruzi strains revealed by array comparative genomic hybridization. **BMC genomics**, vol. 12, no. 1, p. 139, 7 Mar. 2011. .

NADALIN, F.; VEZZI, F.; POLICRITI, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. **BMC bioinformatics**, vol. 13 Suppl 14, no. S14, p. S8, 7 Sep. 2012. .

NAGULESWARAN, A.; FERNANDES, P.; BEVKAL, S.; REHMANN, R.; NICHOLSON, P.; RODITI, I. Developmental changes and metabolic reprogramming during establishment of infection and progression of Trypanosoma brucei brucei through its insect host. **PLoS neglected tropical diseases**, vol. 15, no. 9, p. e0009504, Sep. 2021. .

NATH, S.; MALLICK, S. K.; JHA, S. An improved method of genome size estimation by flow

cytometry in five mucilaginous species of Hyacinthaceae. **Cytometry. Part A: the journal of the International Society for Analytical Cytology**, vol. 85, no. 10, p. 833–840, Oct. 2014. .

O'LEARY, N. A.; WRIGHT, M. W.; BRISTER, J. R.; CIUFO, S.; HADDAD, D.; MCVEIGH, R.; RAJPUT, B.; ROBBERTSE, B.; SMITH-WHITE, B.; AKO-ADJEI, D.; ASTASHYN, A.; BADRETDIN, A.; BAO, Y.; BLINKOVA, O.; BROVER, V.; CHETVERNIN, V.; CHOI, J.; COX, E.; ERMOLAEVA, O.; … PRUITT, K. D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. **Nucleic acids research**, vol. 44, no. D1, p. D733–45, 4 Jan. 2016. .

OLTMANN, J.; HESELMEYER-HADDAD, K.; HERNANDEZ, L. S.; MEYER, R.; TORRES, I.; HU, Y.; DOBERSTEIN, N.; KILLIAN, J. K.; PETERSEN, D.; ZHU, Y. J.; EDELMAN, D. C.; MELTZER, P. S.; SCHWARTZ, R.; GERTZ, E. M.; SCHÄFFER, A. A.; AUER, G.; HABERMANN, J. K.; RIED, T. Aneuploidy, TP53 mutation, and amplification of MYC correlate with increased intratumor heterogeneity and poor prognosis of breast cancer patients. **Genes, chromosomes & cancer**, vol. 57, no. 4, p. 165–175, Apr. 2018. .

OPPEZZO, P.; OBAL, G.; BARAIBAR, M. A.; PRITSCH, O.; ALZARI, P. M.; BUSCHIAZZO, A. Crystal structure of an enzymatically inactive trans-sialidase-like lectin from Trypanosoma cruzi: the carbohydrate binding mechanism involves residual sialidase activity. **Biochimica et biophysica acta**, vol. 1814, no. 9, p. 1154–1161, Sep. 2011. .

ORÓSTICA, K. Y.; VERDUGO, R. A. chromPlot: visualization of genomic data in chromosomal context. **Bioinformatics** , vol. 32, no. 15, p. 2366–2368, 1 Aug. 2016. .

ORR, B.; GODEK, K. M.; COMPTON, D. Aneuploidy. **Current biology: CB**, vol. 25, no. 13, p. R538–42, 29 Jun. 2015. .

OTTO, T. D. **IPA - Improve long read (Pacbio) Assemblies**. [*S. l.*: *s. n.*], 2017. Available at: https://github.com/ThomasDOtto/IPA.

PAGE, A. J.; DE SILVA, N.; HUNT, M.; QUAIL, M. A.; PARKHILL, J.; HARRIS, S. R.; OTTO, T. D.; KEANE, J. A. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. **Microbial genomics**, vol. 2, no. 8, p. e000083, Aug. 2016. .

PEARMAN, W. S.; FREED, N. E.; SILANDER, O. K. Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. **BMC bioinformatics**, vol. 21, no. 1, p. 220, 29 May 2020. .

PFAU, S. J.; AMON, A. A System to Study Aneuploidy In Vivo. **Cold Spring Harbor symposia on quantitative biology**, vol. 80, p. 93–101, 2015. .

POLLARD, M. O.; GURDASANI, D.; MENTZER, A. J.; PORTER, T.; SANDHU, M. S. Long reads: their purpose and place. **Human molecular genetics**, vol. 27, no. R2, p. R234–R241, 1 Aug. 2018. .

PRIETO BARJA, P.; PESCHER, P.; BUSSOTTI, G.; DUMETZ, F.; IMAMURA, H.; KEDRA, D.; DOMAGALSKA, M.; CHAUMEAU, V.; HIMMELBAUER, H.; PAGES, M.; STERKERS, Y.; DUJARDIN, J.-C.; NOTREDAME, C.; SPÄTH, G. F. Haplotype selection as an adaptive mechanism in the protozoan pathogen Leishmania donovani. **Nature ecology & evolution**, vol. 1, no. 12, p. 1961–1969, Dec. 2017. .

QIN, M.; WU, S.; LI, A.; ZHAO, F.; FENG, H.; DING, L.; RUAN, J. LRScaf: improving draft genomes using long noisy reads. **BMC genomics**, vol. 20, no. 1, p. 955, 9 Dec. 2019. .

QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics** , vol. 26, no. 6, p. 841–842, 15 Mar. 2010. .

R CORE TEAM. **R: A language and environment for statistical computing**. [*S. l.*: *s. n.*], 2022. Available at: https://www.r-project.org/.

REIS-CUNHA, J. L.; BAPTISTA, R. P.; RODRIGUES-LUIZ, G. F.; COQUEIRO-DOS-SANTOS, A.; VALDIVIA, H. O.; DE ALMEIDA, L. V.; CARDOSO, M. S.; D'ÁVILA, D. A.; DIAS, F. H. C.; FUJIWARA, R. T.; GALVÃO, L. M. C.; CHIARI, E.; CERQUEIRA, G. C.; BARTHOLOMEU, D. C. Whole genome sequencing of Trypanosoma cruzi field isolates reveals extensive genomic variability and complex aneuploidy patterns within TcII DTU. **BMC genomics**, vol. 19, no. 1, p. 816, 13 Nov. 2018. .

REIS-CUNHA, J. L.; BARTHOLOMEU, D. C. Trypanosoma cruzi Genome Assemblies: Challenges and Milestones of Assembling a Highly Repetitive and Complex Genome. **Methods in molecular biology** , vol. 1955, p. 1–22, 2019. .

REIS-CUNHA, J. L.; COQUEIRO-DOS-SANTOS, A.; PIMENTA-CARVALHO, S. A.; MARQUES, L. P.; RODRIGUES-LUIZ, G. F.; BAPTISTA, R. P.; ALMEIDA, L. V. de; HONORATO, N. R. M.; LOBO, F. P.; FRAGA, V. G.; GALVÃO, L. M. da C.; BUENO, L. L.; FUJIWARA, R. T.; CARDOSO, M. S.; CERQUEIRA, G. C.; BARTHOLOMEU, D. C. Accessing the Variability of Multicopy Genes in Complex Genomes using Unassembled Next-Generation Sequencing Reads: The Case of Trypanosoma cruzi Multigene Families. **mBio**, vol. 13, no. 6, p. e0231922, 20 Oct. 2022. .

REIS-CUNHA, J. L.; RODRIGUES-LUIZ, G. F.; VALDIVIA, H. O.; BAPTISTA, R. P.; MENDES, T. A. O.; DE MORAIS, G. L.; GUEDES, R.; MACEDO, A. M.; BERN, C.; GILMAN, R. H.; LOPEZ, C. T.; ANDERSSON, B.; VASCONCELOS, A. T.; BARTHOLOMEU, D. C. Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct Trypanosoma cruzi strains. **BMC genomics**, vol. 16, no. 1, p. 499, 4 Jul. 2015. .

RHOADS, A.; AU, K. F. PacBio Sequencing and Its Applications. **Genomics, proteomics & bioinformatics**, vol. 13, no. 5, p. 278–289, Oct. 2015. .

ROBERTS, R. J.; CARNEIRO, M. O.; SCHATZ, M. C. The advantages of SMRT sequencing. **Genome biology**, vol. 14, no. 7, p. 405, 3 Jul. 2013. .

ROBINSON, P.; ZEMO JTEL, T. Integrative genomics viewer (IGV): Visualizing alignments and variants. **Computational Exome and Genome Analysis**. [*S. l.*]: Chapman and Hall/CRC, 2017. p. 233–245.

ROGERS, M. B.; HILLEY, J. D.; DICKENS, N. J.; WILKES, J.; BATES, P. A.; DEPLEDGE, D. P.; HARRIS, D.; HER, Y.; HERZYK, P.; IMAMURA, H.; OTTO, T. D.; SANDERS, M.; SEEGER, K.; DUJARDIN, J.-C.; BERRIMAN, M.; SMITH, D. F.; HERTZ-FOWLER, C.; MOTTRAM, J. C. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. **Genome research**, vol. 21, no. 12, p. 2129–2142, Dec. 2011. .

SANTOS, R. A. C. dos; DOS SANTOS, R. A. C.; GOLDMAN, G. H.; RIAÑO-PACHÓN, D. M. ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. **Bioinformatics**,

vol. 33, no. 16, p. 2575–2576, 2017. DOI 10.1093/bioinformatics/btx204. Available at: http://dx.doi.org/10.1093/bioinformatics/btx204.

SCHENKMAN, S.; EICHINGER, D.; PEREIRA, M. E.; NUSSENZWEIG, V. Structural and functional properties of Trypanosoma trans-sialidase. **Annual review of microbiology**, vol. 48, p. 499–523, 1994. .

SIMÃO, F. A.; WATERHOUSE, R. M.; IOANNIDIS, P.; KRIVENTSEVA, E. V.; ZDOBNOV, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics** , vol. 31, no. 19, p. 3210–3212, 1 Oct. 2015. .

STERKERS, Y.; LACHAUD, L.; CROBU, L.; BASTIEN, P.; PAGÈS, M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. **Cellular microbiology**, vol. 13, no. 2, p. 274–283, Feb. 2011. .

STROPE, P. K.; SKELLY, D. A.; KOZMIN, S. G.; MAHADEVAN, G.; STONE, E. A.; MAGWENE, P. M.; DIETRICH, F. S.; MCCUSKER, J. H. The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. **Genome research**, vol. 25, no. 5, p. 762–774, May 2015. .

SZPEITER, B. B.; FERREIRA, J. I. G. da S.; ASSIS, F. F. V. de; STELMACHTCHUK, F. N.; PEIXOTO, K. da C., Junior; AJZENBERG, D.; MINERVINO, A. H. H.; GENNARI, S. M.; MARCILI, A. Bat trypanosomes from Tapajós-Arapiuns Extractive Reserve in Brazilian Amazon. **Revista brasileira de parasitologia veterinaria = Brazilian journal of veterinary parasitology: Orgao Oficial do Colegio Brasileiro de Parasitologia Veterinaria**, vol. 26, no. 2, p. 152–158, Apr-Jun 2017. .

TALAVERA-LÓPEZ, C.; MESSENGER, L. A.; LEWIS, M. D.; YEO, M.; REIS-CUNHA, J. L.; MATOS, G. M.; BARTHOLOMEU, D. C.; CALZADA, J. E.; SALDAÑA, A.; RAMÍREZ, J. D.; GUHL, F.; OCAÑA-MAYORGA, S.; COSTALES, J. A.; GORCHAKOV, R.; JONES, K.; NOLAN, M. S.; TEIXEIRA, S. M. R.; CARRASCO, H. J.; BOTTAZZI, M. E.; … ANDERSSON, B. Repeat-Driven Generation of Antigenic Diversity in a Major Human Pathogen, Trypanosoma cruzi. **Frontiers in cellular and infection microbiology**, vol. 11, p. 614665, 3 Mar. 2021. .

THAM, C.-Y.; POON, L.; YAN, T.; KOH, J. Y. P.; RAMLEE, M. K.; TEOH, V. S. I.; ZHANG, S.; CAI, Y.; HONG, Z.; LEE, G. S.; LIU, J.; SONG, H. W.; HWANG, W. Y. K.; TEH, B. T.; TAN, P.; XU, L.; KOH, A. S.; OSATO, M.; LI, S. High-throughput telomere length measurement at nucleotide resolution using the PacBio high fidelity sequencing platform. **Nature communications**, vol. 14, no. 1, p. 281, 17 Jan. 2023. .

THOMAS, M. C.; MACIAS, F.; ALONSO, C.; LÓPEZ, M. C. The biology and evolution of transposable elements in parasites. **Trends in parasitology**, vol. 26, no. 7, p. 350–362, Jul. 2010. .

TORRES, E. M.; DEPHOURE, N.; PANNEERSELVAM, A.; TUCKER, C. M.; WHITTAKER, C. A.; GYGI, S. P.; DUNHAM, M. J.; AMON, A. Identification of aneuploidy-tolerating mutations. **Cell**, vol. 143, no. 1, p. 71–83, 1 Oct. 2010. .

TØRRESEN, O. K.; STAR, B.; MIER, P.; ANDRADE-NAVARRO, M. A.; BATEMAN, A.; JARNOT, P.; GRUCA, A.; GRYNBERG, M.; KAJAVA, A. V.; PROMPONAS, V. J.; ANISIMOVA, M.; JAKOBSEN, K. S.; LINKE, D. Tandem repeats lead to sequence assembly errors and impose

multi-level challenges for genome and protein databases. **Nucleic acids research**, vol. 47, no. 21, p. 10994–11006, 2 Dec. 2019. .

TSAI, I. J.; OTTO, T. D.; BERRIMAN, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. **Genome biology**, vol. 11, no. 4, p. R41, 13 Apr. 2010. .

VAN DIJK, E. L.; AUGER, H.; JASZCZYSZYN, Y.; THERMES, C. Ten years of next-generation sequencing technology. **Trends in genetics: TIG**, vol. 30, no. 9, p. 418–426, 1 Sep. 2014. .

VAN LOO, P.; NORDGARD, S. H.; LINGJÆRDE, O. C.; RUSSNES, H. G.; RYE, I. H.; SUN, W.; WEIGMAN, V. J.; MARYNEN, P.; ZETTERBERG, A.; NAUME, B.; PEROU, C. M.; BØRRESEN-DALE, A.-L.; KRISTENSEN, V. N. Allele-specific copy number analysis of tumors. **Proceedings of the National Academy of Sciences of the United States of America**, vol. 107, no. 39, p. 16910–16915, 28 Sep. 2010. .

WALKER, B. J.; ABEEL, T.; SHEA, T.; PRIEST, M.; ABOUELLIEL, A.; SAKTHIKUMAR, S.; CUOMO, C. A.; ZENG, Q.; WORTMAN, J.; YOUNG, S. K.; EARL, A. M. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. **PIoS one**, vol. 9, no. 11, p. e112963, 19 Nov. 2014. .

WANG, A.; WANG, Z.; LI, Z.; LI, L. M. BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. **Bioinformatics** , vol. 34, no. 12, p. 2019–2028, 15 Jun. 2018. .

WANG, W.; PENG, D.; BAPTISTA, R. P.; LI, Y.; KISSINGER, J. C.; TARLETON, R. L. Strain-specific genome evolution in Trypanosoma cruzi, the agent of Chagas disease. **PLoS pathogens**, vol. 17, no. 1, p. e1009254, Jan. 2021a. .

WANG, W.; PENG, D.; BAPTISTA, R. P.; LI, Y.; KISSINGER, J. C.; TARLETON, R. L. Strain-specific genome evolution in Trypanosoma cruzi, the agent of Chagas disease. **PLoS pathogens**, vol. 17, no. 1, p. e1009254, Jan. 2021b. .

WATSON, M.; WARR, A. **Errors in long-read assemblies can critically affect protein prediction**. **Nature biotechnology**. [*S. l.*: *s. n.*], Feb. 2019.

WEATHERLY, D. B.; BOEHLKE, C.; TARLETON, R. L. Chromosome level assembly of the hybrid Trypanosoma cruzi genome. **BMC genomics**, vol. 10, p. 255, 1 Jun. 2009. .

WEIS, C. L.; PAIS, M.; CANO, L. M.; KAMOUN, S.; BURBANO, H. A. nQuire: a statistical framework for ploidy estimation using next generation sequencing. **BMC bioinformatics**, vol. 19, no. 1, p. 122, 4 Apr. 2018. .

WICKHAM, H. Ggplot2. **Wiley interdisciplinary reviews. Computational statistics**, vol. 3, no. 2, p. 180–185, Mar. 2011. .

WICKHAM, H.; WICKHAM, M. H. Package tidyverse. **the 'Tidyverse**, [*s. d.*]. .

WORLD HEALTH ORGANIZATION. Chagas disease (also known as American trypanosomiasis). 13 Apr. 2022. Available at: https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis). Accessed on: 2022.

YU, Z.; LIU, Y.; SHEN, Y.; WANG, M.; LI, A. CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. **Bioinformatics** , vol. 30, no. 18, p. 2576–2583, 15 Sep. 2014. .

ZHANG, H.; JAIN, C.; ALURU, S. A comprehensive evaluation of long read error correction methods. **BMC genomics**, vol. 21, no. Suppl 6, p. 889, 21 Dec. 2020. .

ZHU, Y. O.; SHERLOCK, G.; PETROV, D. A. Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation. **G3** , vol. 6, no. 8, p. 2421–2434, 9 Aug. 2016. .

ZIMIN, A. V.; SALZBERG, S. L. The SAMBA tool uses long reads to improve the contiguity of genome assemblies. **PLoS computational biology**, vol. 18, no. 2, p. e1009860, Feb. 2022. .

ZINGALES, B. Trypanosoma cruzi genetic diversity: Something new for something known about Chagas disease manifestations, serodiagnosis and drug sensitivity. **Acta tropica**, vol. 184, p. 38–52, Aug. 2018. .

ZINGALES, B.; MILES, M. A.; CAMPBELL, D. A.; TIBAYRENC, M.; MACEDO, A. M.; TEIXEIRA, M. M. G.; SCHIJMAN, A. G.; LLEWELLYN, M. S.; LAGES-SILVA, E.; MACHADO, C. R.; ANDRADE, S. G.; STURM, N. R. The revised Trypanosoma cruzi subspecific nomenclature: rationale, epidemiological relevance and research applications. **Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases**, vol. 12, no. 2, p. 240–253, Mar. 2012. .

ANEXOS

ANEXO I: Relação de artigos científicos publicados e patente depositada durante o período do doutorado não relacionados à tese.

1. Goes WM, Brasil CRF, Reis-Cunha JL, Coqueiro-Dos-Santos A, Grazielle-Silva V, de Souza Reis J, Souto TC, Laranjeira-Silva MF, Bartholomeu DC, Fernandes AP, Teixeira SMR. Complete assembly, annotation of virulence genes and CRISPR editing of the genome of Leishmania amazonensis PH8 strain. Genomics. 2023 May 30;115(5):110661.

2. Coqueiro-Dos-Santos A, Bento GA, Barral A, de Oliveira CI, Bartholomeu DC. Draft Genome Sequence of the Protozoan Parasite Leishmania braziliensis Strain BA788, Isolated from a Clinical Case in Bahia State, Brazil. Microbiol Resour Announc. 2022 Dec 15;11(12):e0024522.

3. Reis-Cunha JL, Coqueiro-Dos-Santos A, Pimenta-Carvalho SA, Marques LP, Rodrigues-Luiz GF, Baptista RP, Almeida LV, Honorato NRM, Lobo FP, Fraga VG, Galvão LMDC, Bueno LL, Fujiwara RT, Cardoso MS, Cerqueira GC, Bartholomeu DC. Accessing the Variability of Multicopy Genes in Complex Genomes using Unassembled Next-Generation Sequencing Reads: The Case of Trypanosoma cruzi Multigene Families. mBio. 2022 Dec 20;13(6):e0231922.

4. Valdivia HO, Roatt BM, Baptista RP, Ottino J, Coqueiro-Dos-Santos A, Sanders MJ, Reis AB, Cotton JA, Bartholomeu DC. Replacement of Leishmania (Leishmania) infantum Populations in an Endemic Focus of Visceral Leishmaniasis in Brazil. Front Cell Infect Microbiol. 2022 Jun 24;12:900084.

5. Leão AC, Viana LA, Fortes de Araujo F, de Lourdes Almeida R, Freitas LM, CoqueiroDos-Santos A, da Silveira-Lemos D, Cardoso MS, Reis-Cunha JL, Teixeira-Carvalho A, Bartholomeu DC. Antigenic diversity of MASP gene family of Trypanosoma cruzi. Microbes Infect. 2022 Sep;24(6-7):104982.

6. Burle-Caldas GA, Dos Santos NSA, de Castro JT, Mugge FLB, Grazielle-Silva V, Oliveira AER, Pereira MCA, Reis-Cunha JL, Dos Santos AC, Gomes DA, Bartholomeu DC, Moretti NS, Schenkman S, Gazzinelli RT, Teixeira SMR. Disruption of Active

TransSialidase Genes Impairs Egress from Mammalian Host Cells and Generates Highly Attenuated Trypanosoma cruzi Parasites. mBio. 2022 Feb 22;13(1):e0347821.

7. Gazzinelli-Guimarães AC, Nogueira DS, Amorim CCO, Oliveira FMS, Coqueiro-DosSantos A, Carvalho SAP, Kraemer L, Barbosa FS, Fraga VG, Santos FV, de Castro JC, Russo RC, Akamatsu MA, Ho PL, Bottazzi ME, Hotez PJ, Zhan B, Bartholomeu DC, Bueno LL, Fujiwara RT. ASCVac-1, a Multi-Peptide Chimeric Vaccine, Protects Mice Against Ascaris suum Infection. Front Immunol. 2021 Dec 21;12:788185.

8. Viana de Almeida L, Luís Reis-Cunha J, Coqueiro-Dos-Santos A, Flávia Rodrigues-Luís G, de Paula Baptista R, de Oliveira Silva S, Norma de Melo M, Castanheira Bartholomeu 96 D. Comparative genomics of Leishmania isolates from Brazil confirms the presence of Leishmania major in the Americas. Int J Parasitol. 2021 Nov;51(12):1047-1057.

9. Fantin RF, Fraga VG, Lopes CA, de Azevedo IC, Reis-Cunha JL, Pereira DB, Lobo FP, de Oliveira MM, Dos Santos AC, Bartholomeu DC, Fujiwara RT, Bueno LL. New highly antigenic linear B cell epitope peptides from PvAMA-1 as potential vaccine candidates. PLoS One. 2021 Nov 2;16(11):e0258637

10. Silva TED, Barbosa FS, Magalhães LMD, Gazzinelli-Guimarães PH, Dos Santos AC, Nogueira DS, Resende NM, Amorim CC, Gazzinelli-Guimarães AC, Viana AG, Geiger SM, Bartholomeu DC, Fujiwara RT, Bueno LL. Unraveling Ascaris suum experimental infection in humans. Microbes Infect. 2021 Sep-Oct;23(8):104836.

PATENTE

1. Recombinant chimeric protein, immunogenic composition against ascaridiasis and uses. BR102020026982.

ANEXO II: Paper submetido referente ao capítulo 2 da tese de doutorado

CADIn - Chromosomal Amplification and Deletion Inference tool

Anderson Coqueiro-dos-Santos[1], Samuel Alexandre Pimenta-Carvalho[1], Gabriela Flávia Rodrigues-Luiz[2], João Luís Reis-Cunha[3], Daniella C. Bartholomeu[1]

[1] Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais (UFMG), Brazil; [2] Experimental Medicine Research Cluster (EMRC), University of Campinas (UNICAMP), Campinas, Brazil; [3] Department of Biology, University of York, United Kingdom.

Highlights

- CADIn is a free, easy-to-use, automated tool that requires only three files as input.
- It estimates genome ploidy and chromosome somy using both variant allele frequency and read depth coverage analyses.
- It generates publication-level figures, removes chromosomal regions with outlier coverages, and validates ploidy/somy variations statistically.
- It can detect aneuploidies in genomes with a low SNP count, such as those of *Leishmania* spp.

Abstract

Recent studies indicate that ploidy variations may facilitate adaptation to specific environmental conditions by modulating gene expression levels through changes in gene dosage, generation of diversity and selection of beneficial haplotypes. Next-Generation Sequencing (NGS) data has been increasingly used to estimate ploidy. Here, we present CADIn, a free, easy-to-use, and automated tool that uses both allele frequencies of heterozygous single nucleotide polymorphisms (SNPs) and read-depth coverage (RDC) analyses to detect genome ploidy and chromosomal somy variations based on NGS data with a single command. CADIn generates publication-quality figures, removes chromosomal regions with outlier coverages, and validates ploidy variations statistically. By applying CADIn to simulated and real NGS data from *Saccharomyces cerevisiae* and *Leishmania* sp., the latter of which is known to have a low frequency of heterozygous

SNPs, we successfully detected polyploidy and aneuploidies. Our findings highlight the importance of using independent evidence when estimating ploidy. CADIn is available at github.com/coqueiro-dos-santos/CADIn.

Keywords

Ploidy estimation; aneuploidy, chromosome copy number variation; Next Generation Sequencing; read depth coverage; allele frequency

1. Introduction

In order to survive, some organisms can temporarily alter their chromosome dosage, a process known as chromosome copy number variation (CCNV) [1,2], which can affect the entire set of chromosomes (polyploidy) or only a few (aneuploidy) [3]. While aneuploidy is usually deleterious for most multicellular organisms, leading to imbalanced gene expression levels [3,4], it can also function as an evolutionary mechanism for unicellular eukaryotes [1,5], facilitating the increase and accumulation of mutations that contribute to genetic diversity despite the instability caused by the higher copy number [6,7]. Variations in chromosome copy number may also generate phenotypic variability by modulating transcript abundance through changing gene dosage [8]. In addition, chromosome duplication followed by the loss of another copy of the same chromosome can allow the elimination of deleterious alleles and therefore the selection of beneficial haplotypes, a process known as haplotype selection [9]. These processes emphasize the role of aneuploidy as a driver for evolutionary adaptation and survival under selective pressure in different environments [10–12].

Ploidy levels can be estimated using a variety of methodologies, each with their own advantages and disadvantages. Flow cytometry techniques are extensively employed, particularly in biomedical and plant studies [13–16]. However, its implementation is hindered by the need to isolate sufficient nuclear material without causing damage, the presence of debris coatings and the low detection rate by fluorescence imaging [14,17]. The use of Next-Generation Sequencing (NGS) data for ploidy measurement is an alternative method that is more practical due to the ease and

low cost of sequencing, as well as the availability of numerous publicly accessible read libraries in databases [18].

Among the available NGS-based strategies for ploidy estimation, variant calling and read depth (RD) assessment are widely used. Variant calling analyzes heterozygous single nucleotide polymorphisms (SNPs) by quantifying the frequency of point variants in a given sample [19–21], enabling the differentiation of disomic (0.5, 0.5), trisomic (0.33, 0.66), tetrasomic (0.25, 0.75) and other chromosome states. Alternatively, the average read depth of a chromosome relative to the average read depth of the entire genome, can be used for estimate ploidy. However, each methodology exhibits inherent limitations when used independently: variant calling proves challenging for organisms with low variation rates [22], while estimating ploidy based on read depth becomes difficult for highly repetitive genomes. Additionally, the lack of statistical evaluation, particularly for read depth, renders the obtained results unsupported.

A number of tools have been developed to estimate CCNV based on NGS data. ASCAT, AbsCN-seq, and CLImAT have been applied for cancer sample analysis [23–25], with CLImAT being a fee-based tool, requiring MATLAB license. ConPADE, which is better suited for polyploid genomes, such as those of plants, requires the generation of contigs and is sensitive to mapping quality [26]. ploidyNGS [19] and nQuire [21] are specifically optimized for organisms with smaller genomes such as yeast, with the former limited to allele frequency analysis and the latter requiring the input of multiple command-lines during the analysis.

Here, we present CADIn (Chromosomal Amplification and Deletion Inference tool), a free, easy-to-use, and automated tool that estimates genome ploidy and aneuploidies with a single command. CADIn employs a combination of allele frequency analysis of heterozygous SNPs and RD variations. This integrated methodology improves the accuracy of ploidy predictions and permits the estimation of CCNVs in genomes with low heterozygosity. CADIn generates publication-quality images, removes chromosomal regions with outlier coverages, enhancing the accuracy of CCNV estimations based on the RD approach, and validates ploidy variations statistically.

2. Materials and methods

2.1. Software Description

CADIn is a pipeline developed for evaluating ploidy and somy through variant calling and read depth coverage (RDC) analysis. It offers a single execution command via the terminal and requires only three categories of input data: a) one or more BAM files containing mapped NGS reads; b) the reference genome file used for read mapping in FASTA format; and c) an annotation file in GFF3 format. The BAM file can be generated by mapping reads to the reference genome or transcriptome using tools such as BWA [27] or Bowtie [28]. Prior to using CADIn, there is no need to index, filter, or sort any data, as these processes will be handled when required. However, after mapping, users can filter the reference file to analyze only regions or chromosomes of interest. Initial screening for mapping quality (-q) is undertaken with a default value of 30. GFF3 files must be filtered by feature type (gene, exon, CDS, mRNA, etc.), with “gene” set as the default option; however, this can be readily modified using the -f flag. It is crucial that the chromosome ID remains consistent between annotation and reference files, and conversion from GFF3 to BED file format will be conducted automatically when required.

Variant calling analysis is one method utilized by CADIn to estimate ploidy. This technique compares the mapped reads to the reference genome in search of single-nucleotide variants. Considered are only heterozygous SNPs with two (and only two) variants. All variants must have a phred quality score higher than 10 and at least five supporting reads, although these parameters are adjustable using the -k and -d flags, respectively. The proportion of reads that support each variant is computed, grouped by chromosome, and depicted on a frequency graph. The graph’s distribution pattern is used to infer somy. A bell-shaped curve distribution, with the maximum number of variants centered around 0.5, is indicative of disomy, since each variant accounts for 50% of the total count. Similarly, distribution peaks around 0.33 and/or 0.66 indicate a trisomic state, while peaks around 0.25, 0.5 and/or 0.75 point to a tetrasomic state, and so on. If the user is not interested in performing variant calling, the -v flag followed by 0 can be used to suppress it (1 is the default option, enabling the analysis).

The read depth count of annotated regions serves as the foundation for the other methodology implemented by CADIn to assess ploidy. By computing each gene position, CADIn estimates the gene's depth by calculating the median or mean value (users can choose using the -m flag) of the number of mapped reads across all positions. Genes with coverage below 50% will be discarded, and this threshold can be adjusted using the -l flag. The RDC of the entire genome is used to normalize the depth of the remaining genes. RDC can be calculated based on either the depth of all genes or the total depth of the mapped genome, as specified by the -p flag. After normalization, CADIn iteratively applies the Grubbs test to each chromosome to eliminate outlier regions. This step enables the removal of highly repetitive genomic regions, such as those rich in multigene families, which could otherwise artificially inflate the RDC count and hamper somy estimation. CADIn then generates boxplots and heatmaps of RDC values, which must be compared due to observed differences and variations among chromosomes within the same sample. The baseline value of 1 represents the sample ploidy. Deviations from this baseline indicate ploidy fluctuations and potential aneuploidies. To statistically assess the significance of these differences (p-value < 0.05), CADIn applies the Mann-Whitney-Wilcoxon rank test, testing for scores below 0.5 and 1, as well as scores above 1, 1.5, 2, 2.5, and 3. All chromosomes in the sample are subjected to pairwise comparisons, with the Bonferroni correction applied to facilitate an asymptotic test. The outcomes of these comparisons are depicted on a heatmap, with p-values below 0.01, 0.05, and 0.10 highlighted.

2.2. Datasets used for validation

To evaluate CADIn's performance, both simulated and real datasets were used. Synthetic paired-end NGS genomic reads from the S288C strain of *Saccharomyces cerevisiae* were generated using the ART software [29]. The simulated dataset included distinct NGS coverages per chromosome. The genome coverage of the libraries for chromosomes 2, 3, 4, and 5 was 75x, 100x, 125x, and 150x, respectively, while the coverage for the remaining chromosomes was 50x. These simulations were performed using Illumina HiSeq 1000 (100 bp), HiSeq 2000 (100 bp), HiSeq 2500 (125 bp and 150 bp), and HiSeqX TruSeq (150 bp), to evaluate the performance of various platforms and

read lengths. Real data validations were conducted using five clinical samples of *S. cerevisiae* and two *Leishmania* species (*Leishmania infantum* and *Leishmania major*). Read libraries were downloaded from the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA) [30] (Table S1), quality-checked using FastQC (version 0.11.8) [31] and trimmed using Trimmomatic (version 0.33) [32]. Genome (FASTA format) and annotation (GFF format) files for *S. cerevisiae* S288C (GCA_000146045.2), *L. major* (Friedlin strain) and *L. infantum* (JPCM5 strain) were obtained from NCBI and TriTrypDB (release 46) [33]. All read libraries were then mapped to their respective reference genomes using BWA (version 0.7.12) [27], and the resulting BAM files, along with the reference genome and annotation data for each species, were used as input for CADIn, which was executed with the default parameters. Mitochondrial reference data were excluded before ploidy estimations. Finally, CADIn's performance was compared to that of nQuire using the same dataset. For this specific analysis, the mapped reads were sorted and filtered with a mapping quality threshold of 30. In addition, BED-formatted annotation files containing the coordinates for each chromosome were generated.

2.3. Implementation

All analyses were conducted on a laptop with 8 GB of RAM, 256 GB of disk space, and an Intel i5 processor with 2 cores. The analyses were performed using a single processing core and were not parallelized. CADIn was tested on Debian- and Red Hat-based Linux systems, as well as on MacOS. The time taken from the submission of each dataset to the receipt of results was measured (Table S6).

3. Results

3.1. Simulated data

CADIn's capability to detect ploidy variations was initially assessed using simulated *S. cerevisiae* genomic read data. For this purpose, read depth coverages for chromosomes 2, 3, 4, and 5 were artificially increased to 75x, 100x, 125x, and 150x, respectively (see methods). As a result, the expected somy variations were observed for these chromosomes, whereas no somy alterations were observed for the remaining

chromosomes (Fig. 1). Moreover, there was no observed bias in the somy values when changing the read length or sequencing platform (Fig. 1, Fig. S1, Fig. S2, Table S2).

3.2 Real datasets

CADIn was also validated using real genomic reads from *S. cerevisiae* and *Leishmania* sp. Initially, the overall genome ploidy of *S. cerevisiae* isolates was evaluated using allelic frequency analysis, considering only the heterozygous SNPs from all chromosomes. As shown in Fig. 2, *S. cerevisiae* isolates presented different genome ploidies. CBS7837 and YJM1098 were diploid, while CBS2919 and YJM466 were triploid, and CBS9564 was tetraploid.

Next, using the same allele frequency approach, the somy of individual chromosomes was evaluated for each isolate. Variations in somy were observed for isolates CBS2919 (chromosome 1), CBS9564 (chromosomes 9 and 13), YJM1098 (chromosome 12), and YJM466 (chromosomes 6 and 9), whereas CBS7837 exhibited no aneuploidy (Fig. 3).

RDC analysis was conducted on the same *S. cerevisiae* dataset to confirm the allelic frequency results. Somy estimations by the RDC methodology were similar to what was observed using the variant calling approach (Fig. 4). In addition, the RDC method exhibited results with statistically significant differences in both individual (Table S3) and pairwise (Fig. S3) comparisons.

To further evaluate CADIn's performance, the genome ploidy and chromosomal somy of *Leishmania infantum* and *Leishmania major* samples were assessed. These parasites are known for their low frequency of heterozygous SNPs [22]. Despite the larger genome sizes (~30 Mb for *Leishmania* and ~12 Mb for *S. cerevisiae*), the substantially lower number of variants in *Leishmania* isolates compared to *S. cerevisiae* isolates presented challenges for ploidy estimation based on allelic frequency analysis. Thus, only a tendency towards a diploid pattern was observed for both *L. infantum* and *L. major* (Fig. S4). The reduced number of variants per chromosome compared to the entire genome further hindered somy estimations (Fig. 5). These findings highlight the limitations of

relying solely on allelic frequency analysis for ploidy and somy estimations in genomes with low variation rates, an approach employed by other existing pipelines [17,23,24].

Due to the impracticality of using the frequency of heterozygous SNPs to assess chromosome somy variations in *L. infantum* and *L. major* isolates, RDC analysis was conducted instead. As depicted in Fig. 6 and Table S4, CADIn successfully estimated somy alterations in *Leishmania* sp., highlighting the importance of employing diverse approaches for somy estimation. In both species, chromosome 31 exhibited supernumerary status, a pattern already documented in the literature [5]. Additionally, *L. infantum* isolates displayed increased copy numbers for chromosomes 6, 8, 9, 17, 22, 25, 33, and 35 (Fig. 6). Notably, pairwise analyses revealed statistically significant somy differences in *Leishmania* samples (Fig. S5).

Next, we compared *Leishmania* sp. somy predictions obtained by CADIn and nQuire. nQuire models the distribution of base frequencies at variable sites using a Gaussian Mixture Model and selects the most plausible ploidy model using maximum likelihood. To do so, nQuire calculates the distances (delta) between each one of three fixed models (diploid, triploid and tetraploid) and the free model. The smallest delta value of a given model indicates the best statistical support and, consequently, the estimated chromosome somy. Based on the observed delta values, all *Leishmania* chromosomes were classified as tetrasomic (Fig. S6 and Table S5), a pattern not found in any other study [5,8,22] and quite different from CADIn somy predictions based on RDC (Fig 5, Table S4). These results reinforce the notion that somy and ploidy estimations on genomes with low frequency of variable sites must be based on RDC analysis when using NGS data.

CADIn's runtime is greatly influenced by the number of reads in the BAM file and, to a lesser extent, by genome size and gene counts. The runtime details for the different samples used in this study can be found in Table S6.

4. Discussion

While aneuploidy may be deleterious in multicellular organisms, causing, for instance, cancer and Down syndrome in humans [34], recently it has been associated with adaptations to stress conditions in unicellular eukaryotes, such as drug resistance [8,12]. As CCNV can be generated in a single round of cell duplication, it is a rapid mechanism to modulate expression levels through changing gene dosage [8,35]. In addition, it is an important source of genetic variation that contributes to the generation of phenotypic diversity [6,7]. Consequently, the development of simple methods to estimate genome ploidy and chromosome somies is a crucial step towards a better understanding of the biological significance of this phenomenon in a broad variety of organisms. Here, we introduce CADIn, a free and automated tool that estimates CCNVs with a single command using NGS data.

We evaluate CADIn performance using *S. cerevisiae* and *Leishmania* sp., organisms in which ploidy and somy variations are well documented [5–8,22]. In a study comprising 144 different *S. cerevisiae* strains [36], Zhu and colleagues demonstrated that over 40 percent of the isolates exhibited overall ploidy variations, while more than one-third of the samples were aneuploid. This extensive genomic plasticity has been previously associated with the ability of *S. cerevisiae* to act as an opportunistic pathogen in humans and contribute to a variety of clinical outcomes. Using both heterozygous SNPs and RDC methods, CADIn has proven to be highly effective in detecting ploidy and somy variations in this organism. Our findings in *S. cerevisiae* are consistent with those reported by Zhu et al. [36], and for samples S01, CBS7837, and CBS9564, our results align with those obtained using another tool [21]. However, CADIn surpasses other pipelines due to its statistical support, which increases confidence in the results.

Some ploidy and somy estimation pipelines employ allelic frequency analyses as their method of choice [19,21,25]. However, for organisms with a low variation rate, such as species belonging to the *Leishmania* genus [8,22], it is difficult to obtain reliable results using only variant calling analysis, as demonstrated by our comparisons with nQuire. Similarly, for organisms with a highly repetitive genome such as the protozoan parasite

Trypanosoma cruzi, which is phylogenetically related to *Leishmania*, variant calling exhibits several limitations in ploidy and somy estimations [37]. In these cases, accounting for RDC values of each gene, followed by their normalization with genome coverage, enables for the identification and subsequent removal of potential outliers [38]. This strategy has already been used in studies with *Trypanosoma brucei* and *T. cruzi*, where the absence of aneuploidies in the former and their presence in the latter were observed [37,39], proving to be a valuable alternative for such organisms.

It is important to emphasize that, for euploid genomes, the allelic frequency approach is recommended to estimate genome ploidy, as the RDC analysis normalizes chromosome somy by the genome coverage. However, RDC analysis is suitable for detecting increases or decreases in somy compared to the overall genome ploidy and is particularly useful for identifying monosomic chromosomes where heterozygous SNPs are not expected. Additionally, CADIn's functionality, which allows the automatic exclusion of genes with outlier coverage and/or the removal of known repetitive genomic sequences/regions by the user, enhances the accuracy of somy estimations based on RDC analysis. This feature is especially important when assessing ploidy/somy in highly repetitive genomes. In short, by combining allele frequency and RDC analyses, CADIn accurately estimated genome ploidy and identified aneuploidies in organisms with different genome complexities. At last, CADIn's statistical support ensures the reliability of ploidy and somy estimations. The tool is robust and simple to install and use, making it a valuable resource for researchers investigating ploidy variations in any sequenced genome using NGS data.

Funding

This work was supported by the Brazilian Higher Education Personnel Improvement Coordination (CAPES) Biologia Computational [grant number 1759/2014], and the Brazilian National Council for Scientific and Technological Development (CNPq) [grant number 305816/2019-5]. DCB is a CNPq research fellow. ACS received a scholarship from CNPq. SAPC received a scholarship from the Research Support Foundation of the State of Minas Gerais (FAPEMIG).

Declaration of Competing Interest

The authors declare there is no conflict of interests.

References

[1] L. Comai, The advantages and disadvantages of being polyploid, Nat. Rev. Genet. 6 (2005) 836–846.

[2] J.L. Freeman, G.H. Perry, L. Feuk, R. Redon, S.A. McCarroll, D.M. Altshuler, H. Aburatani, K.W. Jones, C. Tyler-Smith, M.E. Hurles, N.P. Carter, S.W. Scherer, C. Lee, Copy number variation: new insights in genome diversity, Genome Res. 16 (2006) 949–961.

[3] N.K. Chunduri, Z. Storchová, The diverse consequences of aneuploidy, Nat. Cell Biol. 21 (2019) 54–62.

[4] J.M. Sheltzer, E.M. Torres, M.J. Dunham, A. Amon, Transcriptional consequences of aneuploidy, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 12644–12649.

[5] S.A. Iantorno, C. Durrant, A. Khan, M.J. Sanders, S.M. Beverley, W.C. Warren, M. Berriman, D.L. Sacks, J.A. Cotton, M.E. Grigg, Gene Expression in Leishmania Is Regulated Predominantly by Gene Dosage, MBio. 8 (2017). https://doi.org/10.1128/mBio.01393-17.

[6] A.M. Selmecki, Y.E. Maruvka, P.A. Richmond, M. Guillet, N. Shoresh, A.L. Sorenson, S. De, R. Kishony, F. Michor, R. Dowell, D. Pellman, Polyploidy can drive rapid adaptation in yeast, Nature. 519 (2015) 349–352.

[7] S. Venkataram, B. Dunn, Y. Li, A. Agarwala, J. Chang, E.R. Ebel, K. Geiler-Samerotte, L. Hérissant, J.R. Blundell, S.F. Levy, D.S. Fisher, G. Sherlock, D.A. Petrov, Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast, Cell. 166 (2016) 1585–1596.e22.

[8] F. Dumetz, H. Imamura, M. Sanders, V. Seblova, J. Myskova, P. Pescher, M. Vanaerschot, C.J. Meehan, B. Cuypers, G. De Muylder, G.F. Späth, G. Bussotti, J.R. Vermeesch, M. Berriman, J.A. Cotton, P. Volf, J.C. Dujardin, M.A. Domagalska, Modulation of Aneuploidy in Leishmania donovani during Adaptation to Different In Vitro and In Vivo Environments and Its Impact on Gene Expression, MBio. 8 (2017). https://doi.org/10.1128/mBio.00599-17.

[9] P. Prieto Barja, P. Pescher, G. Bussotti, F. Dumetz, H. Imamura, D. Kedra, M. Domagalska, V. Chaumeau, H. Himmelbauer, M. Pages, Y. Sterkers, J.-C. Dujardin, C. Notredame, G.F. Späth, Haplotype selection as an adaptive mechanism in the protozoan pathogen Leishmania donovani, Nat Ecol Evol. 1 (2017) 1961–1969.

[10] G. Simonetti, S. Bruno, A. Padella, E. Tenti, G. Martinelli, Aneuploidy: Cancer strength or vulnerability?, Int. J. Cancer. 144 (2019) 8–25.

[11] R.T. Todd, A. Selmecki, Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs, Elife. 9 (2020). https://doi.org/10.7554/eLife.58349.

[12] A.H. Yona, Y.S. Manor, R.H. Herbst, G.H. Romano, A. Mitchell, M. Kupiec, Y. Pilpel, O. Dahan, Chromosomal duplication is a transient evolutionary solution to stress, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 21010–21015.

[13] R. Blanco, C.E. Rengifo, M. Cedeño, M. Frómeta, E. Rengifo, Flow Cytometric Measurement of Aneuploid DNA Content Correlates with High S-Phase Fraction and Poor Prognosis in Patients with Non-Small-Cell Lung Cancer, International Scholarly Research Notices. 2013 (2013). https://doi.org/10.1155/2013/354123.

[14] J. Dolezel, J. Greilhuber, J. Suda, Estimation of nuclear DNA content in plants using flow cytometry, Nat. Protoc. 2 (2007) 2233–2244.

[15] D.W. Hedley, M.L. Friedlander, I.W. Taylor, C.A. Rugg, E.A. Musgrove, Method for analysis

of cellular DNA content of paraffin-embedded pathological material using flow cytometry, J. Histochem. Cytochem. 31 (1983) 1333–1335.
[16] S.J. Pfau, A. Amon, A System to Study Aneuploidy In Vivo, Cold Spring Harb. Symp. Quant. Biol. 80 (2015) 93–101.
[17] S. Nath, S.K. Mallick, S. Jha, An improved method of genome size estimation by flow cytometry in five mucilaginous species of Hyacinthaceae, Cytometry A. 85 (2014) 833–840.
[18] E.L. van Dijk, H. Auger, Y. Jaszczyszyn, C. Thermes, Ten years of next-generation sequencing technology, Trends Genet. 30 (2014) 418–426.
[19] R.A.C. dos Santos, R.A.C. dos Santos, G.H. Goldman, D.M. Riaño-Pachón, ploidyNGS: visually exploring ploidy with Next Generation Sequencing data, Bioinformatics. 33 (2017) 2575–2576. https://doi.org/10.1093/bioinformatics/btx204.
[20] A.M. Gross, S.S. Ajay, V. Rajan, C. Brown, K. Bluske, N.J. Burns, A. Chawla, A.J. Coffey, A. Malhotra, A. Scocchia, E. Thorpe, N. Dzidic, K. Hovanes, T. Sahoo, E. Dolzhenko, B. Lajoie, A. Khouzam, S. Chowdhury, J. Belmont, E. Roller, S. Ivakhno, S. Tanner, J. McEachern, T. Hambuch, M. Eberle, R.T. Hagelstrom, D.R. Bentley, D.L. Perry, R.J. Taft, Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease, Genet. Med. 21 (2019) 1121–1130.
[21] C.L. Weiß, M. Pais, L.M. Cano, S. Kamoun, H.A. Burbano, nQuire: a statistical framework for ploidy estimation using next generation sequencing, BMC Bioinformatics. 19 (2018) 122.
[22] M.B. Rogers, J.D. Hilley, N.J. Dickens, J. Wilkes, P.A. Bates, D.P. Depledge, D. Harris, Y. Her, P. Herzyk, H. Imamura, T.D. Otto, M. Sanders, K. Seeger, J.-C. Dujardin, M. Berriman, D.F. Smith, C. Hertz-Fowler, J.C. Mottram, Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania, Genome Res. 21 (2011) 2129–2142.
[23] L. Bao, M. Pu, K. Messer, AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data, Bioinformatics. 30 (2014) 1056–1063.
[24] P. Van Loo, S.H. Nordgard, O.C. Lingjærde, H.G. Russnes, I.H. Rye, W. Sun, V.J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C.M. Perou, A.-L. Børresen-Dale, V.N. Kristensen, Allele-specific copy number analysis of tumors, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 16910–16915.
[25] Z. Yu, Y. Liu, Y. Shen, M. Wang, A. Li, CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data, Bioinformatics. 30 (2014) 2576–2583.
[26] G.R.A. Margarido, D. Heckerman, ConPADE: genome assembly ploidy estimation from next-generation sequencing data, PLoS Comput. Biol. 11 (2015) e1004229.
[27] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics. 25 (2009) 1754–1760.
[28] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
[29] W. Huang, L. Li, J.R. Myers, G.T. Marth, ART: a next-generation sequencing read simulator, Bioinformatics. 28 (2012) 593–594.
[30] R. Leinonen, H. Sugawara, M. Shumway, I.N.S.D. Collaboration, The sequence read archive, Nucleic Acids Res. 39 (2010) D19–D21.
[31] S. Andrews, FastQC: a quality control tool for high throughput sequence data, (2010).
[32] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics. 30 (2014) 2114–2120.
[33] M. Aslett, C. Aurrecoechea, M. Berriman, J. Brestelli, B.P. Brunk, M. Carrington, D.P. Depledge, S. Fischer, B. Gajria, X. Gao, M.J. Gardner, A. Gingle, G. Grant, O.S. Harb, M. Heiges, C. Hertz-Fowler, R. Houston, F. Innamorato, J. Iodice, J.C. Kissinger, E. Kraemer, W. Li, F.J. Logan, J.A. Miller, S. Mitra, P.J. Myler, V. Nayak, C. Pennington, I. Phan, D.F.

Pinney, G. Ramasamy, M.B. Rogers, D.S. Roos, C. Ross, D. Sivam, D.F. Smith, G. Srinivasamoorthy, C.J. Stoeckert Jr, S. Subramanian, R. Thibodeau, A. Tivey, C. Treatman, G. Velarde, H. Wang, TriTrypDB: a functional genomic resource for the Trypanosomatidae, Nucleic Acids Res. 38 (2010) D457–D462.
[34] J. Oltmann, K. Heselmeyer-Haddad, L.S. Hernandez, R. Meyer, I. Torres, Y. Hu, N. Doberstein, J.K. Killian, D. Petersen, Y.J. Zhu, D.C. Edelman, P.S. Meltzer, R. Schwartz, E.M. Gertz, A.A. Schäffer, G. Auer, J.K. Habermann, T. Ried, Aneuploidy, TP53 mutation, and amplification of MYC correlate with increased intratumor heterogeneity and poor prognosis of breast cancer patients, Genes Chromosomes Cancer. 57 (2018) 165–175.
[35] R.S.N. Fehrmann, J.M. Karjalainen, M. Krajewska, H.-J. Westra, D. Maloney, A. Simeonov, T.H. Pers, J.N. Hirschhorn, R.C. Jansen, E.A. Schultes, H.H.H.B.M. van Haagen, E.G.E. de Vries, G.J. te Meerman, C. Wijmenga, M.A.T.M. van Vugt, L. Franke, Gene expression analysis identifies global gene dosage sensitivity in cancer, Nat. Genet. 47 (2015) 115–125.
[36] Y.O. Zhu, G. Sherlock, D.A. Petrov, Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation, G3 . 6 (2016) 2421–2434.
[37] J.L. Reis-Cunha, R.P. Baptista, G.F. Rodrigues-Luiz, A. Coqueiro-Dos-Santos, H.O. Valdivia, L.V. de Almeida, M.S. Cardoso, D.A. D'Ávila, F.H.C. Dias, R.T. Fujiwara, L.M.C. Galvão, E. Chiari, G.C. Cerqueira, D.C. Bartholomeu, Whole genome sequencing of Trypanosoma cruzi field isolates reveals extensive genomic variability and complex aneuploidy patterns within TcII DTU, BMC Genomics. 19 (2018) 816.
[38] T.J. Treangen, S.L. Salzberg, Repetitive DNA and next-generation sequencing: computational challenges and solutions, Nat. Rev. Genet. 13 (2011) 36–46.
[39] L.V. Almeida, A. Coqueiro-Dos-Santos, G.F. Rodriguez-Luiz, R. McCulloch, D.C. Bartholomeu, J.L. Reis-Cunha, Chromosomal copy number variation analysis by next generation sequencing confirms ploidy stability in Trypanosoma brucei subspecies, Microb Genom. 4 (2018). https://doi.org/10.1099/mgen.0.000223.

Figure legends

Figure 1. Estimated somy for simulated *S. cerevisiae* samples using RDC analysis. The genome coverage of the libraries for chromosomes 2, 3, 4, and 5 was 75x, 100x, 125x, and 150x, respectively, while the coverage for the remaining chromosomes was set to 50x. A) Each boxplot represents a chromosome (identification number on the x-axis), with somy values depicted on the y-axis. The red line indicates the baseline chromosome copy number per haploid genome. Boxplots were generated using normalized read depth values for all genes on a chromosome. B) Heatmap depicting somy estimations, with each row representing a chromosome and each column representing a distinct sample. The median value of the normalized read depth was utilized for all genes on each chromosome. The color scale indicates somy fluctuations relative to the baseline (white represents 1). HS10, HiSeq 1000; HS20, HiSeq 2000; HS25-1, HiSeq 2500 with 125bp; HS25-2, HiSeq 2500 with 125bp; HSXt, HiSeqX TruSeq.

Figure 2. CADIn ploidy estimations of different *S. cerevisiae* samples based on the allele frequency of heterozygous SNPs across all chromosomes. The x-axis represents the allele frequency ratio of heterozygous positions from 0 (0 %) to 1 (100 %), while the y-axis indicates the allele frequency count in the genome.

Figure 3. Somy estimations of each chromosome in different *S. cerevisiae* samples obtained by allele frequency analysis of heterozygous SNPs. The x-axis represents the allele frequency ratio of heterozygous positions from 0 (0 %) to 1 (100 %), while the y-axis indicates the allele frequency count in the chromosome.

Figure 4. Somy estimations of each chromosome in different *S. cerevisiae* samples using RDC analysis. Each boxplot represents a chromosome (identification number on the x-axis), with somy values depicted on the y-axis. The red line indicates the baseline chromosome copy number per haploid genome (value of 1). Boxplots were generated using normalized read depth values for all genes on a chromosome.

Figure 5. Somy estimations in *L. infantum* and *L. major* isolates using the allelic frequency approach. The x-axis represents the allele frequency ratio of heterozygous positions from 0 (0 %) to 1 (100 %), while the y-axis indicates the allele frequency count in the chromosome. Each box represents a chromosome, and those without heterozygous SNPs are not represented.

Figure 6. Somy estimations of each chromosome in LinJ *L. infantum* and LmjF *L. major* using RDC analysis. Each boxplot represents a chromosome, with somy values depicted on the y-axis. The red line indicates the baseline chromosome copy number per haploid genome (value of 1). Boxplots were generated using normalized read depth values for all genes on a chromosome.

Supplementary data

Figure S1. Chromosome somy estimations in *S. cerevisiae* based on RDC analysis. *S. cerevisiae* simulated data for HiSeq 2500 with 125bp (HS25-1) e 150bp (HS25-2) length reads. The genome coverage of the libraries for chromosomes 2, 3, 4, and 5 was 75x,

100x, 125x, and 150x, respectively, while the coverage for the remaining chromosomes was set to 50x. Each boxplot represents a chromosome (identification number on the x-axis), with somy values depicted on the y-axis. The red line indicates the baseline chromosome copy number per haploid genome. Boxplots were generated using normalized read depth values for all genes on a chromosome.

Figure S2. Heatmap of the significance values obtained through the pairwise Wilcox test, evaluated between chromosomes. *S. cerevisiae* simulated data for HiSeq 1000 (HS10), HiSeq 2000 (HS20), HiSeq 2500 with 125bp (HS25-1) e 150bp (HS25-2) length reads, and HiSeqX TruSeq (HSXt). The genome coverage of the libraries for chromosomes 2, 3, 4, and 5 was 75x, 100x, 125x, and 150x, respectively, while the coverage for the remaining chromosomes was set to 50x. In gray, chromosomal somies with no statistical difference; and in blue, chromosomal somies statistically different.

Figure S3. Heatmap of the significance values obtained through the pairwise Wilcox test for *S. cerevisiae* samples, evaluated between chromosomes of the same sample. In gray, chromosomal somies with no statistical difference; and in blue, chromosomal somies statistically different.

Figure S4. Genome ploidy estimations of LinJ *Leishmania infantum* and LmjF *Leishmania major.* isolates based on allele frequency analysis of heterozygous SNPs. The x-axis represents the allele frequency ratio of heterozygous positions from 0 (0 %) to 1 (100 %), while the y-axis indicates the allele frequency count in the chromosome.

Figure S5. Heatmap of the significance values obtained through the pairwise Wilcox test for *Leishmania sp.* samples, evaluated between chromosomes of the same species. In gray, chromosomal somies with no statistical difference; and in blue, chromosomal somies statistically different.

Figure S6. nQuire results represented in a graphic with delta values for *Leishmania infantum* sample. Those values were obtained after Gaussian Mixture Model analysis (y-axis), the values correspond to the diploidy, triploidy, and tetraploidy test (x-axis). Among the three assessments, the one with the lowest values will have the better confidence to be accurate.

Figure1

Click here to access/download;Figure;Fig1.tif

Layer Group

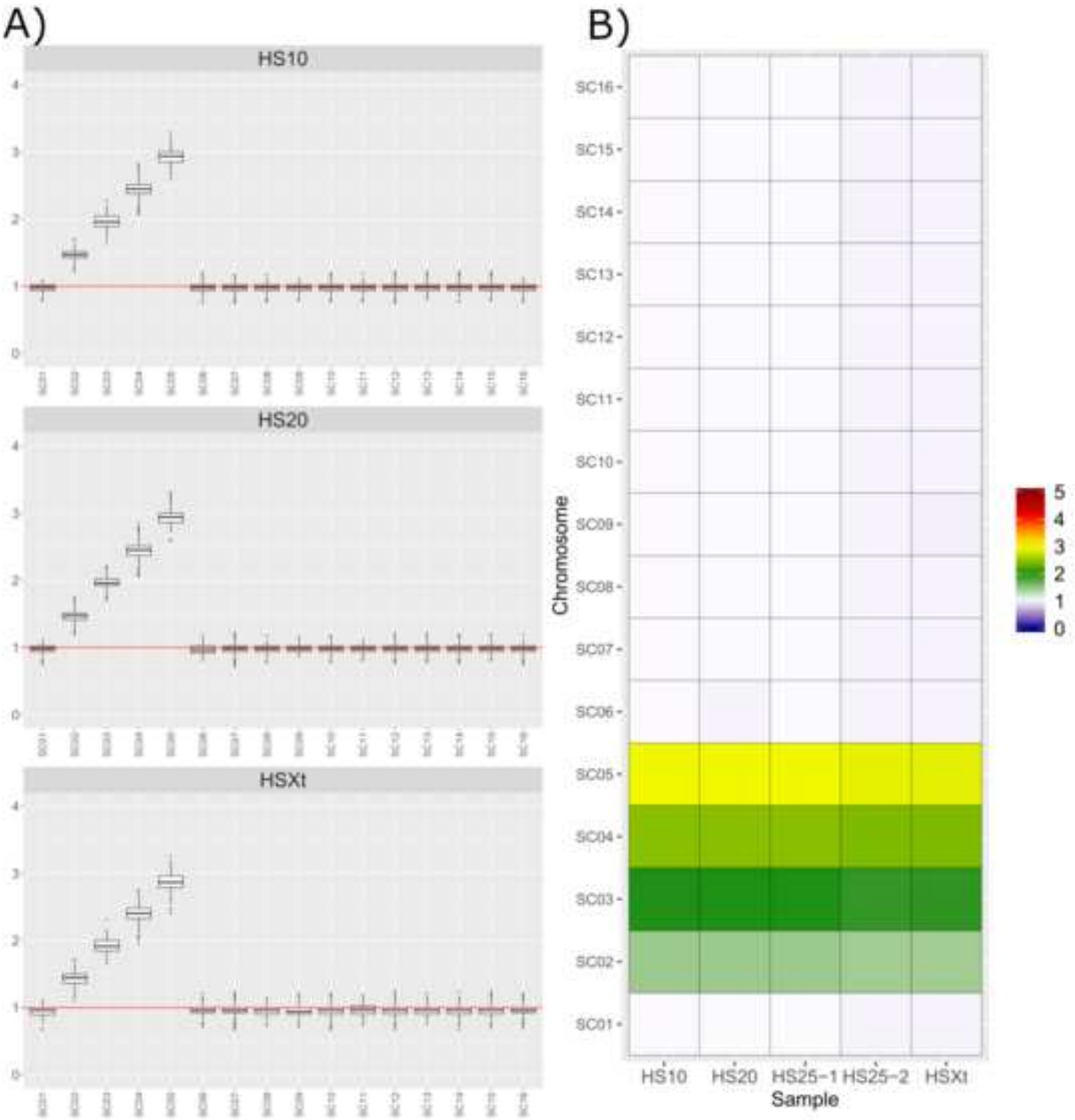
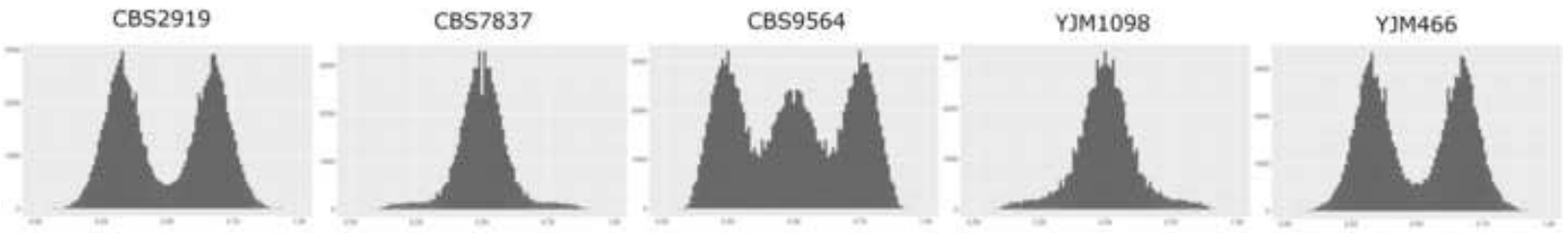


Figure2

Click here to access/download;Figure;Fig2.tif



Layer #1

Figure3

Click here to access/download;Figure;Fig3.tif

-

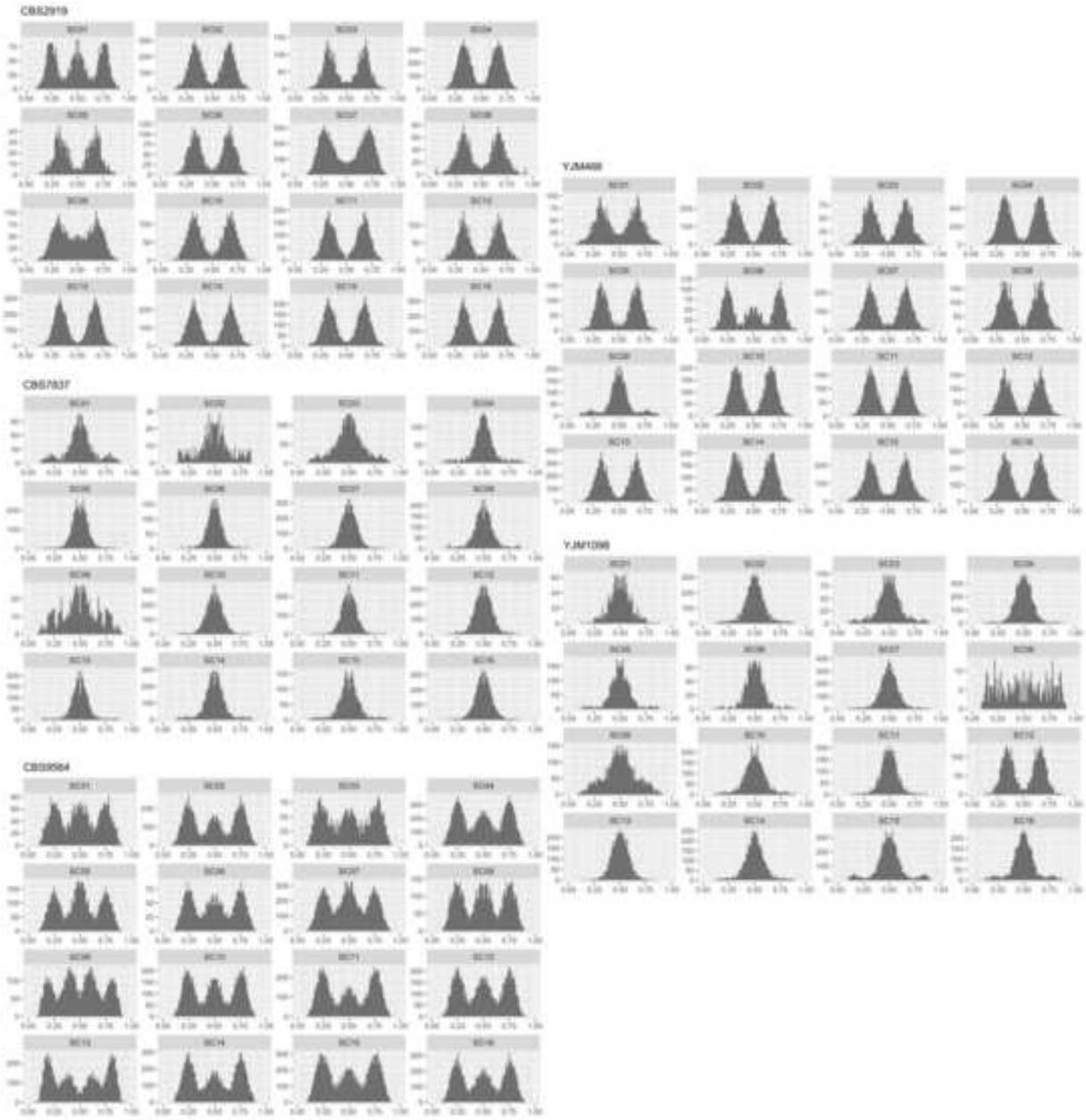


Figure4

Click here to access/download;Figure;Fig4.tif

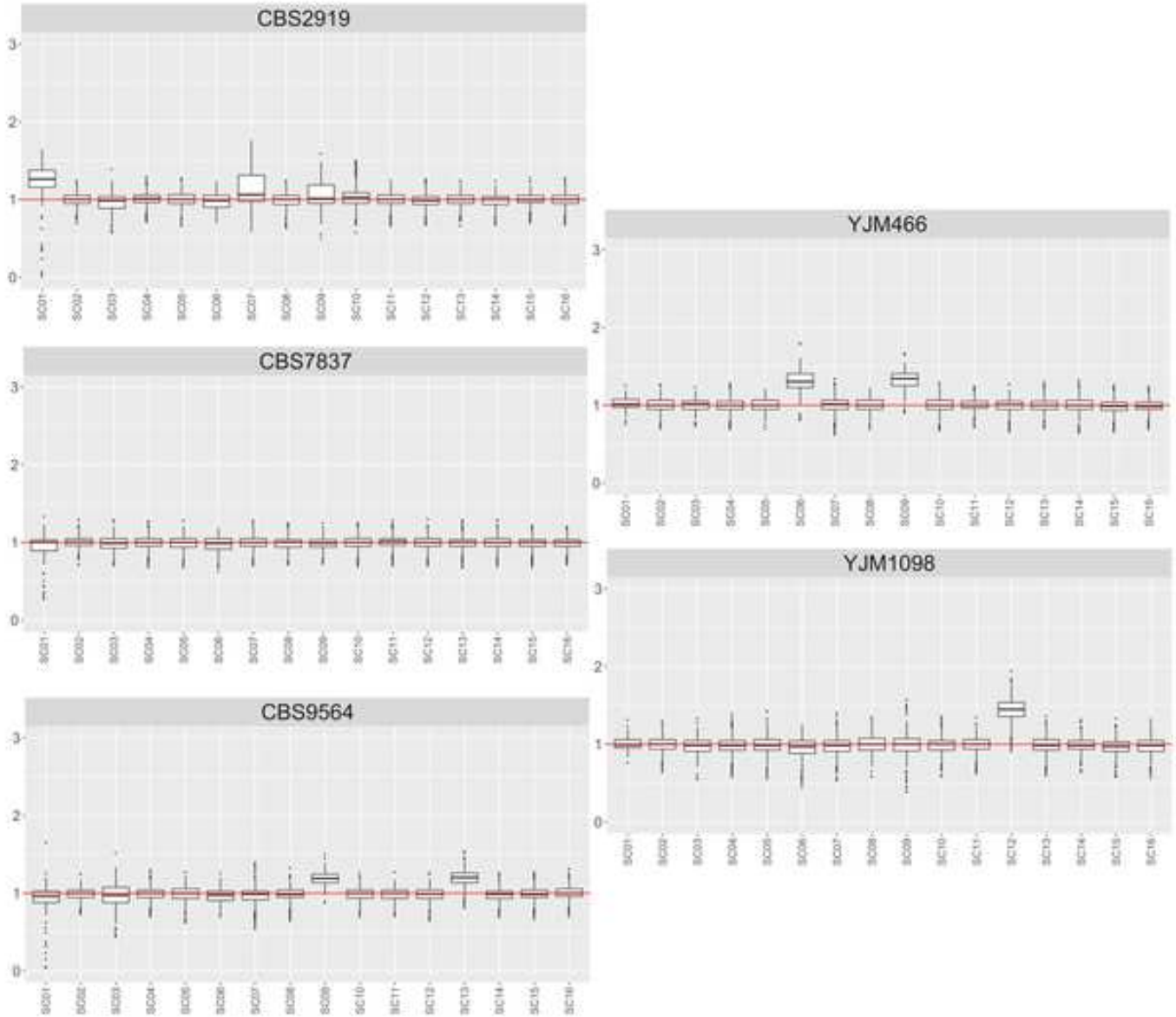
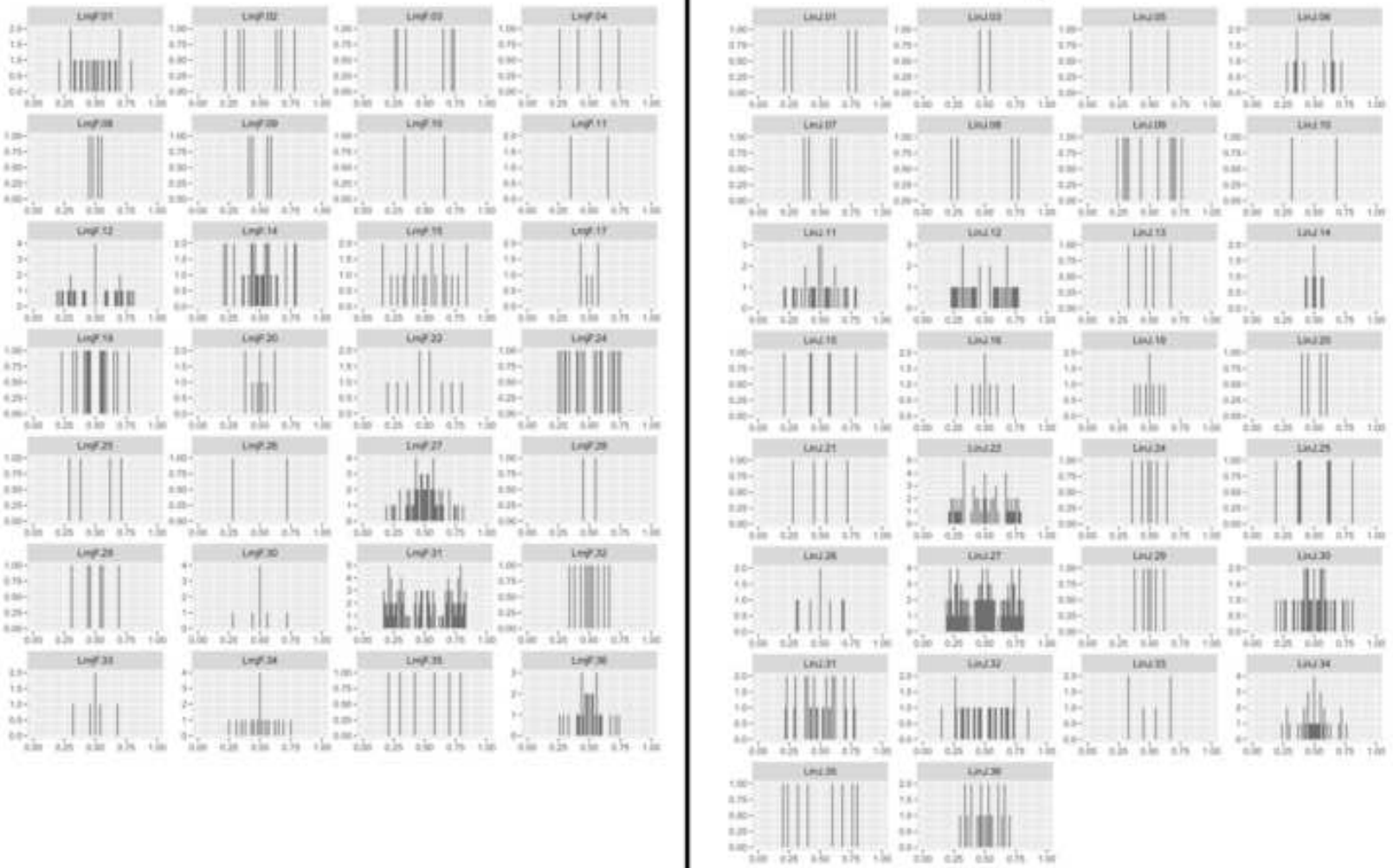


Figure5

Click here to access/download;Figure;Fig5.tif



Layer #1

Figure6

Click here to access/download;Figure;Fig6.tif

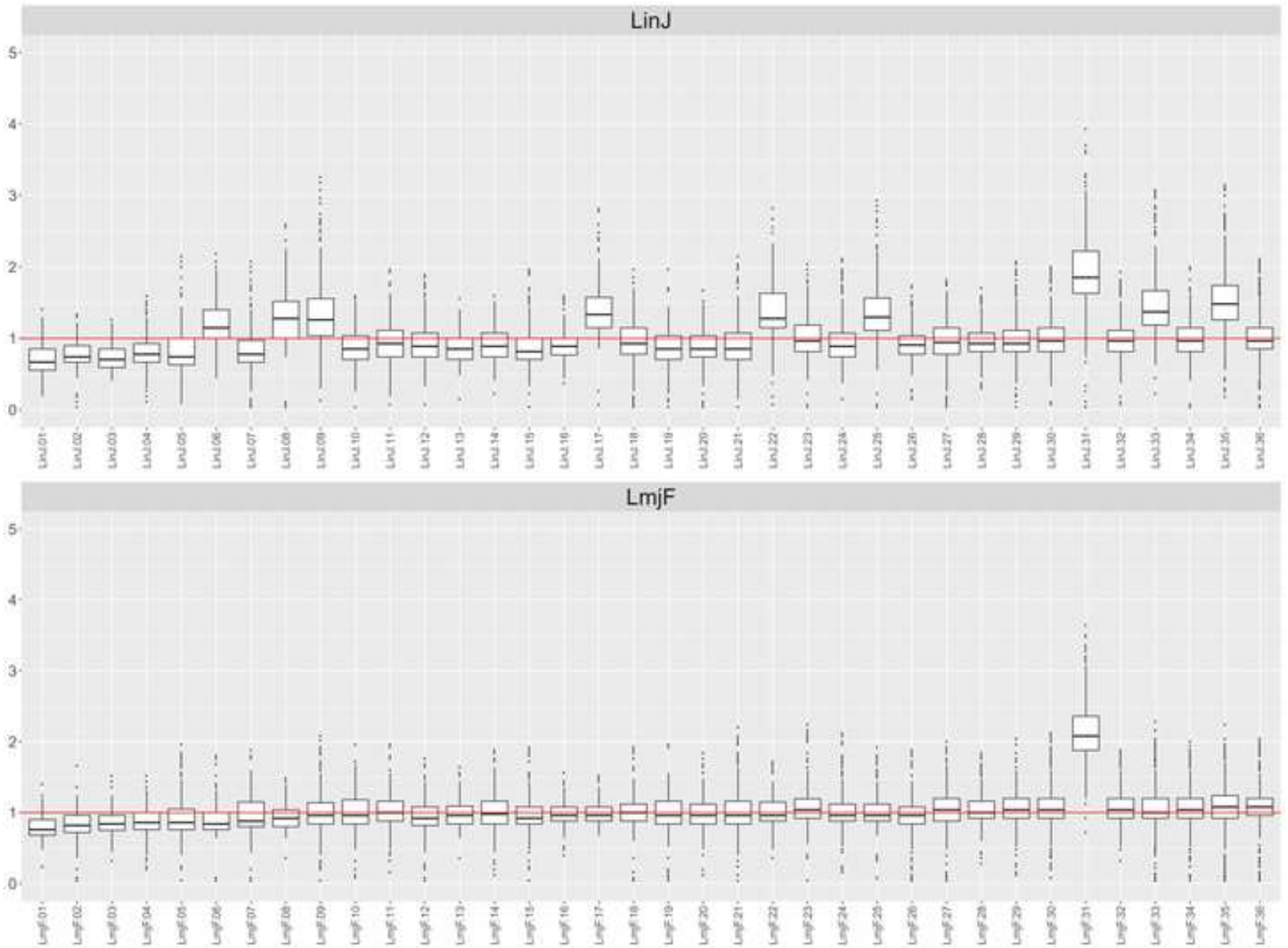


Fig S1

Fig S2

Fig S3

Fig S4

Fig S5

Fig S6


Click here to access/download
Supplementary Material
Supplementary data - Figure S6.pdf

Tables S1-S6