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

Enrico Giovanelli Tacconi Gimenez

Re-sequenciamento de *Corynebacterium pseudotuberculosis* e a busca por mecanismos de tropismo pelo hospedeiro

Belo Horizonte

2023

Enrico Giovanelli Tacconi Gimenez

Comparação de métodos de sequenciamento do genoma de *Corynebacterium pseudotuberculosis* 162

Dissertação apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Bioinformática

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo
Coorientador: Dr. Marcus Vinícius Canário Viana

Belo Horizonte

2023

043 Gimenez, Enrico Giovanelli Tacconi.
Comparação de métodos de sequenciamento do genoma de Corynebacterium pseudotuberculosis 162 [manuscrito] / Enrico Giovanelli Tacconi Gimenez. – 2023.
50 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo. Coorientador: Dr. Marcus Vinícius Canário Viana.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Sequenciamento de Nucleotídeos em Larga Escala. 3. Corynebacterium pseudotuberculosis. 4. Genômica. I. Azevedo, Vasco Ariston de Carvalho. II. Viana, Marcus Vinícius Canário. 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 INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

ENRICO GIOVANELLI TACCONI GIMENEZ

Às quatorze horas do dia **29 de setembro de 2023**, reuniu-se, através de videoconferência, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho do discente Enrico Giovanelli Tacconi Gimenez intitulado: **"Comparação de Métodos de Sequenciamento do Genoma de Corynebacterium Pseudotuberculosis 162"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Vasco Ariston de Carvalho Azevedo**, 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
Dr. Vasco Ariston de Carvalho Azevedo - Orientador	Universidade Federal de Minas Gerais	Aprovado
Dr. Marcus Vinicius Canário Viana - Coorientador	Universidade Federal de Minas Gerais	Aprovado
Dr. Bruno Silva Andrade	Universidade Estadual do Sudoeste da Bahia	Aprovado
Dr. Sandeep Tiwari	Universidade Federal da Bahia	Aprovado

Pelas indicações, o candidato foi considerado: Aprovado

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

Belo Horizonte, 29 de setembro de 2023.

Documento assinado eletronicamente por **Marcus Vinicius Canário Viana**, **Usuário Externo**, em 29/09/2023, às 15:43, 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 **Sandeep Tiwari**, **Usuário Externo**, em 29/09/2023, às 15:44, 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 **Bruno Silva Andrade**, **Usuário Externo**, em 29/09/2023, às 15:51, 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 **Vasco Ariston de Carvalho Azevedo**, **Professor do Magistério Superior**, em 02/10/2023, às 08:09, 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 **2667203** e o código CRC **BE7972B6**.

Referência: Processo nº 23072.259332/2023-55 SEI nº 2667203

AGRADECIMENTOS

Agradeço primeiramente e a Deus que me permitiu iniciar e concluir o presente trabalho, assim como ter possibilitado que eu tenha encontrado pessoas boas, dedicadas e pacientes e inteligentes durante meu percurso.

Agradeço a todos que contribuíram para a dissertação de mestrado, especialmente ao Prof. Dr. Vasco Ariston de Carvalho Azevedo pela orientação e por ter confiado a mim este trabalho, ao Dr. Marcus Vinícius Canário Viana pela coorientação e pela paciência e dedicação em me orientar e me ajudar, e ao Dr. Thiago de Jesus Sousa por ter me guiado inicialmente e permitido que eu participasse deste projeto.

Agradeço à minha família, aos meus amigos e à Gabriela pelo apoio.

RESUMO

Corynebacterium pseudotuberculosis é o agente causador da linfadenite caseosa (CLA) em vários animais, incluindo camelídeos, ruminantes, cavalos e humanos. A CLA pode resultar em perdas econômicas graves, especialmente em camelídeos, que são muito valorizados como animais de companhia. Pouco se sabe sobre os mecanismos envolvidos no tropismo por diferentes hospedeiros. A linhagem Cp162 isolada de camelo é a única deste hospedeiro que teve o seu genoma sequenciado. Este estudo teve como objetivo avaliar a evolução das montagens do genoma da linhagem Cp162, identificar possíveis genes relacionados ao tropismo por seu hospedeiro e compreender a diversidade genômica de *C. pseudotuberculosis* por meio de análises de genômica comparativa e filogenia. Desde a primeira montagem, houve um aumento de 88 Kb, 121 genes codificadores de proteínas, redução no número de pseudogenes e a correção de duas inversões e um rearranjo. Em comparação com 129 outros genomas da espécie, Cp162 possui quatro genes exclusivos, que codificam duas transposases e dois genes truncados. Três genes exclusivamente ausentes desta linhagem são *lysG* e dois que codificam "*NUDIX domain protein*", e uma proteína hipotética. Nenhum gene pôde ser associado ao tropismo ao hospedeiro camelo, o que sugere que o tropismo poderia estar associado a polimorfismos de sequência ao invés de presença e ausência de genes, e mais genomas deste hospedeiro precisam ser analisados. A análise mostrou que o pangenoma de *C. pseudotuberculosis* é fechado, mas existem muitos genes desconhecidos neste genoma que podem estar associados ao tropismo para outros hospedeiros.

Palavras-chave: Sequenciamento de nova geração, *Corynebacterium pseudotuberculosis,*

Montagem *ab initio,* pangenômica.

ABSTRACT

Corynebacterium pseudotuberculosis is the causative agent of caseous lymphadenitis (CLA) in several animals, including camelids, ruminants, horses and humans. CLA can result in severe economic losses, especially in camelids, which are highly valued as companion animals. Little is known about the mechanisms involved in tropism for different hosts. Cp162 strain isolated from camel is the only one of this host that had its genome sequenced. This study aimed to evaluate the evolution of genome assemblies of the Cp162 lineage, identify possible genes related to host tropism and understand the genomic diversity of *C. pseudotuberculosis* through comparative genomics and phylogeny analyses. Since the first assembly, there was an increase of 88 kb, 121 protein coding genes, a reduction in the number of pseudogenes and correction of two inversions and one rearrangement. Compared to 129 other genomes of the species, Cp162 has four unique genes, which encode two transposases and two truncated genes. Three genes uniquely missing from this lineage are *lysG* and two that encode "NUDIX domain protein", and a hypothetical protein. No genes could be associated with the camel host tropism, which suggests that the tropism could be associated with sequence polymorphisms rather than the presence and absence of genes, and more genomes from this host need to be analyzed. The analysis showed that the *C. pseudotuberculosis* pangenome is closed, but there are many unknown genes in this genome that may be associated with tropism for other hosts.

Keywords: Next Generation Sequencing, *Corynebacterium pseudotuberculosis, ab initio* assembly, pangenomics

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

BLAST	*Basic Local Alignment Search Tool*
CDS	*Coding Sequence*
CLA	*Caseous Lymphadenitis*
COG	*Cluster of Orthologous Groups*
Cp	*Corynebacterium pseudotuberculosis*
Cpequi	*Corynebacterium pseudotuberculosis* biovar equi
Cpovis	*Corynebacterium pseudotuberculosis* biovar ovis
GO	*Gene Ontology*
GEI	*Genomic Island*
GWAS	*Genome-Wide Association Studies*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
LCA	Linfadenite caseosa
OLC	*Overlap–Layout–Consensus*
PCR	*Polymerase Chain Reaction*
PLD	*Phospholipase D*
RASTtk	*Rapid Annotation using Subsystems Technology tool kit*
rRNA	RNA ribossomal
SNP	*Single Nucleotide Polymorphism*
tRNA	*Transfer Ribonucleic Acid*
VFDB	*Virulence Factors Database*
WGS	*Whole Genome Sequencing*
MLST	*Multilocus Sequencing Typing*
wgMLST	*Whole Genome Multilocus Sequencing Typing*
VTNRs	Números Variáveis de Repetições em tandem
Indels	Inserções e deleções

SUMÁRIO

1. INTRODUÇÃO

1.1 A pesquisa sobre *Corynebacterium pseudotuberculosis*

O Laboratório de Genética Celular e Molecular (LGCM) da Universidade Federal de Minas Gerais e colaboradores possuem várias linhas de pesquisa bem consolidadas, que estudam vários aspectos de *Corynebacterium*, como mecanismos de patogenicidade, métodos de diagnósticos e candidatos vacinais, utilizando metodologias da genômica estrutural e funcional. O primeiro genoma completo sequenciado pelo grupo foi o da C. *pseudotuberculosis* biovar Ovis linhagem 1002, publicada em 2010. A linhagem 162 de *C. pseudotuberculosis* biovar Equi foi isolada de um camelo, em 1999 no Reino Unido, e teve seu primeiro sequenciamento completo em 2012 realizado pelo grupo (HASSAN et al., 2012b), utilizando a plataforma SOLiD v3 Plus (Applied Biosystems). Esta é a única linhagem isolada de camelo que possui seu genoma completamente sequenciado. Várias versões desta linhagem foram depositadas no NCBI, e, ao longo delas, erros de montagem e sequenciamento foram sendo descobertas, o que pode impactar diretamente na sua localização em relação às demais linhagens. A última versão do genoma foi ordenada por mapa óptico e sequenciada utilizando uma plataforma mais moderna de sequenciamento, Illumina HiSeq 2500 (SOUSA et al., 2019b), portanto, podemos explorar melhor este genoma e buscar por *features* que permitam entender melhor a relação patógeno-hospedeiro.

1.2 Colaboradores

Dr. Bertran Brenig, pesquisador e professor da University of Göttingen, Alemanha.

Dr. Marcus Vinicius Canário Viana, pesquisador da UFMG, Brasil.

Dr. Thiago de Jesus Sousa, pesquisador do LACEN-ES, Brasil.

1.3 Estrutura da dissertação

Este manuscrito é dividido em Introdução, Revisão Bibliográfica, Justificativa e Objetivos, Artigo de pesquisa, Discussão Geral, Conclusão e Perspectivas, e um Apêndice.

A Revisão Bibliográfica aborda a evolução das tecnologias de sequenciamento genômico, conceitos e análises das áreas de genômica, e biologia de *C. pseudotuberculosis*;

2. REVISÃO DE LITERATURA

2.1 Genômica

2.1.1. Sequenciamento de genomas

O genoma é toda informação compreendida no material genético de um organismo (ALBERTS et al., 2002), e é estudado desde que Gregor Mendel inaugurou os estudos da genética. A genômica é o estudo dos aspectos do genoma, como conteúdo, estrutura, evolução dos genomas (GIBSON; MUSE, 2009). A genômica teve seu início quando o bacteriófago ϕx174 teve seu genoma completamente sequenciado, em 1977, por Sanger (SANGER et al., 1977). A genômica passou, a partir daí, a ser cada vez mais desenvolvida e tornou-se muito popular na década de 90, com a publicação do Projeto Genoma Humano, que apenas conseguiu publicar seu genoma completo em 2022 (NURK et al., 2022). Essa evolução lenta e gradativa, aconteceu graças ao desenvolvimento de diversas técnicas em biologia molecular e computação, principalmente o desenvolvimento de métodos de sequenciamento desses genomas. O sequenciamento é a técnica de identificar a ordem da sequência nucleotídica de uma molécula de DNA ou RNA, com o objetivo obter a informação sobre a molécula sequenciada (FIETTO; LAMËGO, 2015).

O sequenciamento pode ser usado para fins acadêmicos e clínicos, desde o entendimento do funcionamento dos organismos, caracterização de genes e regiões do genoma, até aplicações na medicina personalizada e o monitoramento em tempo real de epidemias e pandemias (LOMAN; PALLEN, 2015).

O sequenciamento começou na década de 70, com a publicação de dois métodos. O primeiro foi o método Maxam-Gilbert, que utilizava reações de degradação da extremidade 5’ do DNA isolado e fragmentado. Essas reações eram específicas para os nucleotídeos guanina e citosina, mas não eram específicos para adenina e timina. Desta forma, adenina e timina deviam ser inferidos indiretamente. Depois das reações, utilizava-se uma marcação da extremidade 5’ para que, em um gel de agarose, pudessem ser observados os fragmentos de acordo com seu tamanho. Era um procedimento mais simples comparando-se com o método de Sanger, pois podia ser realizado diretamente no DNA isolado e fragmentado, mas sua desvantagem é a utilização do fosfato radioativo (MAXAM; GILBERT, 1977). Logo em seguida, Sanger e colaboradores publicaram seu método que ficou amplamente conhecido (SANGER;

NICKLEN; COULSON, 1977). Este utilizava um procedimento similar à PCR, mas além de todos os ingredientes da PCR, em uma quantidade menor, utilizavam-se ddNTPs, nucleotídeos com uma hidroxila faltante, ou seja, dideoxinucleotídeos trifosfatados, que serviam para terminar precocemente o processo de síntese de DNA feito pela DNA polimerase. Inicialmente, os ddNTPs eram marcados com isótopos radioativos de fósforo, a fim de observar os fragmentos no gel de agarose e identificar as bases nitrogenadas. Posteriormente, o método passou por aprimoramentos que permitiram a redução do número de reações de quatro para uma, apenas, com a utilização de fluoróforos específicos para cada ddNTP, e a utilização de eletroforese capilar, que permitiu a automatização dos processos.

As revoluções do sequenciamento foram o sequenciamento *shotgun*, de alto rendimento ou massivamente paralelo, e de molécula única. A primeira grande revolução, conhecida como sequenciamento de *shotgun*, utilizava ainda o método de Sanger para sequenciar o genoma completo de microrganismos, porém fragmentando-o em diversos pedaços menores, a fim de possibilitar as leituras dos fragmentos completos para posterior montagem. Essa estratégia, inaugurada por Craig Venter, Hamilton Smith e colaboradores para sequenciar uma linhagem de *Haemophilus influenzae* possibilitou o sequenciamento de genomas completos, que era inviável por conta do elevado tempo de execução da técnica de Sanger e do tamanho das leituras geradas. A segunda revolução na genômica foi o sequenciamento de alto rendimento (*high throughput sequencing*) ou *Next-Generation Sequencing* (NGS), responsável por reduzir significativamente os custos do sequenciamento. Diversas plataformas invadiram o mercado com inovações que ampliaram o uso do sequenciamento na clínica, principalmente. Alguns exemplos são o HiSeq, da Illumina e o Ion Torrent, da Thermo Fisher. A terceira revolução ocorreu com o desenvolvimento de tecnologias de sequenciamento de molécula única, conhecidas também como sequenciamentos de leitura longa (*long-reads*). A primeira plataforma foi lançada pela empresa Pacific Biosciences, denominada Pacbio, e utiliza a tecnologia de sequenciamento *single-molecule real-time* (SMRT). Posteriormente, uma alternativa ao equipamento da Pacific Biosciences foi o MinION, desenvolvido pela empresa Oxford Nanopore Technologies, que apresentou um equipamento extremamente compacto capaz de oferecer sequenciamentos de leituras longas (LOMAN; PALLEN, 2015).

Todas as tecnologias de NGS têm em comum uma etapa anterior ao sequenciamento propriamente dito, que é a preparação da amostra a ser sequenciada, o que chamamos de biblioteca. Essa biblioteca pode ser preparada de diversas formas, mas de forma geral há o

isolamento do DNA do organismo ou da amostra a ser sequenciada, a fragmentação da molécula a ser sequenciada e sua ligação com sequências conhecidas, que são os adaptadores. A partir daí, podemos utilizar dois tipos diferente de bibliotecas. As bibliotecas de leituras do tipo *paired-end*, são obtidas do sequenciamento das extremidades de um mesmo fragmento. Com este tipo de biblioteca podemos ter uma estimativa do tamanho do inserto, ou seja, o tamanho entre as duas leituras pareadas, o que pode facilitar o processo de montagem do genoma. Outro tipo de biblioteca é a chamada de *mate-pair,* que é obtida a partir de um fragmento circularizado, e que, portanto, pode oferecer distâncias maiores entre leituras (NAGARAJAN; POP, 2013).

Um dos primeiros métodos de sequenciamento da segunda revolução na genômica foi o sequenciamento por ligação, sendo o equipamento SOLiD um dos primeiros equipamentos, depois do Polony, a utilizar este método. Nele, primeiro os fragmentos gerados a partir da amostra são ligados a adaptadores, que são sequências conhecidas de oligonucleotídeos. Depois disso, *probes* se anelam ao fragmento ligado ao adaptador. As *probes* possuem duas bases conhecidas, logo depois da sequência do adaptador, seguido de bases universais, ou degeneradas. Cada *probe* é ligada a um fluoróforo diferente, a depender das duas bases conhecidas. Durante a reação, essa *probe* se anela ao adaptador, no primeiro ciclo, e a DNA polimerase começa a extensão do fragmento, clivando a *probe* na sua porção 5', emitindo um sinal luminoso e expondo um fosfato 5' na fita de DNA. A hibridização ocorre novamente e ciclicamente, até que todo o fragmento tenha sido polimerizado. Neste momento, ainda não é possível realizar o *basecalling*, que é a chamada das bases sequenciadas, pois ainda não conhecemos toda a sequência do fragmento. Depois que ocorre a primeira síntese completa do fragmento complementar ao fragmento da amostra, todas as *probes* e adaptadores são removidos da amostra e o processo se reinicia. Após alguns ciclos, temos uma cobertura alta de todas as bases da nossa sequência-alvo, e então o sequenciamento se encerra, dando início à etapa do *basecalling*, quando deve ocorrer a leitura dos dados luminosos gerados em cada ciclo de sequenciamento. Um algoritmo então realiza essa leitura e nos devolve um arquivo com as leituras e suas respectivas qualidades associadas. Uma das grandes desvantagens dessa tecnologia é o tamanho muito pequeno das leituras, que pode ser entre 35 e 85bp.

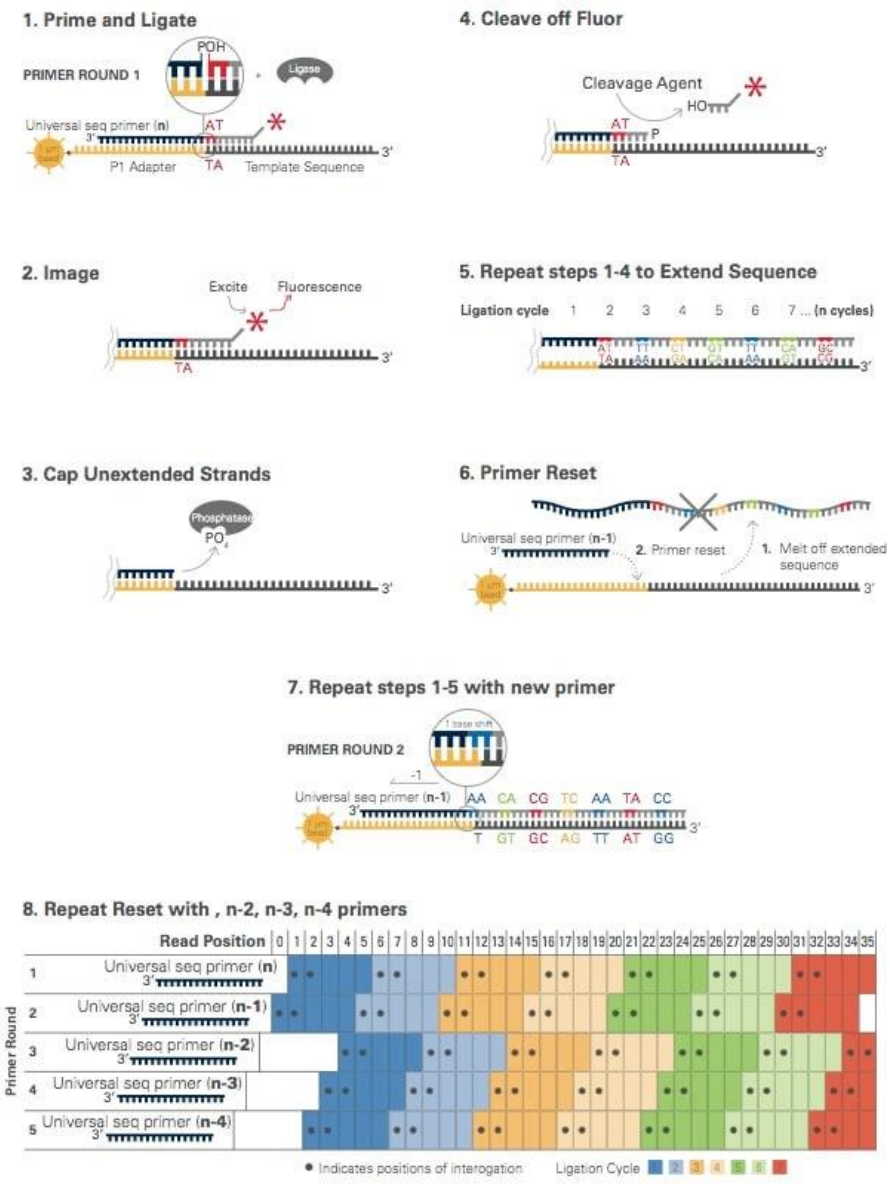


Figura 1: Metodologia de sequenciamento Sequencing by Oligonucleotide Ligation and Detection (SOLiD) (VALOUEV et al., 2008).

O método de sequenciamento mais utilizado hoje em dia é o de sequenciamento por síntese, com a Illumina dominando o mercado. Esse método consiste na obtenção dos dados durante a etapa de síntese da fita complementar à fita-molde, na amostra. Assim como nas tecnologias anteriores, há uma etapa anterior ao sequenciamento onde o DNA da amostra é fragmentado e ligado a adaptadores de sequência conhecida e amplificado em uma *flowcell* onde os fragmentos ficarão presos para serem sequenciados posteriormente. Adicionalmente, há uma etapa de amplificação em ponte, quando os clusters são formados na *flowcell*. Esses *clusters* são agrupamentos de fragmentos da amostra, e isso é feito para que haja uma amplificação do sinal luminosos emitido quando uma nova base nitrogenada é adicionada etapa

de síntese, tornando mais fácil a detecção e diferenciação do sinal de cada nucleotídeo. Durante o sequenciamento por síntese, um iniciador (*primer*) se anela a este adaptador e a síntese da fita complementar é feito pela enzima Taq-polimerase. Quatro nucleotídeos modificados são adicionados à reação. Eles possuem um bloqueio semelhante ao bloqueio utilizado no sequenciamento de Sanger e um fluoróforo específico para cada um. Sua extremidade 3'-OH impede o alongamento da fita assim que o nucleotídeo é incorporado, dando tempo para que a fluorescência seja emitida e o equipamento detecte este sinal (GUO et al., 2008; JU et al., 2006). Os nucleotídeos que não foram incorporados são removidos e uma câmera escaneia a *flowcell* em busca dos sinais emitidos pelos fluoróforos dos nucleotídeos incorporados. Depois disso o bloqueio no nucleotídeo é removido junto com o fluoróforo ligado, permitindo o reinício do ciclo de sequenciamento. O Illumina HiSeq tem um output de até 600Gb por corrida, e é um dos equipamentos que proporciona o sequenciamento mais barato do mercado (LIU et al., 2012). Ele utiliza dois softwares, HiSeq Control System (HCS) e Real Time Analyzer (RTA), que determinam o número e a localização de cada *cluster* presente na *flowcell*. Como os fluoróforos utilizados para a detecção de cada um dos 4 tipos de nucleotídeos podem interferir uns nos outros, é possível que a distribuição dos nucleotídeos ao longo da sequência interfira na leitura das bases e, consequentemente, na qualidade do sequenciamento (GOODWIN; MCPHERSON; MCCOMBIE, 2016; LIU et al., 2012). Apesar de não sofrer tanto com erros de homopolímeros, a plataforma tende a sub-representar regiões com alto grau GC ou AT, além de apresentar erros de substituição, que podem chegar até 3,8%, dependendo da localização do erro na leitura (DOHM et al., 2008).

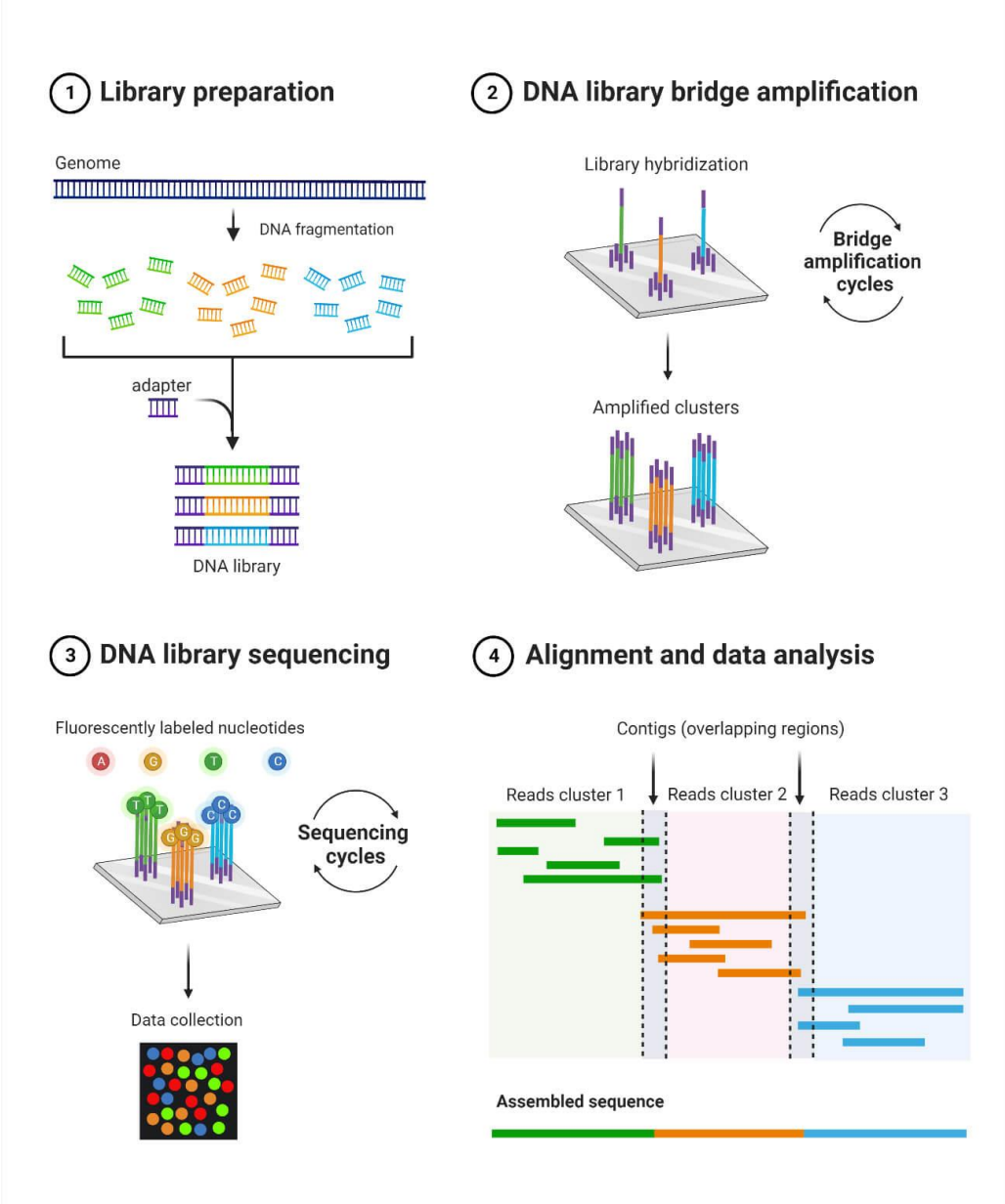


Figura 2 Metodologia de Sequenciamento por Síntese (Ion Torrent e Illumina) (RODRIGUEZ; KRISHNAN, 2023).

Outra marca que utiliza também o método de sequenciamento por síntese é a Thermo Fisher, com os equipamentos IonTorrent (https://www.thermofisher.com/br/en/home/brands/ion-torrent.html). Durante a reação de síntese da fita complementar, sempre que um nucleotídeo é adicionado à fita em construção, um próton é liberado, diminuindo o pH do meio e gerando, portanto, uma diferença de potencial elétrico. Essa diferença é captada pelo equipamento e interpretada como a adição de um nucleotídeo. Durante o processo, cada nucleotídeo é adicionado ciclicamente à reação, um de cada vez, então o equipamento é capaz de registrar a base sequenciada. Quando uma base

nitrogenada é adicionada mais de uma vez em sequência à fita em síntese, em regiões de repetição, a voltagem registrada pelo equipamento será proporcionalmente maior. Em outras palavras, o sinal emitido é o dobro caso dois nucleotídeos iguais sejam incorporados consecutivamente, o triplo, se três forem adicionados e assim por diante (FLUSBERG et al., 2010). Sua grande vantagem é a rapidez em relação às outras plataformas. Isso ocorre por ele não necessitar de câmera e fluorescência. O problema com essa tecnologia é sua ineficiência em sequenciar regiões de homopolímeros, que são regiões com repetições entre 6 e 8 bases. Nessas regiões é onde mais ocorrem os erros do tipo *indel*, quando uma base é inserida ou eliminada do sequenciamento (LOMAN et al., 2012).

Mais recentemente novas tecnologias foram desenvolvidas e ficaram conhecidas como a terceira geração de sequenciamento. O *output* dessas plataformas são arquivos com sequências longas, facilitando a montagem do genoma, principalmente em regiões repetitivas. As duas empresas que lideram este mercado atualmente são a Pacific Biosciences, com o sequenciador PacBio, e a Oxford Nanopore Technologies, com seus sequenciadores MinION. Ambas entregam dados de leituras longas, mas funcionam de maneiras diferentes (GOODWIN; MCPHERSON; MCCOMBIE, 2016).

O PacBio funciona com a fixação da polimerase e a circularização da molécula-alvo, permitindo que um feixe de luz incida diretamente sobre o nucleotídeo incorporado durante a reação do sequenciamento e que o mesmo fragmento seja sintetizado diversas vezes. A fluorescência emitida, bem como o tempo de emissão, é registrada, identificando o nucleotídeo incorporado (RONAGHI et al., 1996). Mesmo sendo o equipamento ideal para montagens *ab initio*, um erro conhecido como *single-pass* pode ocorrer quando a molécula é sequenciada uma única vez, o que pode gerar uma taxa de erro de até 15%.

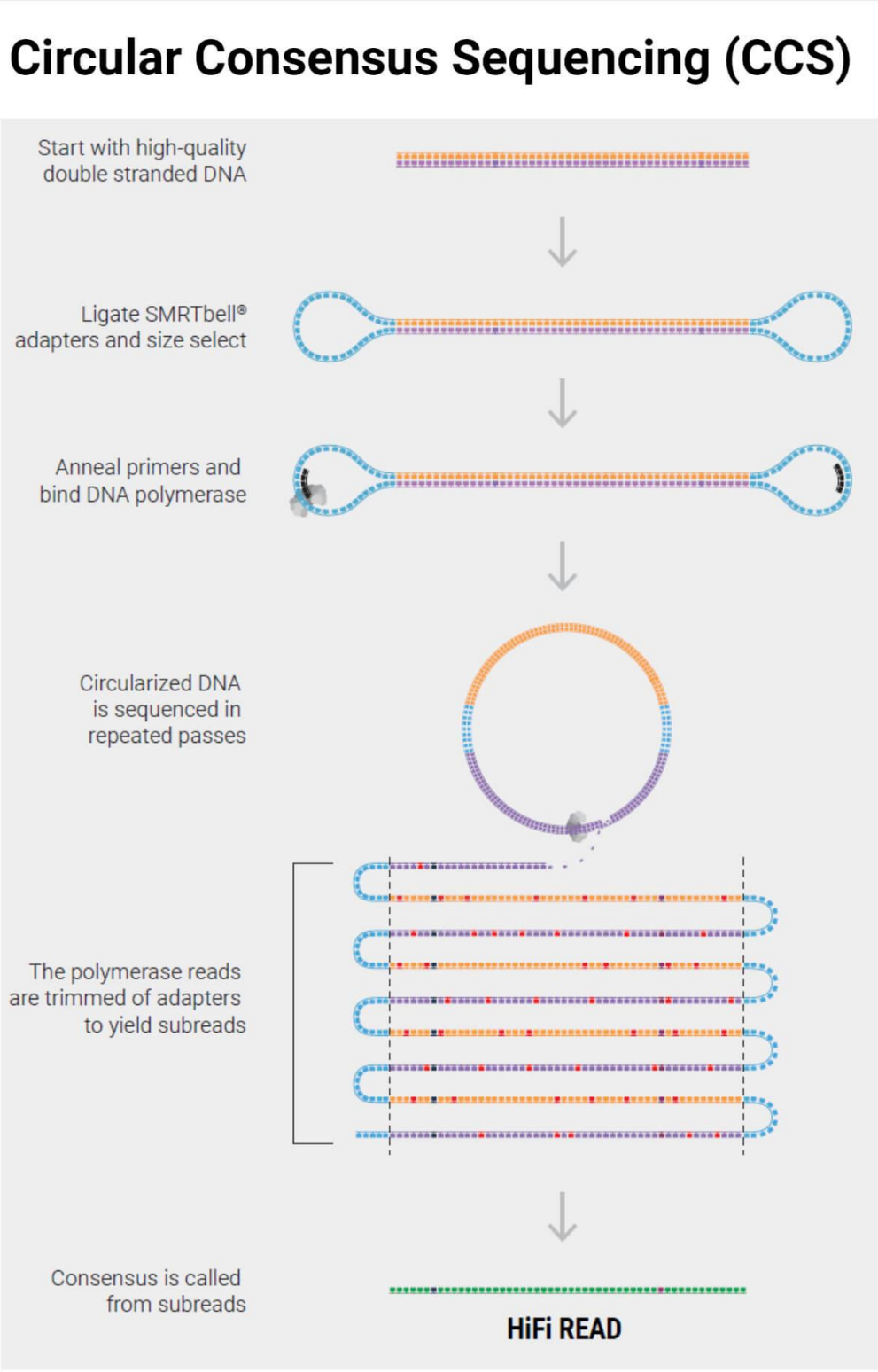


Figura 3 Metodologia de Sequenciamento de Molécula Única utilizada pelo PacBio (PACBIO, 2022).

MinION utiliza um poro que contém uma proteína motora fixa. À medida que a molécula atravessa o poro, a corrente elétrica que passa pelo mesmo é alterada, o que permite

identificar a base específica que está no local, uma vez que cada nucleotídeo, inclusive nucleotídeos modificados, possuem sua própria assinatura de alteração na corrente elétrica (GOODWIN; MCPHERSON; MCCOMBIE, 2016). Esse equipamento pode ter uma taxa de erro ainda maior, chegando a 30%.

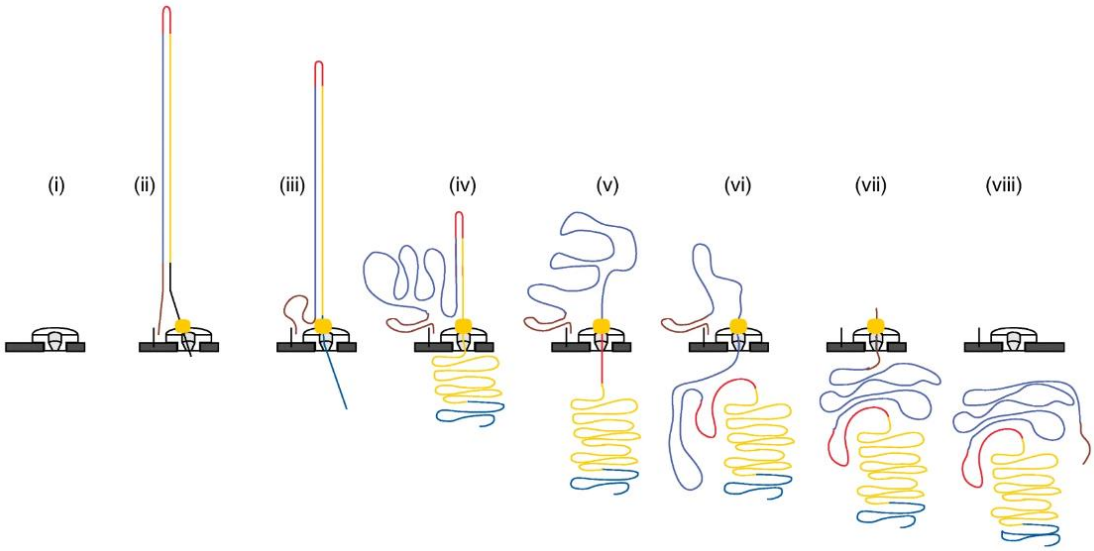


Figura 4 Passos na translocação de DNA através do nanoporo: (i) canal aberto; (ii) dsDNA com adaptador de liderança (azul), motor molecular ligado (laranja) e adaptador de pinça (vermelho) é capturado pelo nanoporo; a captura é seguida pela translocação do (iii) adaptador de liderança, (iv) fita molde (dourada), (v) adaptador de pinça, (vi) fita complementar (azul escuro) e (vii) adaptador de cauda (marrom); e (viii) o status retorna ao canal aberto. Adaptado de (JAIN et al., 2016).

2.1.2. Montagem de genomas

A montagem de genomas é o processo de obtenção de uma imagem do genoma real do organismo sequenciado a partir dos dados fornecidos pelo sequenciador utilizado. Utilizamos as leituras geradas pelo equipamento para formar um consenso a partir de algoritmos que compararão as leituras entre si ou contra um genoma previamente montado, para então formar porções maiores, chamadas *contigs*, que são as sequências contíguas de nucleotídeos. O sequenciamento não é capaz de diretamente nos fornecer a informação do genoma completo por conta dos vieses de cada sequenciador e também por conta de regiões de repetição no genoma que dificultam o sequenciamento e o trabalho dos programas de montagem, por se tratar de sequências que quase idênticas que estão presentes em diferentes partes do mesmo genoma (NAGARAJAN; POP, 2013). Não é recomendável também que essas lacunas geradas

pelo sequenciamento sejam ignoradas, uma vez que podem ser importantes para a identificação do gênero e espécie do organismo, como *clusters* de RNA ribossomal (rRNA).

Os *contigs* formados serão ordenados, formando o que chamamos de *scaffolds*, que possuem a informação da ordem dos genes, mas nem sempre possuem a informação das sequências nas lacunas onde o sequenciador não foi capaz de obter informação ou onde o montador não pôde distinguir a qual região determinadas leituras pertenciam.

Existem duas estratégias para montar um genoma, que são: montagem por referência e montagem *ab initio* (do latim "do início", ou também *de novo* para alguns autores). Na montagem por referência utilizamos um genoma completo previamente sequenciado para ordenar nossas leituras. Para isso utilizamos sempre o genoma de um organismo que seja mais próximo filogeneticamente do novo organismo-alvo, para garantir que o maior número de leituras possível seja alinhado ao genoma de referência. Essa abordagem tem sido deixada de lado por gerar um genoma, de diversas formas, enviesado. Quando fazemos uma montagem por referência, a ordenação dos genes do nosso genoma ficará igual à ordenação do genoma utilizado como referência, mesmo que tenham ocorrido eventos de inversão ou rearranjo no genoma de estudo. Outro erro que podo ocorrer é não representação de regiões que não estejam presentes no genoma de referência, mas apenas no genoma em montagem. Isso pode ocorrer quando as leituras do sequenciamento não encontram regiões de alinhamento no genoma de referência, e ficarão de fora do genoma final.

Para montagem por referência, utiliza-se o algoritmo da tabela *hash*, que salva em uma tabela as médias e medianas das frequências de cada nucleotídeo, para que seja calculado a seguir a cobertura do sequenciamento, assim como as sequências idênticas encontradas entre as chamadas "sementes" da referência e as leituras do sequenciamento (RAMOS et al., 2011). Os *softwares* mais utilizados para essa estratégia são Bowtie (LANGMEAD; SALZBERG, 2012) e o BWA (LI; DURBIN, 2009).

Ao optar pela abordagem *ab initio* para montar genomas, utilizamos apenas os dados gerados pelo sequenciador para chegar ao nosso genoma final. Para isso, podemos usar dois algoritmos, o *Overlap-Layout-Consensus* (OLC), ou o grafo *de Bruijn*. Qualquer um dos dois algoritmos assume *a priori* que há uma sobreposição entre as leituras, e que podemos usar essa sobreposição para ordená-las (MILLER; KOREN; SUTTON, 2010).

No algoritmo OLC, três etapas são realizadas no ordenamento das leituras. A primeira etapa consiste na comparação entre todas as leituras, gerando o que se chama *overlap*, e as porções de sobreposição são identificadas. Na segunda etapa do algoritmo, é criado um grafo onde os nós são as leituras e as arestas são as sobreposições. Ao final é gerada uma sequência consenso que passe por todos os nós do grafo. Caso seja possível passar por todos os nós do grafo uma única vez (Caminho Hamiltoniano), você terá seu genoma completo. Mas por conta da própria natureza das tecnologias de sequenciamento, algumas lacunas são geradas ao longo do genoma. Por conta disso, em vez do genoma completo, teremos *contigs* que são os melhores caminhos dentro de um conjunto de nós no grafo gerado (POP, 2009). Os programas que utilizam grafos *de Bruijn* para realizar a montagem também buscam por sobreposições, mas não entre as leituras completas. Em vez disso, dividem essas leituras em fragmentos de tamanho k, chamados *k-mers*. Os fragmentos podem ter tamanho padrão ou definido pelo próprio usuário. O algoritmo então vai procurar sobreposições entre esses fragmentos. Encontrando-os, o grafo começa a ser construído de forma que cada aresta representa uma ligação entre *k-mers*. Para determinar o consenso, o algoritmo vai encontrar o caminho Euleriano, passando por cada aresta do grafo uma única vez. (NAGARAJAN; POP, 2013). Um método que foi muito utilizado na ordenação de *contigs* foi o mapa óptico, que utiliza a posição de sítios de restrição no genoma. Criado em 1993 Schwartz e colaboradores (SCHWARTZ et al., 1993) e aprimorado diversas vezes depois (ASTON; MISHRA; SCHWARTZ, 1999; SAHA; RAJASEKARAN, 2014; SAMAD et al., 1995). Tem várias aplicações como tipagem de microrganismos (KOTEWICZ et al., 2007, 2008), estudo de alterações genômicas (FERREIRA et al., 2016) e ordenação de *contigs* gerados em montagens ab initio (ONMUS-LEONE et al., 2013). Essa técnica funciona em seis etapas: alongamento da fita de DNA, fragmentação por enzimas de restrição, marcação fluorescente dos fragmentos, registro das imagens desses fragmentos, cálculo do tamanho e ordenamento dos fragmentos. Primeiro, em uma lâmina de vidro, a fita de DNA é alongada pela força de cargas elétricas negativas dos fosfatos da mesma. Em seguida é escolhida uma enzima de restrição para fazer os cortes constantes ao longo da fita. Os fragmentos gerados são marcados com sonda fluorescente, possibilitando que, no próximo passo, seja possível que com um microscópio de alta precisão, seja feita uma imagem da fita fragmentada e, em seguida, o software Argus (Argus Imaging System, OpGen) possa calcular o tamanho de cada fragmento, a ordem dos mesmos e os pontos de clivagem, gerando assim um mapa de restrição daquele genoma. Esses dados são utilizados a seguir para ordenação dos *contigs*, alinhando cada ponto de clivagem do mapa óptico com as sequências de restrição dos

scaffolds gerados na etapa anterior de montagem. Este é o método mais confiável de ordenação de *contigs*, uma vez que utiliza dados provenientes diretamente do genoma físico. Mesmo assim, alguns erros ainda podem ocorrer como clivagens incompletas, clivagens não-específicas e erros de reconhecimento do tamanho dos fragmentos, que é mensurada a partir da quantidade de fluorescência emitida por cada fragmento.

O mapa óptico não é mais utilizado para fechamento de *gaps* pois a plataforma não é mais comercializada, e pelo uso de plataformas de sequenciamento com leituras longas, que podem auxiliar os montadores na finalização da montagem do genoma.

Após a ordenação dos *contigs* e geração dos *scaffolds* ainda pode ser necessário, que um trabalho extra seja realizado para fechar lacuna (*gaps*) que ainda estão presentes no genoma. Para isso podemos utilizar dois programas, o GabBlaster (DE SÁ et al., 2016) e o Gfinisher (GUIZELINI et al., 2016), que utilizam *contigs* de montagens alternativas para fechar os *gaps* da montagem principal. Alternativamente ou até de forma complementar, podemos utilizar o CLC Genomics Workbench (Qiagen) para mapear todas as leituras contra um genoma de referência, identificar a região de gap no genoma de referência e utilizar o consenso entre as leituras do sequenciamento para fechar o gap do seu genoma-alvo. Neste caso é importante observar se há cobertura suficiente nas leituras que que possa suportar a evidência de que aquela região realmente existe no seu genoma-alvo, e que não se está inserindo uma região que não existe apenas pelo fato de ela estar presente no seu genoma de referência. Também é possível fechar essas lacunas por ressequenciamento (LEHRI; SEDDON; KARLYSHEV, 2017; PEONA et al., 2021), ou mesmo a utilização de *contigs* de outras montagens (SOUSA et al., 2019a).

Uma vez finalizada, a qualidade da montagem de um genoma é avaliada por diferentes métricas (CHRISTENSEN, 2018). Primeiramente precisamos verificar a cobertura do sequenciamento, que é calculada pela seguinte fórmula

$$\frac{(\textit{Número de Leituras} \cdot \textit{Média dos Tamanhos das Leituras})}{\textit{Tamanho do Genoma}}$$

O tamanho do genoma pode ser obtido pela soma do tamanho dos *contigs* gerados ou com o tamanho esperado para outras bactérias da mesma linhagem. Outra métrica importante é o N50, que é o tamanho do *contig* na mediana dos *contigs* ordenados por tamanho decrescente,

ou seja, é o tamanho do *contig* que está na metade do genoma sequenciado, quando os *contigs* são ordenados de forma decrescente pelo seu tamanho. A completude de um genoma é uma métrica que nos diz qual a porcentagem de genes no genoma montado em relação aos genes esperados para aquela montagem.

Outro viés que pode estar presente em um genoma montado é a presença de contaminação e quimerismo (ORAKOV et al., 2021). A contaminação pode ser oriunda do processamento da amostra até o sequenciamento, ou até mesmo do processo computacional da montagem, e pode resultar em *contigs* quiméricos, quando leituras de duas linhagens diferentes são unidas no processo de montagem.

2.1.3. Anotação de genomas

A anotação é atividade que visa definir as propriedades estruturais e funcionais da sequência do genoma. Ela identifica características como genes, pseudogenes, genes de RNA, regiões não traduzidas, elementos móveis e outros. A anotação funcional utiliza a similaridade entre sequências para definir tais características. Desta forma, sequências semelhantes a outras sequências que possuem determinada função, possuem essa função atribuída no genoma anotado. A anotação estrutural tem por objetivo identificar o local das caraterísticas descobertas no genoma, ou seja, visa identificar a ordem dos genes naquele determinado genoma, e isso é feito utilizando algoritmos (BECKLOFF et al., 2012). A anotação da funcional pode ser obtida por ferramentas especializadas para, por exemplo, genes codificadores de proteínas, tRNA, rRNA, transportadores, genes de virulência e resistência, e sistemas CRISPR (SOLDATOS et al., 2015), ou por experimentação.

2.2.2 Identificação de espécies e tipagem

A identificação de espécies e tipagem são importantes para identificação de bactérias por sequências de DNA e pode utilizar alguns genes ou todo o genoma. O gene do rRNA 16S é um marcador universal por estar presente em todas as espécies bacterianas comumente e um valor de pelo menos 97% de identidade sugere a mesma espécie (CHRISTENSEN; OLSEN, 2018a). A identidade média de nucleotídeo (ANI) utiliza todo o genoma e um *cutoff* de ao menos 95% de identidade é tido como sendo ideal para agrupamento de amostras de mesma

espécie (JAIN et al., 2018). A hibridização de DNA-DNA digital utiliza um *cutoff* de 70% (MEIER-KOLTHOFF et al., 2013).

Para identificar linhagens de bactérias e para estudar populações bacterianas, alguns métodos de tipagem foram desenvolvidos, como *Multilocus Sequencing Typing* (MLST) e o *Whole Genome Multilocus Sequencing Typing* (wgMLST). O método MLST utiliza informações de sequência e localização cromossômica de genes *housekeeping* para tipar linhagens bacterianas. Seu sucessor wgMLST utiliza dados de sequenciamento de genoma completo para identificar polimorfismos de nucleotídeo único (SNPs), números variáveis de repetições em tandem (VNTRs), inserções e deleções (Indels) e rearranjos mediados por sequências de inserção. Esses dados são analisados e amostras que possuem as mesmas variações são agrupadas como sendo pertencentes à mesma linhagem (MAIDEN et al., 1998). Quando uma espécie tem seu genoma bem sequenciado e anotado, é possível definir um sistema de tipagem para diversas características importantes como, sorotipo, resistência a antibióticos e virulência (CHRISTENSEN; OLSEN, 2018b).

2.2.4. Análises genômicas comparativas

O alinhamento é um processo fundamental para a comparação de sequências, e talvez o mais utilizado na bioinformática. Ele consiste na comparação entre os caracteres de duas (alinhamento par-a-par) ou mais sequências (alinhamento múltiplo) (LESK, 2008). De um alinhamento, podemos obter diversa métricas que vão nos dizer suas caraterísticas. Uma delas é a identidade, que é a proporção de caracteres idênticos no alinhamento. Outra, a similaridade, é a proporção de caracteres similares como aminoácidos de mesma carga, ou purinas e pirimidinas. O valor *p* é a probabilidade de se obter uma pontuação maior ou igual ao utilizar os mesmos caracteres permutados aleatoriamente. O valor *e* indica a mesma probabilidade do valor *p*, mas para um alinhamento contra as sequências do banco de dados utilizado (JUNQUEIRA; BRAUN; VERLI, 2014).

A comparação de genes requer que sequências homólogas sejam identificadas. Ao buscar por homólogos, nos deparamos com alguns tipos de homologias. São elas a ortologia, quando os genes são originados de um gene ancestral, mas são ambos diferentes entre si e diferentes também de seu ancestral, gerando genes ortólogos; a paralogia, quando os genes são originados de cópias de um mesmo gene, gerando genes parálogos; e a xenologia quando os

genes são adquiridos por uma linhagem por transferência horizontal de uma segunda linhagem, gerando genes xenólogos. Essas relações podem ser inferidas utilizando grafos, árvores ou combinando métodos diferentes, o que chamamos de meta-métodos (ALTENHOFF; GLOVER; DESSIMOZ, 2019).

A análise pangenômica surgiu em 2005 com Tettelin e colaboradores em seu trabalho de sequenciamento de vários sorotipos de *Streptococcus agalatiae* (TETTELIN et al., 2005), buscando por novos candidatos vacinais, definiram o termo "pangenoma" como sendo todos os genes presentes em um clado, normalmente uma espécie. Este pangenoma possui então algumas subdivisões que vão caracterizá-lo. A primeira é o genoma central, ou *core genome*, que corresponde a todos os genes que estão presentes em todos os sequenciamentos disponíveis daquele clado. O genoma *shell* são os genes compartilhados por alguns sequenciamentos, mas não por todos. O *singletons*, são s genes presentes apenas em um dos genomas sequenciados daquele clado.

O pangenoma de uma espécie pode ser aberto ou fechado, e essa designação indica o quanto um pangenoma pode crescer ainda. Um pangenoma aberto significa que muitos genes são adicionados ao pangenoma quando um novo sequenciamento é adicionado à análise. Isso pode indicar que aquela espécie pode conter poucos sequenciamentos ainda, ou que aquela espécie analisada pode transferir facilmente seus genes entre si e/ou colonizar diferentes tipos de ambientes. Um pangenoma aberto tende a crescer quando um novo genoma é sequenciado. O pangenoma fechado significa que poucos genes são adicionados ao pangenoma quando um novo sequenciamento é feito. A estimativa é realizada de acordo com a fórmula $n = \kappa * N^{\gamma}$, na qual n é o número de genes, N é o número de genomas, e κ e γ ($\alpha = \gamma$-1) são parâmetros livres determinados empiricamente. Um valor de $\alpha > 1$ significa que o pangenoma está fechado e um valor de $\alpha < 1$ significa que o pangenoma está aberto. Sabemos, no entanto, que o pangenoma de uma espécie nunca será totalmente descrito, pois as bactérias estão constantemente trocando seu material genético umas com as outras por transferência horizontal (MEDINI et al., 2005; TETTELIN et al., 2008).

O alinhamento de sequências pode ser utilizado para inferir a filogenia de genes e organismos. As sequências são alinhadas e a filogenia pode ser inferida por métodos baseados em distância como *Neihgbor-Joining*, ou um método baseado em caracteres, como Máxima Verossimilhança. O valor de suporte de um nó pode ser calculado por um teste de permutação, como o *bootstrap*. Este método cria pseudo-amostras pela reamostragem de posições do

alinhamento de sequências, constrói novas árvores e calcula a proporção de árvores que possui um determinado nó (CHRISTENSEN; OLSEN, 2018c; YANG; RANNALA, 2012).

2.2.6. Patogenômica

A patogenômica é o estudo de como modificações, ganhos e perdas de genes de um microrganismo influenciam seu grau de virulência (PALLEN; WREN, 2007). Com a patogenômica, podemos detectar alterações no genoma bacteriano que ocorrem durante sua instalação no hospedeiro (SHEPPARD; GUTTMAN; FITZGERALD, 2018).

A patogenômica permite que identifiquemos, por exemplo, marcadores moleculares para diagnóstico de microrganismos (ALMEIDA et al., 2017; BADELL et al., 2019), candidatos ao desenvolvimento de vacinas, utilizando técnicas de vacinologia reversa (RAPPUOLI, 2000; RAPPUOLI et al., 2016) e de descobrimento de novos quimioterápicos (BARH et al., 2011; KUMAR JAISWAL et al., 2017). Todas essas aplicações são eficazes em reduzir o tempo e o custo de desenvolvimento das aplicações citadas.

2.1 Biologia de *Corynebacterium pseudotuberculosis*

2.1.1. Taxonomia e microbiologia

O gênero *Corynebacterium* pertence à família Corynebacteriaceae, ordem Corynebacteriales, classe Actinobacteria, filo Actinobacteria, mais recentemente designado Actinomycecota (OREN; GARRITY, 2021). O gênero possui espécies de bactérias gram-positivas, pleomórficas, imóveis, catalase positivas, anaeróbias facultativas ou aeróbias, com conteúdo G+C alto, variando de 46% a 74%. A parede celular possui arabinose e galactose como açúcares principais e pode conter ácido micólico. A espécie tipo do gênero é a *Corynebacterium diphtheriae*. O gênero possui mais de 1125 espécies de vida livre, comensais e patogênicas, com espécies de interesse biotecnológico, médico e veterinário (BERNARD; FUNKE, 2015; SCHOCH et al., 2020). Neste gênero, as espécies patogênicas são agrupadas no clado chamado “complexo da *C. diphtheriae*”, que é um clado de espécies que produzem a toxina diftérica (DT) quando adquirem o gene *tox* de um profago (BERNARD; FUNKE, 2015; DAZAS et al., 2018). Atualmente, o grupo é composto por seis espécies: *C. belfantii* (DAZAS et al., 2018), *Corynebacterium diphtheriae* (BERNARD; FUNKE, 2015), *C.*

pseudotuberculosis (DORELLA et al., 2006a), *C. rouxii* (BADELL et al., 2020), *C. silvaticum* (DANGEL et al., 2020) e *C. ulcerans* (RIEGEL et al., 1995).

A espécie *Corynebacterium pseudotuberculosis* foi isolada de linfonodos de gado pela primeira vez em 1888 por Edmond Nocard, veterinário e bacteriologista francês. Em seguida, Hugó Preisz coletou essa bactéria em abcessos de ovelha, e foi então designada Bacilo Preisz-Nocard. Apenas mais tarde foi identificada como sendo um membro do gênero *Corynebacterium* e nomeada *C. pseudotuberculosis* . É um patógeno intracelular facultativo, que pode utilizar glicose, frutose, maltose, galactose e manose como fonte de energia. É positiva para produção de urease e vermelho de metila, demonstrando que ela utiliza a via ácida de fermentação da glicose (BERNARD; FUNKE, 2015).

C. pseudotuberculosis pode infectar mamíferos em diversos grupos diferentes, e possui dois biovares, sendo eles, biovar ovis, que é mais comumente encontrado em infecções de caprinos e ovinos e causa linfadenite caseosa (LC), já tendo sido identificado como agente causador de infecções em humanos, enquanto o biovar equi é mais encontrado em infecções de cavalos, búfalos e outros mamíferos e causa diferentes doenças. A identificação da espécie pode ser realizada com API Coryne system (API-bioMérieux, Inc., La Balme les Grottes, France), enquanto os biovares podem ser identificados com teste de redução de nitrato, onde equi é positivo e ovis é negativo (DORELLA et al., 2006b).

2.1.2. Patogenicidade e epidemiologia

A LCA possui prevalência mundial, principalmente na Austrália, Argentina, Nova Zelândia, Estados Unidos da América e em países do Oriente Médio, como Tunísia e Jordânia (AL-RAWASHDEH; AL-QUDAH, 2000; ARSENAULT et al., 2003; BINNS; GREEN; BAILEY, 2002; CONNOR et al., 2000; PATON et al., 2003; SAID; BENZARTI; ABDELKADER, 2002). No Brasil, a região mais afetada é o Nordeste, pela maior concentração de criações de caprinos e ovinos do país (RIBEIRO et al., 2001). Dados da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) mostras que em alguns estados da região Nordeste do Brasil, pode-se encontrar uma prevalência de cerca de 40% da CLA em ovinos (Figura 1), e de até 33% em caprinos (Figura 2).

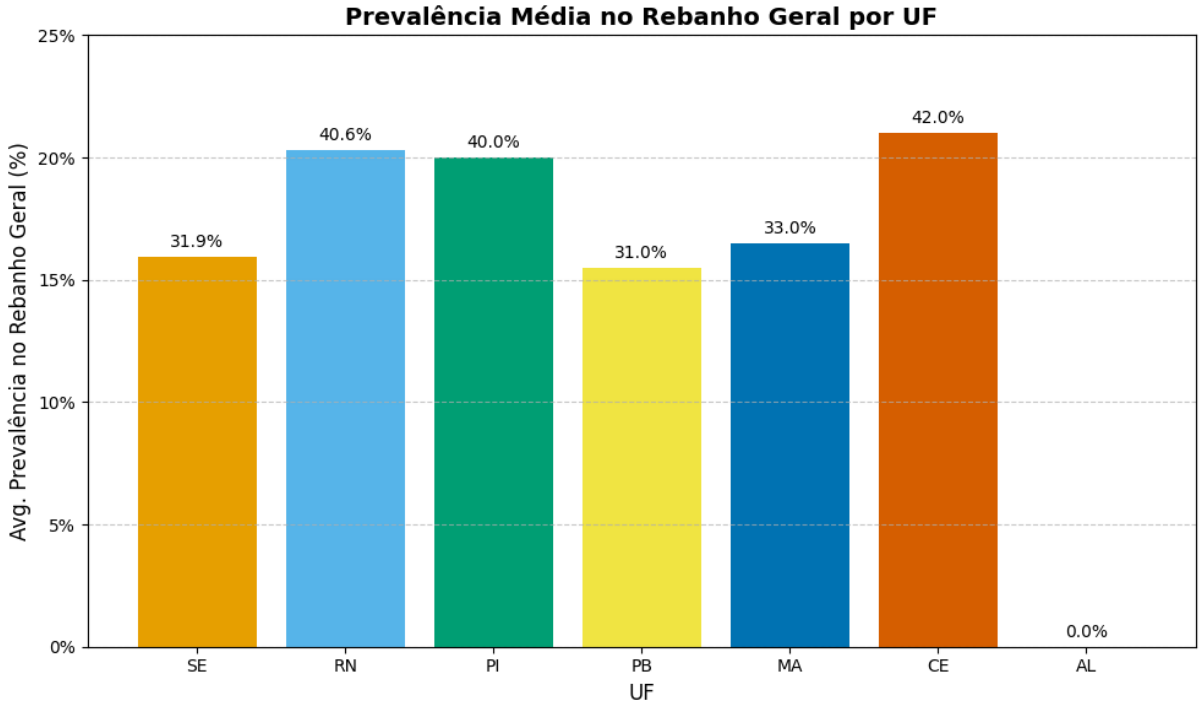


Figura 5: Prevalência da Linfadenite Caseosa em ovinos na região Nordeste do Brasil.

Disponível em (CENTRO DE INTELIGÊNCIA E MERCADO DE CAPRINOS E OVINOS. LINFADENITE CASEOSA (LC), 2023)

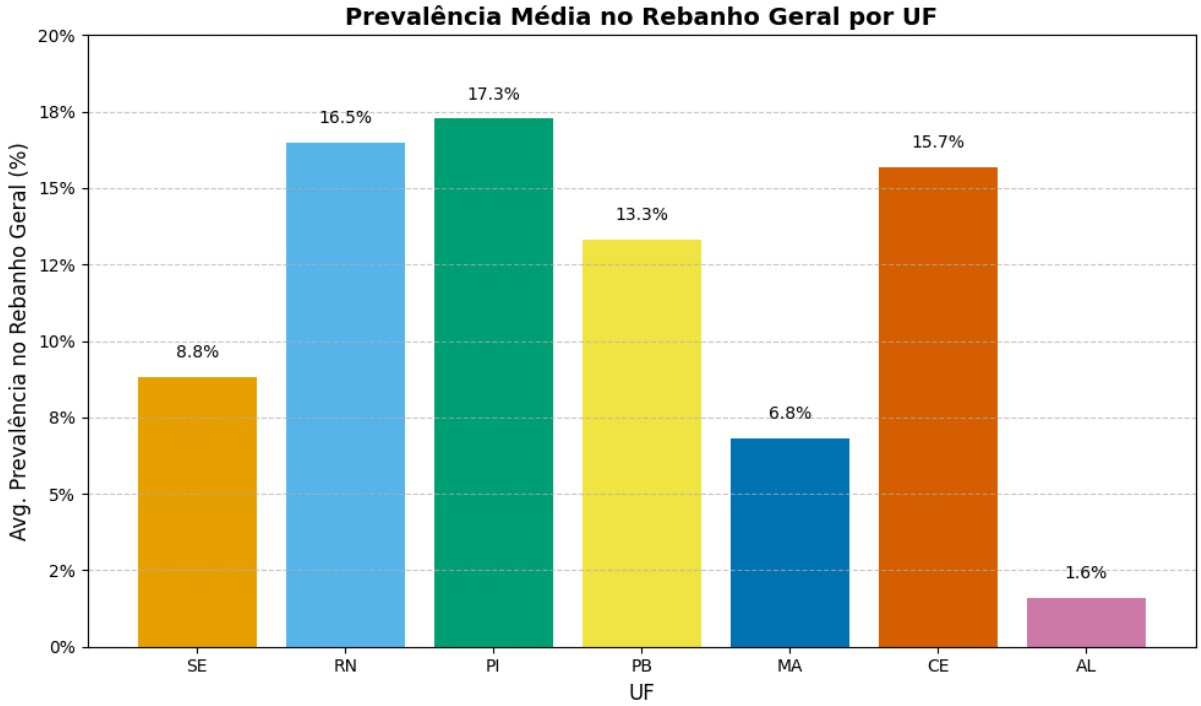


Figura 6: Prevalência da Linfadenite Caseosa em caprinos na região Nordeste do Brasil.

Disponível em: (CENTRO DE INTELIGÊNCIA E MERCADO DE CAPRINOS E OVINOS. LINFADENITE CASEOSA (LC), 2023)

Patogenicidade é a capacidade que um agente biológico pode ter em causar dano a um organismo, e os fatores de virulência são as características deste agente biológico que causam tal dano a um determinado organismo em um determinado nicho, mas não necessariamente em outro nicho (HILL, 2012; TAUCH; BURKOVSKI, 2015). Fatores de virulência são características ou estratégias dos microrganismos patogênicos para facilitar a infecção do hospedeiro e seu estabelecimento naquele organismoClique ou toque aqui para inserir o texto..

A fosfolipase D (PLD) é uma toxina que degrada a esfingomielina da membrana celular do hospedeiro, o que facilita a disseminação do patógeno e causa a morte de macrófagos no hospedeiro. Essa toxina é o principal fator de virulência de *C. pseudotuberculosis* (D'AFONSECA et al., 2008; MCKEAN; DAVIES; MOORE, 2007).

O biovar ovis causa a Linfadenite Caseosa (LC) em caprinos e ovinos, gado e em camelos (TERAB et al., 2021). O biovar equi causa linfangite ulcerativa em equinos (DORELLA et al., 2006a), doença da pele edematosa em búfalos (SELIM, 2001).

2.1.3. Genômica

Várias linhagens de *Corynebacterium pseudotuberculosis* já foram sequenciadas, e o genoma dessa espécie possui entre 2,28 e 2,44 Mb e com conteúdo GC de aproximadamente 52%. O genoma de referência dessa espécie pertence à linhagem 1002 (Número de acesso: NZ_CP012837.1) e possui onze ilhas de patogenicidade preditas, muitas delas presentes apenas em linhagens do biovar ovis. Diversos trabalhos também mostram expressões diferenciais em diversos genes em culturas com restrição de ferro. Alguns desses genes são relacionados à manutenção da homeostasia de ferro no organismo e são positivamente regulados nessas condições, e estão em ilhas genômicas já conhecidas como sendo associadas a fatores de virulência (IBRAIM et al., 2019).

Um trabalho do grupo identificou candidatos vacinais com metodologias de vacinologia reversa, utilizando dados do próprio genoma das espécies (ARAÚJO et al., 2019). No entanto, esse trabalho utilizou a primeira versão do genoma de *C. pseudotuberculosis* da linhagem 162

para suas análises. Uma atualização deste trabalho pode trazer novas perspectivas para os estudos atuais de linfadenite caseosa.

2.1.4. A linhagem 162

A linhagem 162 foi isolada de um camelo, em 1999, no Reino Unido. O animal apresentava abcessos no pescoço no momento da coleta. Ela foi sequenciada pela primeira vez por Hassan, em 2012, utilizando a tecnologia SOLiD v3 Plus (Applied Biosystems), e identificada como pertencente à linhagem equi (HASSAN et al., 2012a). Essa linhagem, apesar de ser classificada como equi por conta de seu fenótipo, parece não se agrupar perfeitamente com outras linhagens equi, por possuir algumas diferenças nos padrões de mutação (OLIVEIRA et al., 2016). Apesar disso, essa classificação foi feita antes dos sequenciamentos com as plataformas mais atuais, Illumina HiSeq e IonTorrent, e do ordenamento do genoma por mapa óptico, realizado por Sousa (SOUSA et al., 2019a). Além disso, é a única linhagem de *C. pseudotuberculosis* classificada como equi que foi isolada de um camelo e está depositada no RefSeq.

3. JUSTIFICATIVA E OBJETIVOS

3.1 Justificativa

C. pseudotuberculosis é uma espécie de bactéria patogênica que pode infectar diversos hospedeiros diferentes. Entender melhor a genômica das diferentes linhagens vai nos ajudar a desenvolver melhores métodos de diagnóstico, tratamento e prevenção das infecções causadas por esse patógeno.

3.2 Objetivo geral

O objetivo deste trabalho é a análise genômica de *C. pseudotuberculosis* para a linhagem 162, com foco nas diferenças entre os sequenciamentos.

3.3 Objetivos específicos

Os objetivos específicos foram:

- Comparar as três montagens da linhagem *C. pseudotuberculosis* realizada com três diferentes plataformas de sequenciamento de genômico.
- Caracterizar os genes dentro das regiões excluídas, invertidas ou rearranjadas;
- Discutir as causas dos erros de sequenciamento e montagem e relacioná-los às tecnologias utilizadas em cada uma das versões.
- Identificar genes que podem estar relacionadas ao tropismo por camelos.

4. ARTIGO

4.1 Artigo – Resequencing of *Corynebacterium pseudotuberculosis* Cp162 and the Search for Host Tropism Mecanisms

Neste capítulo investiguei as características do pangenoma da espécie, bem como as diferenças entre as três versões do sequenciamento da Cp162, os genes exclusivos da linhagem e possíveis correlações entre genes e a infecção em camelos, utilizando análises de pangenômica, GWAS e *scripts in house* para análises manuais.

5. DISCUSSÃO GERAL

As versões do genoma da Cp162 possuem diferenças importantes. O sequenciamento feito com Illumina HiSeq e ordenação dos contigs com mapa óptico produziu um genoma maior, com maior número de CDSs e menos pseudogenes. A ordem dos genes também foi corrigida. Ter a ordem e o conteúdo gênico corretos é importante em estudos de plasticidade genômica (ref). As duas primeiras versões do genoma foram geradas a partir de dados produzidos pelas tecnologias SOLiD e Ion Torrent, que são conhecidas por altas taxas de erros de *indel*, fazendo com que ocorram erros de *frameshift* ou erros na janela de leitura, o que pode levar a um número maior de pseudogenes pelos *softwares* de anotação automática. Portanto, dados gerados a partir dessas tecnologias devem ser utilizados com cautela em estudo que necessitem de genomas completos, como estudo de plasticidade genômica e pangenômica.

Há, no genoma da linhagem, regiões de rearranjos e inversões que podem ter ocorrido justamente por estarem localizadas exatamente entre genes que codificam para transposases (ref). Foram encontrados ao todo nove fatores de virulência na linhagem. Três foram identificados utilizando o PanViTa (*DIP_RS14950, mprA, tufA)*, com a base de dados VFDB, e os outros seis (*pld*, *cpp*, *nanH*, *spaC*, *sodC*, *pknG*), já descritos na literatura, utilizando o BLASTp. Há essa diferença pelo fato de que os seis fatores de virulência não identificados pelo PanViTa não estão presentes na base de dados de Corynebacterium no VFDB (http://www.mgc.ac.cn/cgi-bin/VFs/genus.cgi?Genus=Corynebacterium), ou pelas sequências buscadas não terem um valor de identidade com as sequências da base de dados acima do valor mínimo estipulado pelo PanViTa. Tudo isso demonstra a necessidade de haver uma atualização da base de dados de *Corynebacterium*.

Foram encontradas 13 ilhas genômicas (GEI) na linhagem. A GEI5 foi encontrada apenas nas linhagens Cp162 (camelo), I37 (gado) e G1 (alpaca), no entanto essa ilha pode ter sido adquirida de um ancestral em comum, e não necessariamente estar relacionada ao tropismo por esses hospedeiros, uma vez que as linhagens 262 e I19 também infectam gado. Foi identificado também um profago incompleto, um sistema CRIPR-Cas Tipo I-E e três arranjos CRISPR.

Sendo a Cp162 a única linhagem isolada de um camelo e sequenciada, procuramos em seu genoma por genes que poderiam estar envolvidos com seu tropismo por este hospedeiro. Dentre os genes exclusivos da linhagem foi possível identificar genes de transposases e proteínas truncadas. Nas proteínas ausentes apenas nesta linhagem identificamos a proteína

lysG da superfamília NUDIX, e uma proteína hipotética sem nenhum domínio conservado. Como não há nenhuma relação aparente entre essas proteínas e o tropismo por camelos, próximos estudos podem ser feitos quando novos genomas de bactérias isoladas de camelos forem sequenciados, então análises de GWAS e busca por SNPs poderão ser feitas.

Os resultados de filogenia com os 130 genomas corroboram com resultados de estudos anteriores, em que foi demonstrado que o biovar ovis foi originado do biovar equi, que possui dois hospedeiros exclusivos, cavalos e búfalos, e que possuem adaptações para infectar esses hospedeiros. O pangenoma da espécie é fechado, portanto não é esperado que encontremos novos genes ao sequenciar novos genomas, no entanto muitas proteínas hipotéticas ainda podem ter sua função descoberta. A resistência a rifampicina foi identifica em todos os genomas do grupo (*rpoB* e *rpbA*), sugerindo que este antimicrobiano não seja utilizado no tratamento de animais doentes.

Houve também uma diferença no número de anotações de proteínas hipotéticas em cada uma das versões, com 294 proteínas hipotéticas na primeira versão do genoma, 301 na segunda e 262 na terceira, indicando uma redução desde a primeira versão até a mais recente. Esta reduçãopode ter sido causada por atualização do programa de anotação PGAP e pela redução no número de erros de sequenciamento do tipo *indel*, que causam mudança na matriz de leitura e anotação de fragmentos como genes separados. Uma análise realizada pelo grupo com outras linhagens de *C. pseudotuberculosis*, cujo objetivo incluía a anotação de proteínas hipotéticas, mostrou que a maioria dessas proteínas foram reanotadas como componentes da membrana celular. (ARAÚJO et al., 2020).

6. CONCLUSÃO E PERSPECTIVAS

As análises genômicas realizadas contribuíram para o conhecimento sobre a genômica de *C. pseudotuberculosis*, em especial a linhagem Cp162. O pangenoma da espécie é fechado ($\alpha > 1$), o que sugere que há poucos novos genes a serem identificados com novos sequenciamentos. As análises não mostraram genes que possam estar relacionados à infecção do hospedeiro, no entanto, é necessário que sejam sequenciados mais genomas provenientes deste hospedeiro. Além disso, muitas proteínas da linhagem Cp162 ainda possuem função desconhecida, e estão anotadas como hipotéticas. De acordo com os dados, é sugerido que não se utilize a rifampina para o tratamento de animais doentes, visto que todos os genomas do grupo apresentaram genes de resistência contra este antimicrobiano.

Como perspectivas, esperamos sequenciar mais genomas de bactérias isoladas de camelos infectados, realizar novas análises de GWAS para identificação de genes associados à infecção específica de camelos. Além disso, a anotação das proteínas hipotéticas identificadas na última versão do genoma pode elucidar o tropismo da linhagem 162 pela espécie de camelo.

REFERÊNCIAS

ALBERTS, B. et al. **Molecular Biology of the Cell. 4th edition.** 4th. ed. [s.l.] Garland Science, 2002.

ALMEIDA, S. et al. Quadruplex PCR assay for identification of Corynebacterium pseudotuberculosis differentiating biovar Ovis and Equi. **BMC Veterinary Research**, v. 13, n. 1, p. 1–8, 2017.

AL-RAWASHDEH, O. F.; AL-QUDAH, K. M. Effect of Shearing on the Incidence of Caseous Lymphadenitis in Awassi Sheep in Jordan. **Journal of Veterinary Medicine Series B**, v. 47, n. 4, p. 287–293, maio 2000.

ALTENHOFF, A. M.; GLOVER, N. M.; DESSIMOZ, C. Inferring Orthology and Paralogy. Em: ANISIMOVA, M. (Ed.). **Evolutionary Genomics - Statistical and Computational Methods**. 2. ed. Lausanne, Switzerland, Switzerland: Humana Press, 2019. p. 149–175.

ARAÚJO, C. L. et al. Prediction of new vaccine targets in the core genome of Corynebacterium pseudotuberculosis through omics approaches and reverse vaccinology. **Gene**, v. 702, p. 36–45, jun. 2019.

ARAÚJO, C. L. et al. In silico functional prediction of hypothetical proteins from the core genome of *Corynebacterium pseudotuberculosis* biovar *ovis*. **PeerJ**, v. 8, p. e9643, 26 ago. 2020.

ARSENAULT, J. et al. Prevalence of and carcass condemnation from maedi–visna, paratuberculosis and caseous lymphadenitis in culled sheep from Quebec, Canada. **Preventive Veterinary Medicine**, v. 59, n. 1–2, p. 67–81, maio 2003.

ASTON, C.; MISHRA, B.; SCHWARTZ, D. C. Optical mapping and its potential for large-scale sequencing projects. **Trends in Biotechnology**, v. 17, n. 7, p. 297–302, jul. 1999.

BADELL, E. et al. Improved quadruplex real-time PCR assay for the diagnosis of diphtheria. **Journal of medical microbiology**, v. 68, n. 10, p. 1455–1465, 2019.

BADELL, E. et al. Corynebacterium rouxii sp. nov., a novel member of the diphtheriae species complex. **Research in Microbiology**, v. 171, n. 3–4, p. 122–127, abr. 2020.

BARH, D. et al. In silico subtractive genomics for target identification in human bacterial pathogens. **Drug Development Research**, v. 72, n. 2, p. 162–177, 2011.

BECKLOFF, N. et al. Bacterial Genome Annotation. Em: ALI NAVID (Ed.). **Microbial Systems Biology - Methods and Protocols**. 1. ed. New York, Dordrecht, Heidelberg, London: Springer New York Dordrecht Heidelberg London, 2012. p. 471–503.

BERNARD, A. L.; FUNKE, G. Corynebacterium. Em: **Bergey's Manual of Systematic of Archaea and Bacteria (Online)**. London: John Wiley & Sons, Bergey's Manual Trust, 2015. p. 1–70.

BINNS, S. H.; GREEN, L. E.; BAILEY, M. Postal survey of ovine caseous lymphadenitis in the United Kingdom between 1990 and 1999. **Veterinary Record**, v. 150, n. 9, p. 263–268, mar. 2002.

CENTRO DE INTELIGÊNCIA E MERCADO DE CAPRINOS E OVINOS. LINFADENITE CASEOSA (LC). **EMBRAPA CAPRINOS E OVINOS**.

CHRISTENSEN, H. (ED.). **Introduction to Bioinformatics in Microbiology**. Cham: Springer International Publishing, 2018.

CHRISTENSEN, H.; OLSEN, J. E. Sequence-Based Classification and Identification of Prokaryotes. Em: CHRISTENSEN, H. (Ed.). **Introduction to Bioinformatics in Microbiology**. 1. ed. Switzerland: Springer Nature Switzerland AG, 2018a. p. 121–134.

CHRISTENSEN, H.; OLSEN, J. E. Sequenced-based typing of prokaryotes. Em: CHRISTENSEN, H. (Ed.). **Introduction to Bioinformatics in Microbiology**. 1. ed. Switzerland: Springer Nature Switzerland AG, 2018b. p. 189–203.

CHRISTENSEN, H.; OLSEN, J. E. Short Introduction to Phylogenetic Analysis of Molecular Sequence Data. Em: CHRISTENSEN, H. (Ed.). **Introduction to Bioinformatics in Microbiology**. 1. ed. Switzerland: Springer Nature, 2018c. p. 103–120.

CONNOR, K. M. et al. Characterization of United Kingdom Isolates of *Corynebacterium pseudotuberculosis* Using Pulsed-Field Gel Electrophoresis. **Journal of Clinical Microbiology**, v. 38, n. 7, p. 2633–2637, jul. 2000.

D'AFONSECA, V. et al. A description of genes of Corynebacterium pseudotuberculosis useful in diagnostics and vaccine applications. **Genetics and Molecular Research**, v. 7, n. 1, p. 252–260, 2008.

DANGEL, A. et al. Corynebacterium silvaticum sp. nov., a unique group of NTTB corynebacteria in wild boar and roe deer. **International Journal of Systematic and Evolutionary Microbiology**, 2020.

DAZAS, M. et al. Taxonomic status of Corynebacterium diphtheriae biovar Belfanti and proposal of Corynebacterium belfantii sp. nov. **International Journal of Systematic and Evolutionary Microbiology**, v. 68, n. 12, p. 3826–3831, 1 dez. 2018.

DE SÁ, P. H. C. G. et al. GapBlaster—A Graphical Gap Filler for Prokaryote Genomes. **PLOS ONE**, v. 11, n. 5, p. e0155327, 12 maio 2016.

DOHM, J. C. et al. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. **Nucleic Acids Research**, v. 36, n. 16, 1 set. 2008.

DORELLA, F. A. et al. Corynebacterium pseudotuberculosis : microbiology, biochemical properties, pathogenesis and molecular studies of virulence. **Veterinary Research**, v. 37, n. 2, p. 201–218, mar. 2006a.

DORELLA, F. A. et al. *Corynebacterium pseudotuberculosis* : microbiology, biochemical properties, pathogenesis and molecular studies of virulence. **Veterinary Research**, v. 37, n. 2, p. 201–218, mar. 2006b.

FERREIRA, A. C. et al. Whole-genome mapping reveals a large chromosomal inversion on Iberian Brucella suis biovar 2 strains. **Veterinary Microbiology**, v. 192, p. 220–225, ago. 2016.

FIETTO, L. G.; LAMËGO, M. R. DE A. História e importância da genômica. Em: MOREIRA, L. M. (Ed.). **Ciências genômicas: fundamentos e aplicações**. 1. ed. Ribeirão Preto: Sociedade Brasileira de Genética, 2015. p. 21–26.

FLUSBERG, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. **Nature Methods**, v. 7, n. 6, p. 461–465, 9 jun. 2010.

GIBSON, G.; MUSE, S. **A Primer of Genome Science Third Edition**. [s.l: s.n.].

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333–351, 17 jun. 2016.

GUIZELINI, D. et al. GFinisher: a new strategy to refine and finish bacterial genome assemblies. **Scientific Reports**, v. 6, n. 1, p. 34963, 10 dez. 2016.

GUO, J. et al. Four-color DNA sequencing with 3′- *O* -modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. **Proceedings of the National Academy of Sciences**, v. 105, n. 27, p. 9145–9150, 8 jul. 2008.

HASSAN, S. S. et al. Whole-genome sequence of Corynebacterium pseudotuberculosis strain Cp162, isolated from camel. **Journal of Bacteriology**, v. 194, n. 20, p. 5718–5719, 2012a.

HASSAN, S. S. et al. **Whole-genome sequence of Corynebacterium pseudotuberculosis strain Cp162, isolated from camel**. **Journal of Bacteriology**, out. 2012b.

HILL, C. Virulence or Niche Factors: What's in a Name? **Journal of Bacteriology**, v. 194, n. 21, p. 5725–5727, 1 nov. 2012.

IBRAIM, I. C. et al. Transcriptome profile of Corynebacterium pseudotuberculosis in response to iron limitation. **BMC Genomics**, v. 20, n. 1, p. 663, 20 dez. 2019.

JAIN, C. et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. **Nature Communications**, v. 9, n. 1, p. 5114, 30 dez. 2018.

JAIN, M. et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. **Genome Biology**, v. 17, n. 1, p. 239, 25 dez. 2016.

JU, J. et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. **Proceedings of the National Academy of Sciences**, v. 103, n. 52, p. 19635–19640, 26 dez. 2006.

JUNQUEIRA, D. M.; BRAUN, R. L.; VERLI, H. Alinhamentos. Em: VERLI, H. (Ed.). **Bioinformática da Biologia à flexibilidade molecular**. 1. ed. Porto Alegre: Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014. p. 39–71.

KOTEWICZ, M. L. et al. Optical maps distinguish individual strains of Escherichia coli O157 : H7. **Microbiology**, v. 153, n. 6, p. 1720–1733, 1 jun. 2007.

KOTEWICZ, M. L. et al. Optical mapping and 454 sequencing of Escherichia coli O157:H7 isolates linked to the US 2006 spinach-associated outbreak. **Microbiology**, v. 154, n. 11, p. 3518–3528, 1 nov. 2008.

KUMAR JAISWAL, A. et al. An In Silico Identification of Common Putative Vaccine Candidates against Treponema pallidum: A Reverse Vaccinology and Subtractive Genomics Based Approach. **International Journal of Molecular Sciences**, v. 18, n. 2, p. 1–15, 14 fev. 2017.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, 4 abr. 2012.

LEHRI, B.; SEDDON, A. M.; KARLYSHEV, A. V. The hidden perils of read mapping as a quality assessment tool in genome sequencing. **Scientific Reports**, v. 7, n. 1, p. 43149, 22 fev. 2017.

LESK, A. M. Alinhamentos e árvores filogenéticas. Em: LESK, A. M. (Ed.). **Introdução à Bioinformática**. 2. ed. Porto Alegre: Artmed, 2008. p. 177–238.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 15 jul. 2009.

LIU, L. et al. Comparison of Next-Generation Sequencing Systems. **Journal of Biomedicine and Biotechnology**, v. 2012, p. 1–11, 2012.

LOMAN, N. J. et al. Performance comparison of benchtop high-throughput sequencing platforms. **Nature Biotechnology**, v. 30, n. 5, p. 434–439, 22 maio 2012.

LOMAN, N. J.; PALLEN, M. J. Twenty years of bacterial genome sequencing. **Nature Reviews Microbiology**, v. 13, n. 12, p. 787–794, 2015.

MAIDEN, M. C. J. et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. **Proceedings of the National Academy of Sciences**, v. 95, n. 6, p. 3140–3145, 17 mar. 1998.

MAXAM, A. M.; GILBERT, W. A new method for sequencing DNA. **Proceedings of the National Academy of Sciences**, v. 74, n. 2, p. 560–564, fev. 1977.

MCKEAN, S. C.; DAVIES, J. K.; MOORE, R. J. Expression of phospholipase D, the major virulence factor of Corynebacterium pseudotuberculosis is regulated by multiple environmental factors and plays a role in macrophage death. **Microbiology**, v. 153, n. 7, p. 2203–2211, 2007.

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

MEIER-KOLTHOFF, J. P. et al. Genome sequence-based species delimitation with confidence intervals and improved distance functions. **BMC Bioinformatics**, v. 14, 2013.

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

NAGARAJAN, N.; POP, M. Sequence assembly demystified. **Nature Reviews Genetics**, v. 14, n. 3, p. 157–167, 2013.

NURK, S. et al. The complete sequence of a human genome. **Science**, v. 376, n. 6588, p. 44–53, abr. 2022.

OLIVEIRA, A. et al. Corynebacterium pseudotuberculosis may be under anagenesis and biovar Equi forms biovar Ovis: a phylogenic inference from sequence and structural analysis. **BMC Microbiology**, v. 16, n. 1, p. 100, 2 dez. 2016.

ONMUS-LEONE, F. et al. Enhanced De Novo Assembly of High Throughput Pyrosequencing Data Using Whole Genome Mapping. **PLoS ONE**, v. 8, n. 4, p. e61762, 17 abr. 2013.

ORAKOV, A. et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. **Genome Biology**, v. 22, n. 1, p. 178, 13 dez. 2021.

OREN, A.; GARRITY, G. M. Valid publication of the names of forty-two phyla of prokaryotes. **International Journal of Systematic and Evolutionary Microbiology**, v. 71, n. 10, 20 out. 2021.

PACBIO. **SEQUENCE WITH CONFIDENCE SMRT ® sequencing-Delivering highly accurate long reads to drive discovery in life science**. [s.l: s.n.].

PALLEN, M. J.; WREN, B. W. Bacterial pathogenomics. **Nature**, v. 449, n. 7164, p. 835–42, 18 out. 2007.

PATON, M. et al. Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks. **Australian Veterinary Journal**, v. 81, n. 1–2, p. 91–95, jan. 2003.

PEONA, V. et al. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. **Molecular Ecology Resources**, v. 21, n. 1, p. 263–286, 10 jan. 2021.

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

RAMOS, R. T. et al. Analysis of quality raw data of second generation sequencers with Quality Assessment Software. **BMC Research Notes**, v. 4, n. 1, p. 130, 18 dez. 2011.

RAPPUOLI, R. Reverse vaccinology. **Current Opinion in Microbiology**, v. 3, n. 5, p. 445–450, 2000.

RAPPUOLI, R. et al. Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. **Journal of Experimental Medicine**, v. 213, n. 4, p. 469–481, 2016.

RIBEIRO, M. G. et al. **PUNÇÃO ASPIRATIVA COM AGULHA FINA NO DIAGNÓSTICO DO CORYNEBACTERIUM PSEUDOTUBERCULOSIS NA LINFADENITE CASEOSA CAPRINA**. Botucatu - SP: [s.n.].

RIEGEL, P. et al. Taxonomy of Corynebacterium diphtheriae and related taxa, with recognition of Corynebacterium ulcerans sp. nov. nom. rev. **FEMS microbiology letters**, v. 126, n. 3, p. 271–6, 1 mar. 1995.

RODRIGUEZ, R.; KRISHNAN, Y. The chemistry of next-generation sequencing. **Nature Biotechnology**, v. 41, n. 12, p. 1709–1715, 16 dez. 2023.

RONAGHI, M. et al. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. **Analytical Biochemistry**, v. 242, n. 1, p. 84–89, nov. 1996.

SAHA, S.; RAJASEKARAN, S. Efficient and scalable scaffolding using optical restriction maps. **BMC Genomics**, v. 15, n. S5, p. S5, 14 jul. 2014.

SAID, M. BEN; BENZARTI, M.; ABDELKADER, A. **Epidemiological and clinical studies of ovine caseous lymphadenitis**. [s.l: s.n.]. Disponível em: <https://www.researchgate.net/publication/8630027>.

SAMAD, A. H. et al. Mapping the genome one molecule at a time — optical mapping. **Nature**, v. 378, n. 6556, p. 516–517, nov. 1995.

SANGER, F. et al. Nucleotide sequence of bacteriophage φX174 DNA. **Nature**, v. 265, n. 5596, p. 687–695, fev. 1977.

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

SCHOCH, C. L. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. **Database**, v. 2020, 1 jan. 2020.

SCHWARTZ, D. C. et al. Ordered Restriction Maps of *Saccharomyces cerevisiae* Chromosomes Constructed by Optical Mapping. **Science**, v. 262, n. 5130, p. 110–114, out. 1993.

SELIM, S. A. Oedematous Skin Disease of Buffalo in Egypt. **Journal of Veterinary Medicine Series B**, v. 48, n. 4, p. 241–258, 24 maio 2001.

SHEPPARD, S. K.; GUTTMAN, D. S.; FITZGERALD, J. R. Population genomics of bacterial host adaptation. **Nature reviews. Genetics**, v. 19, n. 9, p. 549–565, 4 set. 2018.

SOLDATOS, T. G. et al. How to learn about gene function: text-mining or ontologies? **Methods**, v. 74, p. 3–15, mar. 2015.

SOUSA, T. DE J. et al. Re-sequencing and optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes. **Scientific Reports**, v. 9, n. 1, p. 1–11, 2019a.

SOUSA, T. DE J. et al. Re-sequencing and optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes. **Scientific Reports**, v. 9, n. 1, 1 dez. 2019b.

TAUCH, A.; BURKOVSKI, A. Molecular armory or niche factors: virulence determinants of Corynebacterium species. **FEMS Microbiology Letters**, v. 67, n. 2, p. fnv185, 7 out. 2015.

TERAB, A. M. A. et al. Pathology, bacteriology and molecular studies on caseous lymphadenitis in Camelus dromedarius in the Emirate of Abu Dhabi, UAE, 2015-2020. **PLOS ONE**, v. 16, n. 6, p. e0252893, 8 jun. 2021.

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

TETTELIN, H. et al. Comparative genomics: the bacterial pan-genome. **Current Opinion in Microbiology**, v. 11, n. 5, p. 472–477, out. 2008.

VALOUEV, A. et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. **Genome Research**, v. 18, n. 7, p. 1051–1063, jul. 2008.

YANG, Z.; RANNALA, B. Molecular phylogenetics: Principles and practice. **Nature Reviews Genetics**, v. 13, n. 5, p. 303–314, 2012.

APÊNDICE

APÊNDICE A – Linhas de comando

1. ncbi-datasets-cli

```
conda run -n datasets datasets download genome accession --assembly-version latest –
include genome --inputfile 2_assembly_ids.txt --filename genome_data.zip
```

2. CheckM2

```
conda run -n checkm2 checkm2 predict --threads $(nproc) -x fasta --input
3_genome_data/fasta --output-directory 4_checkm
```

3. GUNC

conda run -n gunc gunc run --threads $(nproc) --file_suffix fasta --input_dir 3_genome_data/fasta -r /mnt/4tb_1/gunc_db/gunc_db_progenomes2.1.dmnd --temp_dir gunc_temp --out_dir 5_gunc

4. GTDB-Tk

conda run -n gtdbtk gtdbtk classify_wf --cpus $(nproc) --extension .fasta --genome_dir 3_genome_data/fasta --out_dir 6_gtdbtk/gtdbtk

5. Panaroo

panaroo-qc -t $(nproc) -i 3_genome_data/fasta/*.fasta --graph_type all --ref_db refseq.genomes.k21s1000.msh -o 8_panaroo/qc

find $PWD/3_genome_data/gff/ -name *.gff | parallel "echo -e {}'\t'$PWD/3_genome_data/fasta/{/.}.fasta" > gff_fasta_paths_panaroo.tsv

panaroo -t $(nproc) --remove-invalid-genes --clean-mode sensitive -a core --core_threshold 0.95 --aligner mafft -i gff_fasta_paths_panaroo.tsv -o 8_panaroo/sensitive

panaroo -t $(nproc) --remove-invalid-genes --clean-mode strict -a core --core_threshold 0.95 --aligner mafft -i gff_fasta_paths_panaroo.tsv -o 8_panaroo/strict

6. IQ-TREE2

#Gerar alinhamento do genoma core

conda activate panaroo

panaroo -t $(nproc) --remove-invalid-genes --clean-mode strict -a core --core_threshold 0.95 --aligner mafft -i ../gff_fasta_paths_panaroo_tree.tsv -o panaroo_tree_rooted

#iqtree

iqtree2 -T $(nproc) -B 1000 -s panaroo_tree_rooted/core_gene_alignment_filtered.aln -o NCTC7910 -pre Cp_tree_rooted

7. PanViTa

python3 ~/panvita/panvita.py -card -vfdb -bacmet *.gbff

APÊNDICE B – Artigo em colaboração

Artigo de pesquisa

Título: Probiogenomic study suggests that Leuconostoc mesenteroides strains Isolated from Human Breast Milk are Potential Probiotics

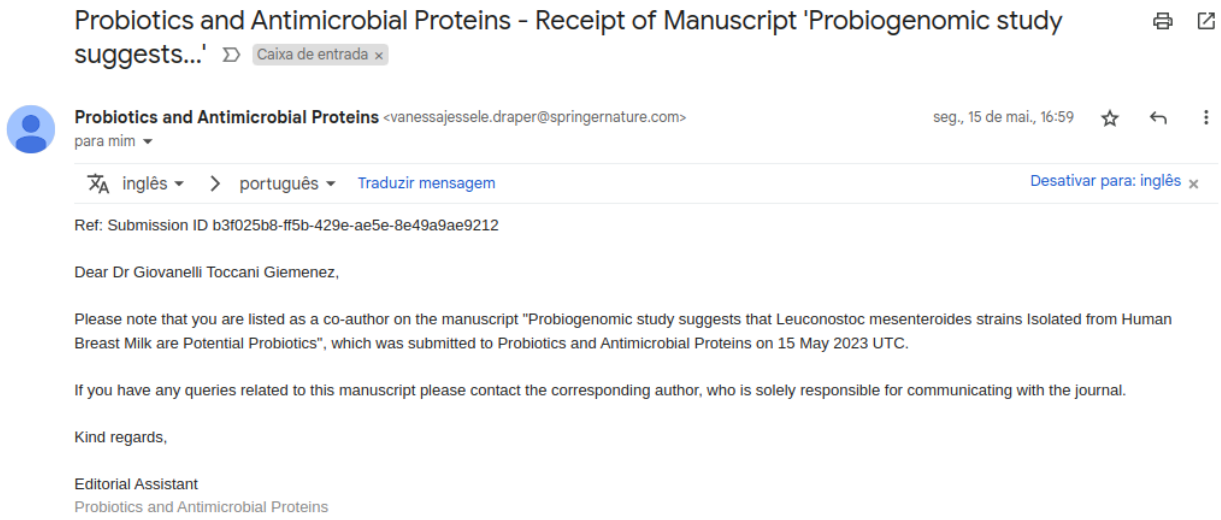
Autores: Juan Carlos Ariute[1], Nina Dias Coelho-Rocha[2], Carlos Willian Dias Dantas[3], Larissa Amorim Tourinho de Vasconcelos[2], Rodrigo Profeta[2], Thiago de Jesus Sousa[2], Ane de Souza Novaes[4], Bruno Galotti[4], Lucas Gabriel Gomes[2], Enrico Giovanelli Tacconi Gimenez[2], Carlos Diniz[1], Mariana Vieira Dias[1], Luís Cláudio Lima de Jesus[2], Arun Kumar Jaiswal[2], Sandeep Tiwari[5:], Rodrigo Carvalho[5], Ana Maria Benko-Iseppon[6], Bertram Brenig[7], Vasco Azevedo[2], Debmalya Barh[8], Flaviano S. Martins[4,*] and Flavia Aburjaile[1,*]

ID de Submissão: b3f025b8-ff5b-429e-ae5e-8e49a9ae9212

Revista: Probiotics and Atimicrobial Proteins

Contribuição: Montagem, anotação dos genomas e escrita do artigo

Ano: 2023



Probiotics and Antimicrobial Proteins - Receipt of Manuscript 'Probiogenomic study suggests...' Caixa de entrada ×

Probiotics and Antimicrobial Proteins <vanessajessele.draper@springernature.com> seg., 15 de mai., 16:59

para mim

inglês › português Traduzir mensagem Desativar para: inglês ×

Ref: Submission ID b3f025b8-ff5b-429e-ae5e-8e49a9ae9212

Dear Dr Giovanelli Toccani Giemenez,

Please note that you are listed as a co-author on the manuscript "Probiogenomic study suggests that Leuconostoc mesenteroides strains Isolated from Human Breast Milk are Potential Probiotics", which was submitted to Probiotics and Antimicrobial Proteins on 15 May 2023 UTC.

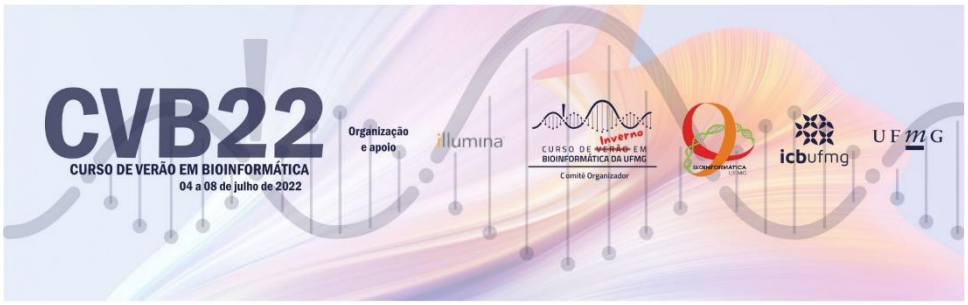
If you have any queries related to this manuscript please contact the corresponding author, who is solely responsible for communicating with the journal.

Kind regards,

Editorial Assistant
Probiotics and Antimicrobial Proteins

APÊNDICE C – Participação em eventos

1. VI Curso de Verão em Bioinformática da UFMG



Certificamos que

Enrico Giovanelli Tacconi Gimenez

participou do VI Curso de Verão em Bioinformática da UFMG com **30 horas** de duração, realizado pelo Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, no período de 04/07/2022 a 08/07/2022.

Glória Regina Franco
Coordenadora do Projeto da Escola de Verão em Bioinformática da UFMG

Aristóteles Góes Neto
Coordenador do Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

2. Minicurso Vacinologia Reversa – VI Curso de Verão em Bioinformática

Certificamos que

Enrico Giovanelli Tacconi Gimenez

participou do minicurso **Vacinologia Reversa**

do VI Curso de Verão em Bioinformática da UFMG com **4 horas** de duração, realizado pelo Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, no período de 04/07/2022 a 08/07/2022.

Glória Regina Franco
Coordenadora do Projeto da Escola de Verão em Bioinformática da UFMG

Aristóteles Góes Neto
Coordenador do Programa Interunidades de Pós-Graduação em Bioinformática da UFMG