UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Biológicas

Programa Interunidades de Pós-graduação em Bioinformática

Thiago de Jesus Sousa

OBTENÇÃO DE GENOMAS COMPLETOS DE ALTA QUALIDADE E A INFLUÊNCIA DESTES EM ANÁLISES GENÔMICAS: uma abordagem para *corynebacterium pseudotuberculosis*

Belo Horizonte

2020

Thiago de Jesus Sousa

OBTENÇÃO DE GENOMAS COMPLETOS DE ALTA QUALIDADE E A INFLUÊNCIA DESTES EM ANÁLISES GENÔMICAS: uma abordagem para *Corynebacterium pseudotuberculosis*

Tese apresentada como requisito parcial para a obtenção do Grau de Doutor pelo Programa de Pós-graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo

Coorientador: Prof. Dr. Siomar Castro Soares

Coorientadora: Dra. Anne Cybelle Pinto Gomide

Belo Horizonte

2020

043 Sousa, Thiago de Jesus.
Obtenção de genomas completos de alta qualidade e a influência destes em análises genômicas [manuscrito]: uma abordagem para *corynebacterium pseudotuberculosis* / Thiago de Jesus Sousa. – 2020.
138 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo. Coorientador: Prof. Dr. Siomar Castro Soares. Coorientadora: Dra. Anne Cybelle Pinto Gomide.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Corynebacterium pseudotuberculosis. 3 Sequenciamento de Nucleotídeos em Larga Escala. 4. Genômica. I. Azevedo, Vasco Ariston de Carvalho . II. Soares, Siomar Castro. III. Gomide, Anne Cybelle Pinto. IV. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. V. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Jéssica Patrícia Silva de Sá – CRB 6 - 3430

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DE DEFESA DE TESE

Às treze horas e trinta minutos do dia **30 de junho de 2020**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Obtenção de genomas completos de alta qualidade e a influência destes em análises genômicas: uma abordagem para Corynebacterium pseudotuberculosis"**, requisito para obtenção do grau de Doutor 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:

Prof./Pesq.	**Instituição**	**Indicação**
Dr. Vasco Ariston de Carvalho Azevedo- Orientador	UFMG	Aprovado
Dra. Anne Cybelle Pinto - Coorientador	UFMG	Aprovado
Dr. Siomar de Castro Soares - Coorientador	UFTM	Aprovado
Dr. Ricardo Portela	UFBA	Aprovado
Dr. Pedro Marcus Pereira Vidigal	UFV	Aprovado
Dr. Henrique Cesar Pereira Figueiredo	UFMG	Aprovado
Dr. Aristóteles Góes Neto	UFMG	Aprovado
Dr. Eric Roberto Guimarães Rocha Aguiar	UESC	Aprovado

Pelas indicações, o candidato foi considerado: Aprovado

O resultado final foi comunicado publicamente o 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, 30 de junho de 2020.

Documento assinado eletronicamente por **Vasco Ariston de Carvalho Azevedo**, **Professor do Magistério Superior**, em 30/06/2020, às 15:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Coordenador(a) de curso de pós-graduação**, em 30/06/2020, às 17:12, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **ERIC ROBERTO GUIMARAES ROCHA AGUIAR**, **Usuário Externo**, em 30/06/2020, às 17:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Ricardo Wagner Dias Portela**, **Usuário Externo**, em 30/06/2020, às 17:43, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Thiago de Jesus Sousa**, **Usuário Externo**, em 30/06/2020, às 17:52, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Pedro Marcus Pereira Vidigal**, **Usuário Externo**, em 30/06/2020, às 17:55, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Henrique Cesar Pereira Figueiredo**, **Subcoordenador(a)**, em 30/06/2020, às 18:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Siomar de Castro Soares**, **Usuário Externo**, em 30/06/2020, às 18:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Anne Cybelle Pinto Gomide**, **Usuário Externo**, em 30/06/2020, às 18:57, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

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 **0164284** e o código CRC **A7B3D668**.

AGRADECIMENTOS

Agradeço a toda equipe do Laboratório de Genética Celular e Molecular da UFMG por toda ajuda e apoio, tanto pessoal quanto acadêmico. Aos meus orientadores, o Prof. Dr. Vasco Azevedo e a Drª Anne Gomide.

Ao Prof. Dr. Rommel Ramos e todo grupo do Laboratório de Biologia Computacional da UFPA.

Ao Prof. Siomar Soares, Prof. Carlo, Prof. Marcus, Drª Leticia, ao mestrando Leandro e todo o grupo de Bioinformática e Imunologia da UFTM.

Ao Prof. Dr. Ricardo Portela, por ter acreditado em mim desde o começo da graduação na UFBA, e o qual sempre serei grato.

Ao Prof. Dr. Mateus Matiuzzi e ao seu grupo pelo empenho e esforço no isolamento de novas linhagens de *C. pseudotuberculosis* essenciais para este trabalho.

A todos os pós-doutorandos e colaboradores do LGCM, a Anne Cybelle, Flávia Aburjaile, Rodrigo Dias, Francielly Rodrigues, Rodrigo Kato, Marcus Viana, Sandeep Tiwari e Arun Kumar.

Ao grupo de Bioinformática do Laboratório de Genética Celular e Molecular, principalmente a Alessandra Lima, Diego Neres, Mariana e Doglas Parise e Rodrigo Profeta.

À Sheila, Thiago, Natalia e a Fernanda por toda atenção e apoio.

Aos amigos Felipe Pereira e Tarcísio Coutinho, bem como suas respectivas esposas Paula Pereira e Camila Franco, mesmo que distante sei que estão torcendo por mim.

As agências de fomento à Pesquisa, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

À toda minha família, em especial minha mãe Conceição, meu pai Israel, minha vó Maria Almeida, minha Madrinha Edenilce, minha tia Ilza e minha namorada Eula Graciele, por todo apoio familiar e suporte.

À Deus, pois ele está comigo a todo momento.

“Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer” (Alan Mathison Turing).

I – RESUMO

Corynebacterium pseudotuberculosis pertence ao grupo *Corynebacterium*, *Mycobacterium*, *Nocardia* e *Rhodococcus* (CMNR). Esta bactéria é um patógeno de grande relevância, por causar infecção em ruminantes. Entre as doenças causadas por *C. pseudotuberculosis*, destaca-se a linfadenite caseosa cujas formas de tratamento, imunoprofilaxia e diagnóstico têm sido insatisfatórias. Deste modo, torna-se necessária a identificação das bases moleculares dessa bactéria para o desenvolvimento de novas abordagens terapêuticas e profiláticas, tais como vacinas. Assim, o estudo por genômica comparativa de *C. pseudotuberculosis,* conduzido pelo nosso grupo de pesquisa, tem sido fundamental para compreender os mecanismos de virulência e patogenicidade desse micro-organismo. Contudo, devido às limitações tecnológicas e à indisponibilidade de ferramentas computacionais avançadas inerentes à época em que diversos isolados de *C. pseudotuberculosis* foram sequenciados, estes apresentam erros de montagem. Portanto, neste trabalho empregamos o mapa óptico cromossômico, que utiliza enzimas de restrição ao longo de todo o genoma para auxiliar no processo de ordenação dos *contigs*. Por meio dessa, abordagem foi possível obter dados genômicos de alta qualidade e acurácia para serem utilizados como referência, possibilitando maior precisão e confiabilidade nas análises ômicas estruturais e funcionais desta espécie. Assim, onze linhagens representantes dos biovares *ovis* e *equi* foram sequenciadas e re-sequenciadas pelas plataformas Ion Torrent PGM™ e Illumina Hiseq 2500 com *paired-end*, e submetidas também à técnica do mapa óptico. A ordenação dos contigs com base nas informações do mapa óptico permitiu detectar erros de montagem e inversões genômicas em seis linhagens. Além disso, foi identificado que o *cluster* de genes *nar*, que diferencia os biovares da espécie, não estava presente em quatro linhagens depositadas do biovar equi*,* além de um rearranjo cromossômico flanqueado por transposases na *C. pseudotuberculosis* 162. Esses genomas completos e acurados foram essenciais como referência para a montagem de 50 genomas de *C. pseudotuberculosis* biovar *ovis.* Ao analisar esses genomas foi possível evidenciar a ausência do gene da *cutinase like family* na *linhagem 1002* e em seus respectivos mutantes. Além disso, são descritas variantes nucleotídicas pontuais ainda não relatadas nesta espécie, as quais podem indicar a influência de pressões ambientais no genoma. O presente trabalho foi capaz de contribuir para a qualidade dos genomas e ampliar o conhecimento molecular de *C. pseudotuberculosis,* atualizando informações sobre o pan-genoma da espécie.

Palavras-chave: *Corynebacterium pseudotuberculosis.* Ion Torrent. Illumina Hiseq. Mapa óptico. Rearranjo Cromossômico. Genômica Comparativa.

II – ABSTRACT

Corynebacterium pseudotuberculosis belongs to the CMNR group (*Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rhodococcus*). This bacterium has its most significant relevance within veterinarian medicine, the causative agent of infections in ruminants. Among the diseases caused by *C. pseudotuberculosis*, we can highlight Caseous Lymphadenitis whose treatment methods are unsatisfactory. Therefore, it is necessary to identify the molecular bases of this pathogen for the development of new therapeutic and prophylactic approaches, such as vaccines. Thus, the comparative study genomics of *C. pseudotuberculosis* by our research group has been essential for understanding their mechanisms of virulence and pathogenicity. However, because of the old technology limitations and the unavailability of advanced computational methods at the time when several isolates were sequenced, incomplete genomes have been generated. Therefore, in this work, we use the chromosomal optical mapping, which uses restriction enzymes throughout the entire genome to assist in contigs scaffolding. Through this approach, aimed to generate high-quality genomic data, as reference sequences to provide greater precision and reliability for molecular structural and functional analysis of this species. So, 11 strains representing the biovar *ovis* and *equi* were sequenced in Ion Torrent PGM™ and re-sequenced in Illumina Hiseq 2500 platforms using paired-end library, and after submitting to the optical map technique. The scaffolding process of the contigs by the optical map information allowed to detect assembly errors and genomic inversions in six strains. Our results revealed that the *nar* cluster, which determinate the biovar type, was not present in 4 previously deposited strains. It was possible to identify a real chromosomal rearrangement flanked by transposases in *C. pseudotuberculosis* 162. These complete and accurate genomes were essential for the assembly of 50 genomes of *C. pseudotuberculosis* biovar *ovis*. When analyzing these genomes, it was possible to evidence the absence of the "cutinase like family gene" was revealed in *C. pseudotuberculosis* 1002 and 6 more mutant genome strains. Also, point nucleotide variants previously unreported in this species are described, which may show the influence of environmental pressures on the genome. The present work could contribute to the quality of the genomes and expand the molecular knowledge of *C. pseudotuberculosis*, updating information about the pan-genome of the species.

Keywords: *Corynebacterium pseudotuberculosis*. Ion Torrent PGM. Illumina Hiseq 2500. Optical map. Misassemblies. Chromosomal rearrangement. Comparative Genomics.

III - LISTA DE ILUSTRAÇÕES

IV - LISTA DE TABELAS

V - LISTA DE ABREVIAÇÕES

A	Adenina
BLAST	Do inglês, *Basic local alignment search tool.*
C	Citosina
dATP	Do inglês, *Deoxyadenosine triphosphate*
dCTP	Do inglês, *Deoxycytidine triphosphate*
dGTP	Do inglês, *Deoxyguanosine triphosphate*
DNA	Ácido desoxirribonucleico
dTTP	Do inglês, *Deoxythymidine triphosphate*
G	Guanina
Gb	Gigabase
GBK	Do inglês, *GenBank*
GC	Guanina + Citosina
GSSs	Do inglês, *Genomic Survey Sequences*
Kb	Quilo base
LC	Linfandenite caseosa
LU	Linfandenite ulcerativa
LGCM	Laboratório de Genética Celular e Molecular
Mb	Megabase
MPS	Do inglês, *Massive parallel sequencing*
NCBI	Do inglês, *National Center for Biotechnology Information*
NGS	Do inglês, *Next-Generation Sequencing*
OLC	Do inglês, *Overlap-Layout-Consensus*
ORF	Do inglês, *Open Reading Frame*
PacBio	Do inglês, *Pacific Biosciences*
pb	Pares de base
PCR	Reação em cadeia da polimerase
PGAP	Do inglês, *Prokaryotic Genome Annotation Pipeline*
PGM™	Do inglês, *Personal Genome Machine™*
pH	Potencial Hidrogeniônico
RNA	Ácido Ribonucleico
rRNA	RNA Ribossômico
SNPs	Do inglês, *Single-Nucleotide Polymorphism*
SOLiD™	Do inglês, *Sequencing by Oligonucleotide Ligation and Detection*
T	Timina
tRNA	RNA transportador
WGM	Do inglês, *Whole Genome Mapping*
WGS	Do inglês, *Whole Genome Shotgun*
μm	Micrômetro

SUMÁRIO

1. DELINEAMENTO DA TESE

A tese está dividida nas seções descritas abaixo, com destaque para Capítulo I e II.

1) **Prefácio;**
2) **Objetivo geral;**
3) **Capítulo I:**
    a. **Objetivos específicos**;
    b. **Introdução**: A introdução apresenta conceitos sobre sequenciamento, montagem de genomas, mapa óptico, estratégias para fechamento de *gaps*, anotação funcional e estrutural, características microbiológicas, sequenciamento e genômica de *C. pseudotuberculosis*.
    c. **Artigo I:** "*Re-sequencing and optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes*". Esse artigo explora a atualização de genomas de *C. pseudotuberculosis*, com o foco na aplicação da técnica do mapa óptico para detecção de erros de montagem e finalização de genomas acurados e completos.
    d. **Discussão ampliada do Artigo I;**
4) **Capítulo II:**
    a. **Objetivos específicos**;
    b. **Introdução:** Conceitos sobre genômica comparativa, Pan-genômica e programas de análises genômicas.
    c. **Capítulo de livro**: Capítulo 5: "*Pan-genomics of veterinary bacteria and its applications"*, ao qual apresentamos uma revisão da literatura sobre estudos genômicos sobre as principais bactérias de interesse médico veterinário.
    d. **Artigo II:** "*New genomics discoveries through re-sequencing and correction of assembly in Corynebacterium pseudotuberculosis genomes*".
5) **Conclusões;**
6) **Perspectivas;**
7) **Referências**;
8) **Apêndices;**

2. PREFÁCIO

A ciência evolui a partir de questionamentos e resultados fundamentados no método científico. Em 1953, James Watson, Francis Crick e Rosalind Franklin questionaram o modelo de tripla hélice da estrutura molecular do DNA proposto por Linus Pauling. Seu modelo de dupla hélice conseguiu se sobrepor e estabelecer diversos paradigmas, culminando no dogma central da biologia molecular, ou dogma de Crick (PRAY, 2008). Dessa forma, a evolução do conhecimento na área passou por estudos com enzimas de restrição, pela PCR (Reação em Cadeia da Polimerase) inventada por Kary Mullis, pelo sequenciamento por Sanger e suas variações automatizadas.

Desde 1970, os biólogos Paulien Hogeweg e Ben Hesper (PAULIEN HOGEWEG, 2011), começaram a estabelecer a interseção multidisciplinar de áreas que dariam origem a "bioinformática", com o intuito de integrar abordagens envolvendo matemática, física, estatística, ciência da computação, engenharia e biologia. Desde então, a bioinformática passou a fundamentar diversas outras áreas entre as ciências biológicas e exatas, tais como *big data*, modelagem de sistemas biológicos, genômica estrutural, funcional e comparativa, análise de expressão gênica, biologia sintética, e várias outras aplicações que estudam DNA, RNA, proteínas e metabólitos.

Entre 2003 a 2005, as plataformas de sequenciamento chamadas de NGS ou MPS (*Next Generation Sequencing* ou *Massive Parallel Sequencing*) começaram a ficar disponíveis para a comunidade científica. Com o apogeu dessas plataformas, o sequenciamento de genomas ficou cada vez mais acessível e viável, possibilitando estudos de vários organismos, tais como os estudos em pan-genômica, levando a um grande volume de informações de indivíduos de todos os reinos, mas principalmente, de bactérias.

Apesar de todas as tecnologias de sequenciamento e das montagens aplicadas nos genomas para garantir a precisão dos dados, existem evidências de genomas com erros de montagem nos bancos de dados genômicos. Uma extração de DNA com muita degradação do material genético, passando por um sequenciamento que gera leituras com valor de *Phred* menor que 20 e uma montagem feita sem planejamento e programas adequados, leva à dificuldade na construção dos *contigs* e às incertezas na finalização dos genomas. Um erro de montagem pode resultar na exclusão de conteúdo gênico, inversão de regiões e ausência de genes importantes, não refletindo o conteúdo gênico real do organismo, o que pode dificultar a seleção de alvos para diagnóstico e vacinas.

O avanço da tecnologia NGS contribuiu para o estudo genômico de milhares de organismos até então ainda não sequenciados. Contudo, não foi dada a devida atenção aos erros presentes nesses dados, principalmente no âmbito de erros de montagem dos genomas. A grande maioria dos genomas depositados no *GenBank* representa apenas versões *drafts* (rascunhos) ou incompletas. A determinação das sequências completas dos genomas é fundamental para entender a biologia e a função de um organismo. Mesmo os genomas de organismos relativamente pequenos, como bactérias (até 10 milhões de bases), geralmente são submetidos como *drafts*. O problema decorre da falta de uma ferramenta universal e confiável que permita a montagem automática dos *contigs*, principalmente com sequências contendo regiões repetitivas longas (LEHRI; SEDDON; KARLYSHEV, 2017).

Há estudos na literatura científica com relatos de que trabalhos de montagem de genomas utilizando diferentes estratégias são mais confiáveis e mais acurados, permitindo assim que genomas possam ser bem explorados em estudos de evolução do patógeno, influência da pressão seletiva no genoma, e nas comparações entre linhagens e análises de genômica estrutural. Também podem levar à identificação de alvos direcionados para desenvolvimento de profilaxias e diagnósticos mais eficientes.

Ao longo dos anos, estratégias utilizando dados oriundos do mesmo organismo foram desenvolvidas para auxiliar na qualidade da montagem dos genomas. O mapa óptico permite a organização de *contigs* e a estimativa do tamanho dos espaços e suas posições; outra estratégia é o sequenciamento com leituras longas (*long reads)* ou leituras curtas (*short reads)* com biblioteca pareada (*paired-end)*, sendo que essas estratégias são mais confiáveis (WEISSENSTEINER *et al.*, 2017). Outra opção é utilizar um genoma completo de excelente qualidade de outro organismo da mesma espécie com similaridade maior que 95%, para assim evitar erros de montagem e incoerências na ordem do conteúdo gênico. Contudo, esta alternativa sofre influência da plasticidade genômica e de processos de transferência horizontal, principalmente em bactérias. Além disso, pode levar a um viés na montagem dos genomas. Deste modo, uma combinação de vários métodos usados para sequenciamento e montagem pode levar à obtenção de um genoma de alta qualidade (GHURYE; POP, 2019).

Em 2005, a genômica comparativa começou a ganhar destaque e amplitude, devido à redução dos custos de sequenciamento proporcionado pelas plataformas NGS, especialmente no âmbito da pan-genômica introduzida por Tettelin (TETTELIN *et al.*, 2005) ao estudar o total dos genes dos organismos de um conjunto de genomas por meio da similaridade gênica e identificação de ortólogos.

A abordagem da pan-genômica consiste em descrever o conjunto completo de genes de um determinado organismo que, quando comparado com mais linhagens, nos possibilita a identificação do genoma central (*Core genome)*, que constitui o conjunto de genes compartilhados por todos os organismos analisados. Os demais genes, denominados dispensáveis são genes compartilhados por mais de um organismo, mas não por todos os organismos (TETTELIN *et al.*, 2005). No decorrer das pesquisas esse conceito de genes dispensáveis foi segmentado em genes únicos (*singletons genome*) e genoma acessório (*shared genome*) (MEDINI *et al.*, 2005).

No âmbito das pesquisas genômicas relacionadas com *C. pseudotuberculosis*, nosso grupo de pesquisa iniciou estudos a partir do sequenciamento de linhagens dessa espécie, resultando nas análises de Pan-genômica propostas por Soares e colaboradores em 2013 (SOARES, SC *et al.*, 2013). Os genes do genoma central podem ser utilizados no desenvolvimento de novas vacinas ou a descoberta de novos alvos para drogas, através das abordagens de vacinologia reversa e ancoragem molecular (*molecular docking*), respectivamente. Apesar das técnicas de produção animal, considerando o *C. pseudotuberculosis*, como as ações epidemiológicas de monitoramento e as boas práticas de manejo dos animais serem usadas, não se tem garantia da sua correta aplicação. Assim, a técnica de prevenção para doenças infecciosas mais eficiente é a vacinação, mas não há no mercado uma vacina para *C. pseudotuberculosis* com resultados eficientes e de qualidade comercial.

Desde o século 18 que essa abordagem vem sendo desenvolvida, como é o caso da vacina da varíola por Edward Jenner (AMERICA; AMHERST; JENNER, 2005). Na vacinologia clássica, uma das formas de confecção da vacina é baseada na identificação a partir de um conjunto de moléculas do patógeno contendo componentes da parede celular e/ou moléculas secretadas, geralmente isoladas no local da infecção nos animais (RAPPUOLI, RINO, 2007). No contexto de *C. pseudotuberculosis* a fosfolipase-D (*pld*) é um grande exemplo de fator de virulência com ação de exotoxina (SONGER, 1997), usada como alvo de diagnóstico e vacina (TACHEDJIAN *et al.*, 1995). Também já foi alvo a proteína CP40, uma protease imunogênica de 40 kDa secretada sendo reconhecida pelos soros de animais infectados durante a infecção precoce (WALKER *et al.*, 1994).

Com o intuito de identificar alvos de forma mais rápida começou, no início do século XXI, a se identificar alvos com base no genoma do patógeno, após o começo da era genômica. Essa triagem computacional favorece a diminuição do tempo e do custo na busca de potenciais candidatos vacinais, quando comparada ao método tradicional (VILELA RODRIGUES *et al.*, 2019). Sendo assim, Rino Rappuoli (2010) apresentou essa nova proposta, a “vacinologia

reversa", que tem como princípio identificar proteínas que são expostas ao hospedeiro, pois são as mais prováveis de serem reconhecidas pelo sistema imune (RAPPUOLI, R, 2001; SETTE; RAPPUOLI, 2010). Essa abordagem teve seu primeiro grande sucesso na vacina 4CMenB para a *Neisseria meningitidis serogroup B*, causadora da meningite bacteriana em humanos, cujos componentes foram minerados pela estratégia da vacinologia reversa (RAPPUOLI, RINO *et al.*, 2018).

Já as proteínas que são preditas como alvos citoplasmáticos do patógeno passam pela abordagem de ancoragem molecular, com o intuito de identificar possíveis alvos de fármacos para essas estruturas proteicas. Esse direcionamento para proteínas citoplasmáticas é feito devido a participação delas em vias metabólicas importantes para o patógeno e por isso são alvos de drogas. Essa estratégia insere-se nesse contexto para reduzir o número de testes de alvos para drogas, pois tem como base prever os modos de ligação e dos detalhes do reconhecimento molecular entre receptor-ligante. Essa abordagem *in silico* permite a identificação de forma mais rápida e com menor custo de um promissor candidato para alvos de fármacos (THOMSEN; CHRISTENSEN, 2006).

Após 20 anos do início da era NGS, a tecnologia de sequenciamento evoluiu bastante e vem contribuindo para o avanço científico de forma acelerada. Esta evolução atingiu tal ponto que o maior desafio, outrora de se obter um sequenciamento de genomas completos, passou a ser a curadoria da qualidade dos dados e a análise de seu grande volume. Nessa situação, a bioinformática entra como principal recurso por meio de abordagens que mesclam a informação biológica com grafos e inteligência computacional. No âmbito da curadoria dessas informações, podemos citar a prática de sequenciar o mesmo organismo em diferentes plataformas, ou ainda de utilizar diferentes métodos concomitantemente, como, por exemplo, o mapa óptico, que utiliza enzimas de restrição para determinar o padrão da ordem gênica daquela linhagem em específico (OLSEN *et al.*, 2015). Com base nos motivos apresentados idealizamos este trabalho, com o objetivo principal de atualizar os genomas depositados com tecnologias mais recentes e assim levantar a discussão a respeito da qualidade dos genomas depositados no *National Center for Biotechnology Information (*NCBI) e a influência dos erros de montagens em análises de genômica comparativa.

O LGCM, há mais de 15 anos, realiza estudos experimentais com a espécie *C. pseudotuberculosis*, sendo que nos últimos anos os esforços foram concentrados no campo de estudos da Bioinformática.

2.1 COLABORADORES

Este trabalho foi realizado no Laboratório de Genética Celular e Molecular (LGCM) do instituto de Ciências Biológicas (ICB) da Universidade Federal de Minas Gerais (UFMG) coordenado pelo Prof. Vasco Ariston de Carvalho Azevedo. Contamos com vários colaboradores:

a) Prof. Dr. Henrique César Pereira Figueiredo, professor da Escola de Veterinária, Departamento de Medicina Veterinária Preventiva e coordenador do Laboratório (AQUACEN) pela UFMG.

b) Prof. Dr. Artur Silva, pesquisador e professor do Laboratório de Polimorfismos de DNA do Instituto de Ciências Biológicas da UFPA.

c) Prof. Dr. Rommel Thiago Jucá Ramos, pesquisador e professor do Laboratório de Polimorfismos de DNA do Instituto de Ciências Biológicas da UFPA.

d) Prof. Dr. Bertram Brenig, pesquisador e professor do Departamento de Biologia Molecular e Pecuária da Universidade de Göttingen - Alemanha.

e) Prof. Dr. Siomar Soares, professor do Departamento de Microbiologia, Imunologia e Parasitologia e pesquisador do Laboratório de ImunoBioinformática da Universidade Federal do Triângulo Mineiro.

f) Prof. Dr. Mateus Matiuzzi da Costa, professor da Universidade Federal do Vale do São Francisco.

3. OBJETIVO GERAL

- Implementar estratégias para a melhoria de montagens de genomas bacterianos, com o intuito de obter genomas acurados e de maior qualidade, tornando possível a atualização dos bancos de dados e a realização novas análises de pan-genômica de *C. pseudotuberculosis*.

CAPÍTULO I

4. CAPÍTULO I

4.1 Objetivos específicos do capítulo I

- Utilizar a estratégia do mapa óptico para ordenação de quatro novos genomas e correção de seis genomas já depositados de *C. pseudotuberculosis* pertencentes aos biovares *ovis* e *equi*.
- Obter genomas completos de alta acurácia e precisão para utilizá-los como referência na ordenação de novas montagens de novas linhagens de *C. pseudotuberculosis*.

4.2 Estrutura do capítulo I

A primeira parte do capítulo descreve conceitos sobre sequenciamento de DNA, tais como: montagem por referência e *ab initio;* ordenação dos genomas após montagem; estratégias para finalização pelo mapa óptico e por referência; anotação estrutural e funcional; características microbiológicas; e os estudos genômicos de *C. pseudotuberculosis*.

A segunda parte apresenta o Artigo I – ***Re-sequencing and Optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes.*** O artigo em destaque dessa seção foi publicado na revista *Scientific Reports* (*Nature Research*), fator de impacto 4,122 (2018).

Em seguida, uma sessão de discussão ampliada do artigo I contendo resultados que não foram publicados.

4.3 Sequenciamento de DNA

Em meados da década de 70, dois grupos propuseram soluções para o desafio de sequenciar DNA, Alan Coulson e Frederick Sanger com o sistema de "*Plus and Minus*"; e Allan Maxam e Walter Gilbert (MAXAM; GILBERT, 1977) com a técnica de clivagem química, ambas através da eletroforese de gel de poliacrilamida.

Entretanto, essa primeira geração tem seu apogeu em 1977, quando o bioquímico Britânico Frederick Sanger e colaboradores publicaram a técnica "*Chain-termination*" ou método de terminação da cadeia nucleotídica (SANGER; NICKLEN; COULSON, 1977). Essa técnica tinha como fundamento a incorporação de um didesoxiribonucleotídeo, assim impedindo o alongamento da cadeia e após revelação dos fragmentos em gel de eletroforese, era possível realizar a leitura e identificação da sequência de DNA presente naquele fragmento.

O método de Sanger foi utilizado extensivamente durante 40 anos. Em 1995, o primeiro genoma completo bacteriano sequenciado foi descrito, sendo da espécie bacteriana *Haemophilus influenzae* (FLEISCHMANN *et al.*, 1995). A partir desse evento, diversos trabalhos envolvendo estudos genômicos de transferência horizontal de genes, redução de genomas, identificação de alvos para drogas e vacinologia reversa foram realizados (RAPPUOLI, R, 2001; RAPPUOLI, RINO *et al.*, 2016). Estes eventos estabeleceram a primeira geração de sequenciadores, nomeada de "*Whole-genome shotgun sequencing*".

O projeto de sequenciamento do genoma humano em meados dos anos 90 foi um estímulo crucial para desenvolvimento de novas metodologias de sequenciamento com menor custo, menor tempo de execução e maior geração de dados brutos de genomas inteiros. Empresas como a *Life Sciences* (454 Roche), *Solexa Genome Analyzer* (Illumina Miseq e Hiseq) e a *Life Technologies* (Ion Torrent) fomentaram o início da segunda geração de plataformas de sequenciamento, nomeada de *Next Generation Sequencing* (NGS), entre 2005 e 2011 (VAN DIJK *et al.*, 2014).

Como principais representantes da segunda geração, temos o MiSeq e HiSeq (Illumina) e o Ion Torrent (PGM™ e S5™ pela Thermo Fisher Scientific), ambos gerando leituras pequenas. A metodologia aplicada pela Illumina tem como base a síntese da molécula de DNA pela polimerase com nucleotídeos sintéticos marcados com fluoróforos em uma célula de fluxo que funciona como uma matriz sólida. No caso do sequenciamento com *paired-end*, o comprimento do fragmento entre as duas sequências do adaptador é definido como tamanho de inserção (VOELKERDING; DAMES; DURTSCHI, 2009).

Já a tecnologia presente no Ion Torrent PGM^{TM} utiliza um sistema de detecção de próton (H^+) resultante da reação de polimerização do DNA para detecção das bases nucleotídicas. Devido à presença do íon H^+, ocorre a alteração do pH do meio, e sensores detectam essa variação e a convertem em sinal elétrico para o equipamento.

Subsequentemente, o sequenciador adiciona as bases nucleotídicas artificiais (dCTP, dATP, dGTP, dTTP) uma de cada vez, ou seja, a cada lavagem, e identifica a variação do potencial hidrogeniônico (pH). Ao final de cada etapa, todos os sinais são processados e as sequências são identificadas pelo conceito da complementaridade, cujo princípio é a ligação entre as bases adenina com timina (A+T), e guanina com citosina (G+C), por duas e três interações de hidrogênio, respectivamente (VOELKERDING; DAMES; DURTSCHI, 2009).

A terceira geração de sequenciadores emergiu com a proposta de "*Single-molecule*" (Sequenciamento de molécula única), gerando leituras longas. A plataforma PacBio RS (*Pacific Biosciences*) (EID *et al.*, 2009) baseia-se no sequenciamento de molécula única em tempo real, por meio de uma tecnologia inovadora que dispensa a amplificação dos fragmentos do DNA, levando à detecção uniforme do genoma sequenciado.

Por outro lado, com uma abordagem mais prática e dinâmica, ocorre então o MinION (Oxford Nanopore Technologies), no qual a molécula de DNA passa através de nanoporos de proteína e, por meio destes, identifica-se a sequência nucleotídica devido às alterações na conformação e corrente elétrica dentro do nanoporo. Uma característica bastante relevante desse sistema é que, na sua versão mais simples, o MinION tem alguns centímetros de comprimento, sendo uma abordagem altamente eficiente para sequenciamento rápido, principalmente no local de coleta da amostra (OXFORD NANOPORE TECHNOLOGIES, 2018).

4.4 Montagem de genomas

Após o sequenciamento, as leituras são processadas para determinar o nível de qualidade. Por padrão utiliza-se o *Phred score*, que determina a probabilidade de erro por número de base e foi originalmente desenvolvido para uso no sequenciamento do genoma humano. A pontuação de cada base é fornecida pelo sequenciador e depois é aplicada em uma escala logarítmica de base 10 para determinar a probabilidade do erro. Por exemplo, um *Phred score* de 20 equivale a probabilidade de 1 erro a cada 100 bases e acurácia de 99% (EWING *et al.*, 1998).

A etapa de montagem inicia com o intuito de obter um consenso do conjunto total de leituras, para que todas as leituras possam ser sobrepostas a fim de refletir a sequência do DNA original, denominado de *contig* (KOREN *et al.*, 2012). O desafio computacional da montagem está justamente na reconstrução desse genoma e torna-se mais problemático, quando a complexidade do genoma é maior, e as sequências geradas são leituras curtas e/ou com baixa qualidade.

Os erros mais comuns dessa abordagem são inserções, deleções e concatenações de dois blocos gênicos em posições erradas. Alguns fatores como regiões genômicas repetitivas são outro desafio para os algoritmos de montagem, o que dificulta a finalização da montagem e, assim, chegar no resultado de um genoma completo. Ao orientar dois ou mais *contigs*, temos então um *scaffold* ou "super *contig*", isto é importante para indicar as regiões de *gaps* (lacunas no genoma onde não se sabe o conteúdo gênico).

Nesse ponto, se não for possível unir todos os *scaffolds*, o genoma fica como *draft*, ou seja, são representados por mais de um *contig ou scaffold*. Possíveis erros na montagem têm efeito crucial na qualidade do genoma final e refletem na acurácia e confiabilidade para futuras análises. Isso torna-se mais grave quando esses erros são negligenciados, dificultando a obtenção de respostas às questões biológicas mais precisas e conclusivas (BAKER, 2012).

Para aumentar a confiança nos dados gerados pelo sequenciamento e facilitar o processo de montagem, os pesquisadores estimam a cobertura do genoma, que é um cálculo do genoma final no total de leituras. Contudo, podem existir regiões sem cobertura ou com cobertura ruim que dificultam o processo de montagem e, portanto, é recomendado uma cobertura de 30X para genomas bacterianos (SIMS *et al.*, 2014).

Podemos definir a cobertura do genoma que foi sequenciado como cobertura teórica, que é a porcentagem de bases totais oriundas das leituras dividido pelo tamanho esperado do genoma ou a profundidade da cobertura, que é o número de vezes que uma base específica é representada nas leituras alinhadas no genoma final. De modo geral, a montagem pode ser conduzida por duas metodologias: por referência ou *ab initio,* que serão abordadas a seguir (BAKER, 2012) .

4.4.1 Montagem por referência

Na montagem por referência, as leituras são alinhadas e ancoradas em uma sequência de referência, geralmente um genoma mais filogeneticamente mais próximo, pois, quanto mais similar for o genoma de referência, maior o número de leituras alinhadas (SIMS *et al.*, 2014). Uma das principais limitações dessa abordagem está na presença de regiões repetitivas, pois independentemente da quantidade destas regiões do genoma sequenciado, apenas as regiões do genoma de referência serão representadas após o alinhamento. Outra grave consequência é a influência de rearranjos gênicos, podendo não representar particularidades e a real ordenação gênica do genoma sequenciado.

O principal algoritmo usado na montagem por referência é a tabela *hash* (RAMOS *et al.*, 2011). Essa abordagem usa como elemento chave as subsequências obtidas da sequência de busca (*query*). O programa busca alinhar sequências idênticas, conhecidas como sementes da referência, para que elas sejam estendidas. Temos como representantes Bowtie (LANGMEAD, 2013) e BWA (LI, H.; DURBIN, 2009). Com o crescente avanço dos programas que usam a abordagem *ab initio* e devido aos pontos negativos da montagem por referência, esta última está, cada vez mais, em desuso no contexto de montagem de genomas bacterianos.

4.4.2 Montagem *ab initio*

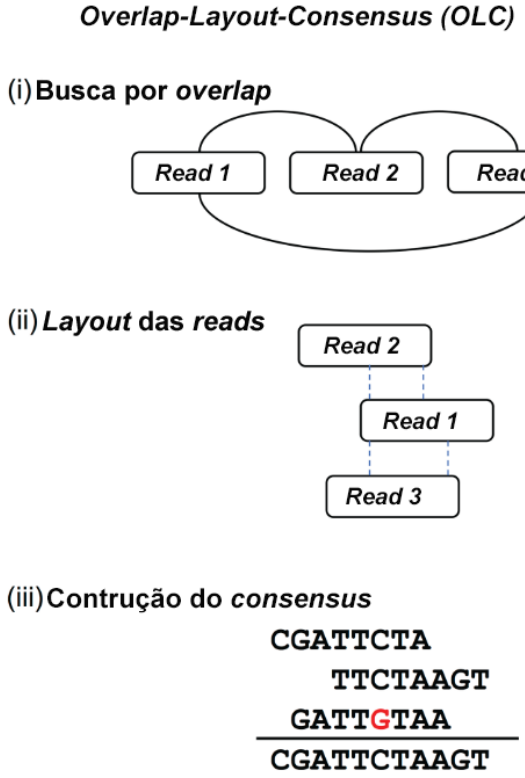
A abordagem *ab initio* utiliza somente os dados brutos gerados no sequenciamento, partindo do pressuposto que existe uma sobreposição entre as sequências nucleotídicas (MILLER; KOREN; SUTTON, 2010). Os algoritmos mais utilizados para este fim são *Overlap-Layout-Consensus* (OLC) e grafo *de Bruijn*.

No OLC o algoritmo é dividido em três partes, sendo a primeira as comparações entre todas as leituras (*Overlap*). Na segunda parte, *layout*, é gerado um grafo de sobreposição entre as leituras por alinhamento par-a-par, em que os nós são as leituras e as sobreposições são as arestas (POP, 2009).

Na terceira parte, *consensus*, a sequência consenso é construída com o alinhamento de múltiplas sequências, para que possa determinar um caminho passe todos os nós. Quando não é possível determinar um único caminho no grafo, são criados os *contigs*, os quais cada um representa o melhor caminho daquele conjunto de nós (Figura 1) (AYLING; CLARK; LEGGETT, 2019; MILLER; KOREN; SUTTON, 2010). O OLC é mais adequado para leituras longas geradas pelas plataformas 454 Roche, Sanger e Ion Torrent™ e está presente nos

programas Newbler (MARGULIES *et al.*, 2005), Edena (MARGULIES *et al.*, 2005) e Mira (DIGUISTINI *et al.*, 2009).

Figura 1 - Representação do algoritmo *Overlap-Layout-Consensus.* (a) Em Sobreposição, Layout, montagem de consenso, (i) são encontradas sobreposições entre leituras e um gráfico de sobreposição é construído (as arestas indicam leituras sobrepostas). (ii) As leituras são dispostas em contigs com base nas sobreposições (linhas tracejadas indicam partes sobrepostas). (iii) A sequência mais provável é escolhida para construir uma sequência de consenso. Polimorfismos (vermelhos) formam ramificações no gráfico.



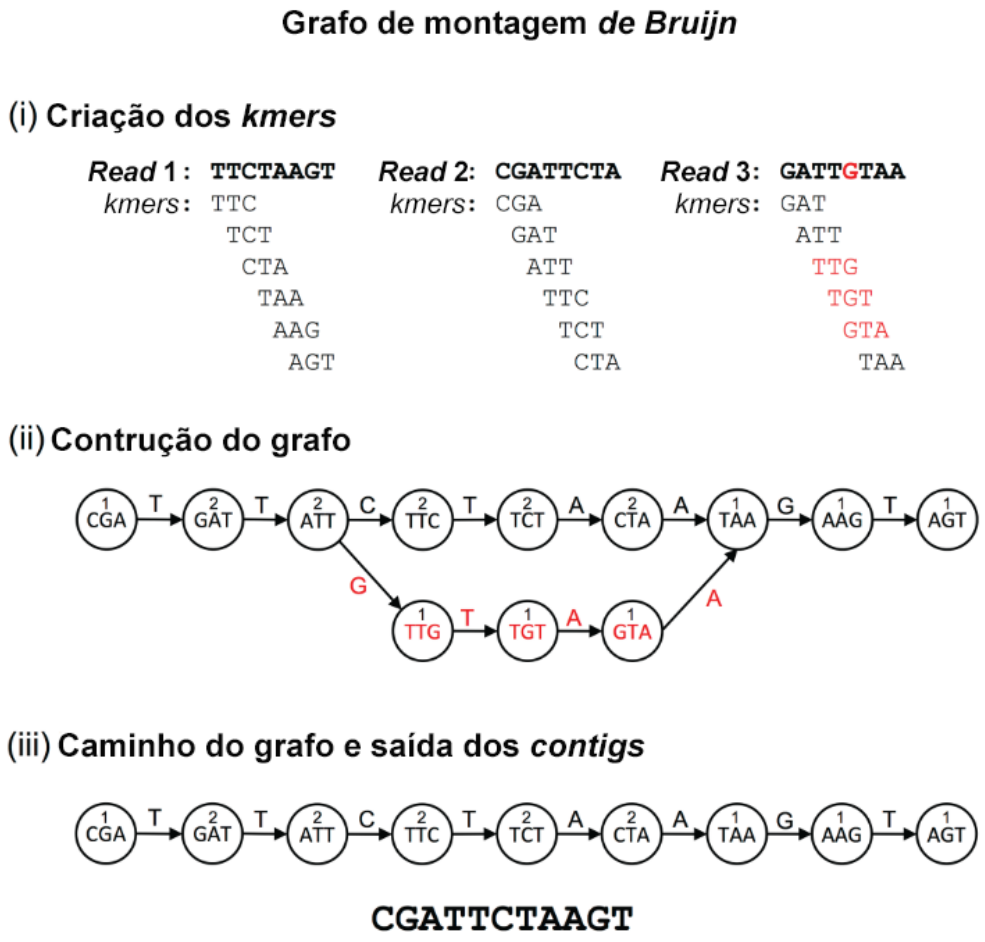
Fonte: Adaptada de Ayling *et al.* 2019

Já a estratégia que utiliza o grafo *de Bruijn* baseia-se na divisão das leituras em *k-mers* (definido pelo usuário ou por padrão), e busca por sobreposição entre *k-mers* adjacentes. Havendo sobreposição entre dois *k-mers* de tamanho k-1, existe uma ligação entre os fragmentos, a qual é expressa no grafo através de adição de uma aresta. Dessa forma, o algoritmo tenta criar grafos em que todos os caminhos entre os nós sejam percorridos, sendo assim, necessária a sobreposição entre as leituras adjacentes (Figura 2) (AYLING; CLARK; LEGGETT, 2019).

A validação da montagem *ab initio* pode ser medida pela quantidade e tamanho dos *contigs* e/ou *scaffolds* (*contigs* unidos e ordenados). Assim, quanto menor a quantidade e maior o *contig,* melhor é a montagem. Além disso existe a métrica do N50, que é o tamanho do *contig* que está em 50% de todos os *contigs* ordenados de forma crescente (MILLER; KOREN;

SUTTON, 2010). O programa QUAST auxilia bastante nesse processo por comparar diversas montagens, podendo ser usado com ou sem genoma referência. Para facilitar a decisão, o QUAST produz um relatório tabular detalhado e vários gráficos (GUREVICH *et al.*, 2013)

Figura 2 - Montagem do grafo *de Bruijn*. (i) as leituras são decompostas em kmers deslizando uma janela de tamanho k pelas leituras. (ii) Os *kmers* se tornam vértices no gráfico de *de Bruijn*, com arestas conectando *kmers* sobrepostos. Polimorfismos (vermelhos) formam ramificações no gráfico. Uma contagem é mantida de quantas vezes um *kmer* é visto, mostrado aqui como números acima dos *kmers*. (iii) Os contigs são construídos caminhando o gráfico a partir dos nós das bordas.



Fonte: Adaptada de Ayling et al. 2019

4.5 Ordenação dos genomas após montagem

Devido às regiões de baixa cobertura ou regiões repetitivas, os montadores não conseguem gerar uma única fita contínua do genoma. Entre as regiões repetitivas, as mais importantes são os operons de rRNA, que codificam as estruturas gênicas 5S, 16S e 23S. Os operons de rRNA bacterianos são regiões com aproximadamente 6000 a 8000 pares de bases

(pb) e com similaridade entre 98.04% e 99.94% (BASHIR *et al.*, 2012). Para que ocorra a concatenação dos *contigs,* torna-se necessária uma etapa de ordenação destas sequências e determinação das regiões entre eles, os *gaps* (NAGARAJAN; POP, 2013).

A partir dos *contigs*, o próximo passo é a ordenação para que o genoma final fique mais próximo possível do real. Para isso, são utilizadas algumas estratégias, tais como a ordenação por referência, a ordenação por bibliotecas pareadas ou pelo mapa óptico. Na ordenação por referência, utiliza-se de outro genoma com maior identidade nucleotídica e/ou mais próximo filogeneticamente como régua, e, por sintenia, os *contigs* são ordenados.

Isso tem um grande ponto negativo, pois essa abordagem impõe no novo genoma a mesma ordem no organismo de referência, dificultando análises sobre inversões gênicas. Já a ordenação por bibliotecas pareadas, o objetivo é usar as informações de distâncias das leituras quando o sequenciamento tem leituras pequenas com *paired-end* (Illumina Hiseq 2500 *paired-end*) ou leituras grandes (PacBio RS III), porém não é garantido que se possa ordenar todos os *contigs*, pois isso depende muito do alinhamento das leituras nos *contigs*. Já com a ordenação por meio de mapa óptico, é possível alinhar mais *contigs* e é mais estratégia mais recomendada, pois o mapa de restrição é feito com a mesma linhagem (ASTON; MISHRA; SCHWARTZ, 1999).

4.6 Estratégia para finalização da montagem pelo mapa óptico

O MapSolver™ (OpGen) foi desenvolvido principalmente para a finalidade de ordenação de *contigs* gerados pela abordagem *ab initio.* Eles são orientados com a fita do mapa óptico que tem as marcações dos sítios de restrição, construindo assim os *scaffolds* (XAVIER *et al.*, 2014). Essa estratégia contribui para geração de *scaffolds* de alta precisão, sendo possível a partir dessa abordagem detectar inversões genômicas que antes eram ocultadas pela estratégia de ordenação por referência.

A estratégia do mapa óptico de genoma completo, ou WGM (*Whole Genome Mapping*), trata-se de uma abordagem que utiliza mapas de restrição de alta resolução para gerar a orientação real do genoma do organismo. É o único método de análises de genomas em larga escala que fornece a visualização completa do genoma estrutural através de uma única imagem (WU *et al.*, 2009). O mapeamento óptico baseia-se na distância dos sítios de restrição para construção dos mapas de alta precisão. Essa é a estratégia em que se obtêm dados com maior

precisão, por ser um resultado físico do genoma, o qual representa a real orientação do cromossomo (SCHWARTZ *et al.*, 1993).

A técnica de mapeamento óptico foi primeiramente desenvolvida por Schwartz e colaboradores em 1993, com o intuito de estudar a ordenação gênica cromossômica de *Saccharomyces cerevisiae* (SCHWARTZ *et al.*, 1993). O mapa óptico é descrito como a nova abordagem para análise de molécula única de DNA, utilizando microscopia de fluorescência para identificar e registrar, permitindo estimar o tamanho dos fragmentos pelas imagens geradas. Várias melhorias foram adicionadas à técnica, principalmente nos quesitos de aquisição de imagens e algoritmos para estimativa do tamanho dos fragmentos (ASTON; MISHRA; SCHWARTZ, 1999; SAHA; RAJASEKARAN, 2014; SAMAD *et al.*, 1995).

A partir disso, o WGM ganhou notoriedade em diversas aplicações na microbiologia clínica, para tipagem de linhagens em caso de epidemias (KOTEWICZ *et al.*, 2008; PETERSEN *et al.*, 2011); na ordenação de *contigs* gerados pela montagem *ab initio* (ONMUS-LEONE *et al.*, 2013); e no estudo de inversões, inserções, deleções, duplicações e instabilidade de genomas bacterianos (CRISTINA *et al.*, 2016; JING *et al.*, 1999; KOTEWICZ *et al.*, 2007; LIN, 1999; SHUKLA *et al.*, 2012).

Deve-se ressaltar que a gama de organismos que podem ser utilizados no WGM vai de procariotos a eucariotos. O estudo com *Mycobacterium avium* subspecies *paratuberculosis* demonstra o sucesso da técnica usada para entender a diversidade do perfil genético entre linhagens (WU *et al.*, 2009). Em 2007 foi determinado o genoma do arroz (*Oryza sativa)* (ZHOU *et al.*, 2007). Sua aplicação, nos últimos anos ganhou mais destaque em genoma de plantas como o trigo (*Triticum aestivum*) (KAPUSTOVÁ *et al.*, 2019) e em mamíferos, como

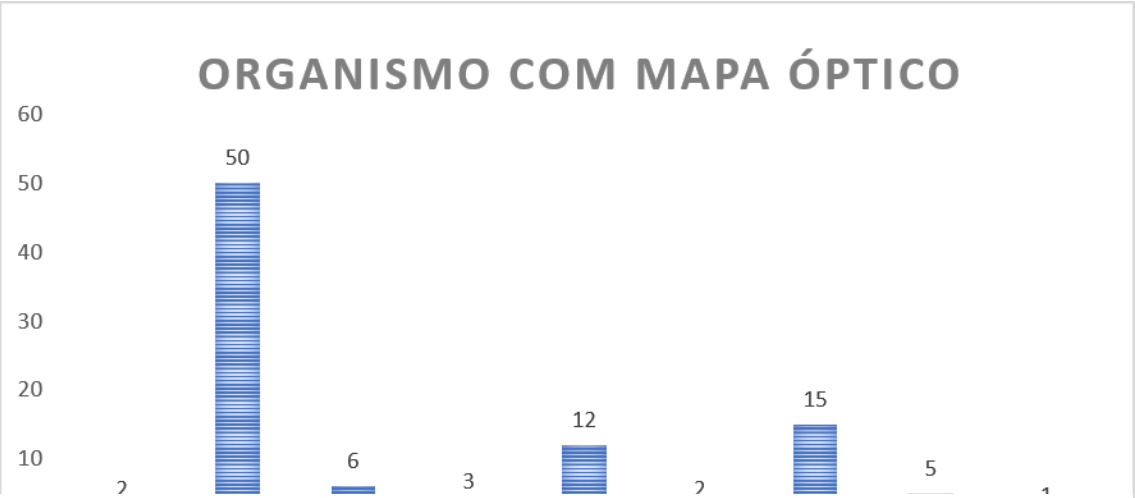


Figura 3 - Distribuição dos organismos em artigos científicos que utiliza a tecnologia do mapa óptico entre 1993 a 2020.

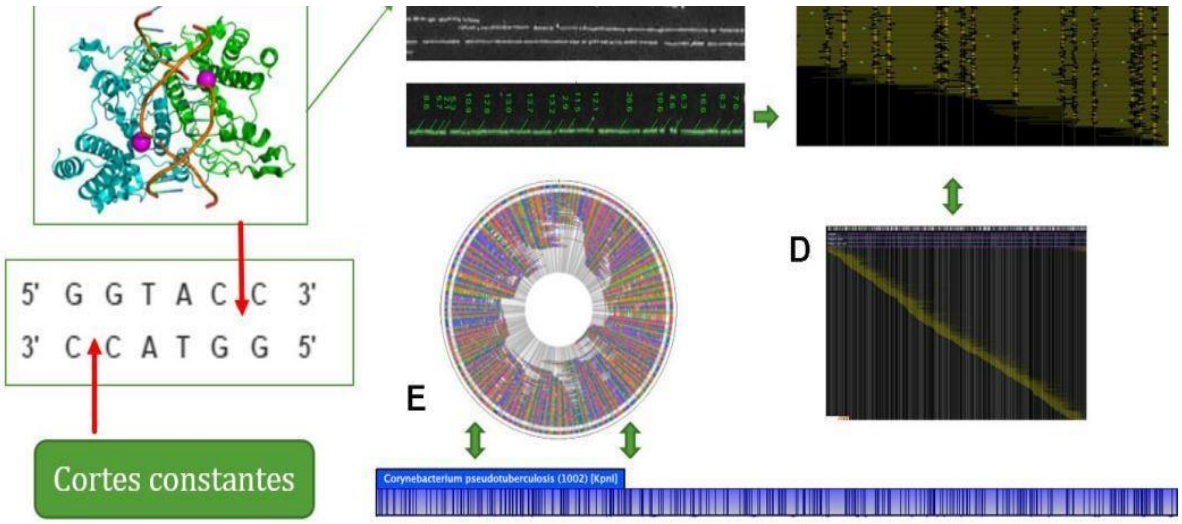
a cabra (*Capra aegagrus)* (DONG *et al.*, 2013) e o cão (*Canis lupus familiaris*) (WANG *et al.*, 2019).

O LGCM foi o pioneiro no uso do mapa óptico em genomas procariotos no Brasil ao aplicar a técnica em 11 linhagens de *C. pseudotuberculosis* entre biovares *equi* e *ovis*. A partir de revisão da literatura e considerando-se os artigos que trabalharam diretamente com mapa óptico no genoma de diferentes organismos, encontrou-se 96 trabalhos entre 1993 e 2020. Deste, mais de 50% dos trabalhos estão voltados para genomas bacterianos, seguido de genomas de plantas e mamíferos (Figura 3).

A construção do mapa óptico tem como princípio o alongamento da fita de DNA pela polarização das cargas negativas dos fosfatos em uma lâmina de vidro (Figura 4B). Em seguida, o genoma é fragmentado por enzimas de restrição, assim como a enzima KpnI, a qual reconhece e cliva a sequência de bases –GTAC– (Figura 4A e 4B). A KpnI executa cortes constantes no genoma, devido a probabilidade de combinação das bases ser 4x4x4x4 = 256 bases.

Os fragmentos são marcados com um corante fluorescente e um microscópio de alta precisão registra as imagens de acordo com a intensidade de fluorescência dos fragmentos (Figura 4B). O programa Argus (*Argus imaging system*, OpGen) calcula o tamanho dos fragmentos, registrando também a ordem e os pontos de clivagem (Figura 4C). Todas as imagens dos fragmentos são sobrepostas (Figura 4D), formando um mapa de restrição do genoma inteiro circular e linearizado com todos os sítios de restrições (Figura 4E), tem-se uma cobertura mínima de 30x (OPGEN, 2016).

Figura 4 - Fluxograma do processo de construção do mapa óptico. (A) Ação de clivagem pela enzima *kpn1* na dupla fita de DNA ao reconhecer a sequência de bases "GTAC". (B) Moléculas de DNA alongadas e fixadas por meio da polarização das cargas negativas e clivagem por endonucleases, com marcação posterior a partir de fluoróforos. (C) O programa Argus (*Argus imaging system*, OpGen) calcula o tamanho dos fragmentos, registrando também a ordem e os pontos de clivagem. (D) Alinhamento de todos os fragmentos reconhecidos pelo programa Argus. (E) Representação do mapa de restrição do genoma inteiro de forma circular.



Contudo, existem limitações inerentes ao método do mapa como é o problema de haver erros no experimento, tais como: clivagem incompleta pelas enzimas em alguns pontos podendo não reconhecer ou recolher um ponto falso; clivagem elevatória não especifica; erro no dimensionamento dos fragmentos, já que ele é uma proporção da intensidade de florescência reconhecida por meio de microscopia; tamanho mínimo do fragmento.

4.7 Finalização da montagem

Mesmo com os *scaffolds* formados, ainda é preciso resolver as regiões de *gaps,* que são regiões não representadas entre os *contigs.* A primeira estratégia para fechar *gaps* é realizar o mapeamento de todas as leituras contra o genoma de referência e detectar a sequência consenso correspondente à localização do *gap*. O programa CLC *Genomics Workbench* (Qiagen) é utilizado nesse tipo de estratégia. Em adição, pode-se utilizar o programa Gfinisher (GUIZELINI *et al.*, 2016) ou o GapBlaster (DE SÁ *et al.*, 2016) que fazem uso de *contigs* gerados em outras montagens com o mesmo conjunto de leituras. Esse método tem vantagens, pois diferentes programas podem construir *contigs* de tamanhos e regiões diferentes com o mesmo conjunto de leituras, e isso ocorre devido ao algoritmo que está implementado nos mesmos.

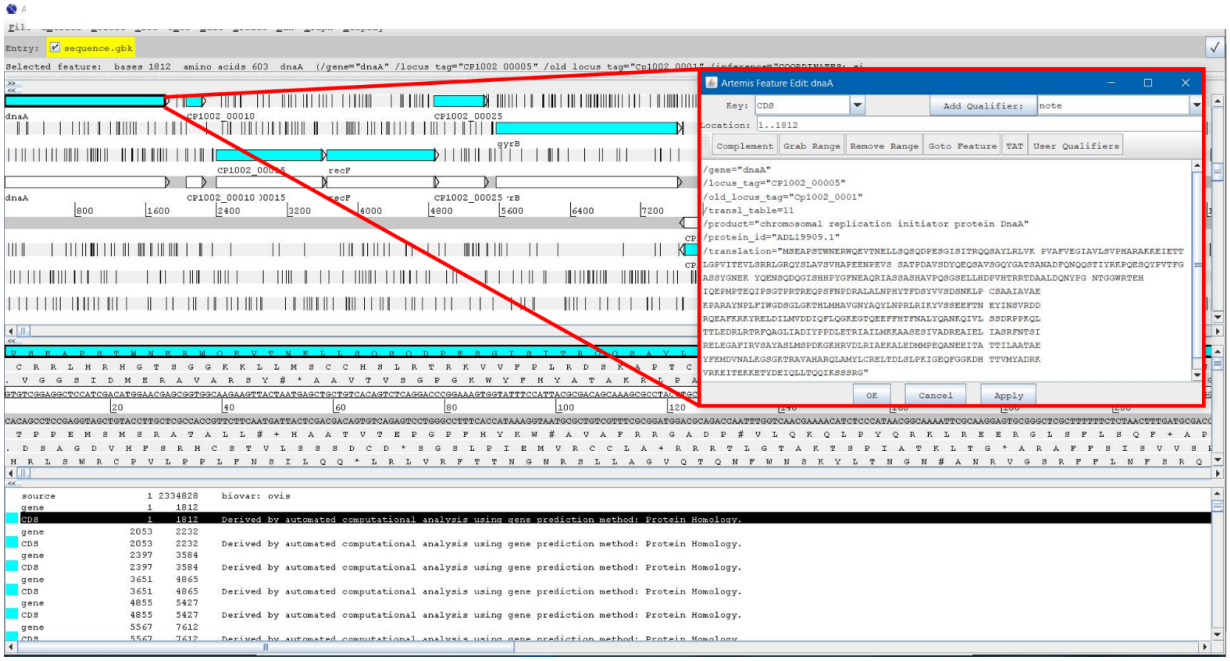
4. 8 Anotação estrutural e funcional

Com a finalização do processo de montagem, é necessário localizar e identificar as regiões de ORF (*Open Read Frame*) e assim, atribuir uma função para essas regiões (Figura 5). Esses processos são denominados predição e anotação gênica, respectivamente, sendo essencial para as análises funcionais posteriores. Isso é feito através da homologia existente entre genes presentes no mesmo grupo de organismos (BRYSON *et al.*, 2005).

Serviços de anotação automática são amplamente utilizados em trabalhos científicos atuais, como por exemplo, o RASTtk (BRETTIN *et al.*, 2015), *Prokaryotic Genome Annotation Pipeline* (PGAP) (TATUSOVA *et al.*, 2016) e o *Rapid Prokaryotic Genome Annotation* (Prokka) (SEEMANN, 2014), sendo os dois últimos uma alternativa para uso local. A última etapa desse processo é a curadoria manual das ORFs, para detecção dos pseudogenes ou erros na anotação pelos preditores automáticos. Essa etapa é crucial e depende de profissionais bem treinados, tendo assim um dado curado e de alta precisão necessário para futuras análises,

principalmente para estudos que envolvem genômica comparativa, pan-genoma, transcriptoma, vacinologia reversa, dentre outros.

Figura 5 - Predição estrutural e funcional do genoma visualizado pelo programa artemis. Em destaque está o gene da *dnaA* na linhagem de *C. pseudotuberculosis* 1002.



O propósito de sequenciar e montar genomas, está em ser possível explorar e descrever evidências muito além das características morfológicas. Assim, ao detalhar todos os genes, proteínas e rRNA e tRNA e realizar análises comparativas entre linhagens e espécies, pode-se supor como atuou o processo evolutivo sobre o mesmo. Dessa forma, pode-se entender como uma bactéria pode adquirir resistência a antibióticos ou um vírus tornar-se capaz de infectar novos hospedeiros.

Uma bactéria que vem sendo estudada em diferentes âmbitos das ômicas em nosso laboratório é a *C. pseudotuberculosis*, a qual é o modelo experimental desta tese.

4.9 Características fenotípicas e estruturais de *C. pseudotuberculosis*

C. pseudotuberculosis pertence ao grupo CMNR (*Corynebacterium*, *Mycobacterium*, *Nocardia* e *Rhodococcus*), caracterizado pelo alto conteúdo GC (46 – 74%) e pela estrutura da parede celular, composta principalmente por peptidoglicanos, arabinogalactanos e ácidos micólicos. Em relação às características morfológicas e bioquímicas, *C. pseudotuberculosis* se comporta como uma bactéria Gram-positiva, pleomórfica e intracelular facultativa. Suas

medidas microscópicas variam de 0,5 a 0,6 μm por 1 a 3 μm (DORELLA, FERNANDA ALVES *et al.*, 2006).

A espécie é dividida em dois biovares: *ovis* e *equi*. Esta divisão é baseada na competência da redução de nitrato, presente nas linhagens do biovar *equi* e ausente no biovar *ovis*. Esta predisposição é decorrente da presença do *cluster nar*, composto pelos genes *moaE*, *molB*, *molA*, *narI*, *narJ*, *narH*, *narG*, *narK*, *narT*, *moeY*, *moaC* (ALMEIDA *et al.*, 2017).

C. pseudotuberculosis é o agente etiológico da Linfadenite Caseosa*,* que provoca diferentes sintomas dependendo do hospedeiro. A infecção em cabras e ovelhas resulta na formação de granulomas nos linfonodos (WINDSOR; BUSH, 2016). Os bovinos infectados apresentam lesões granulomatosas ulcerativas e mastite, enquanto os equinos são diagnosticados com quadro descrito como linfangite ulcerativa (GUEDES *et al.*, 2015).

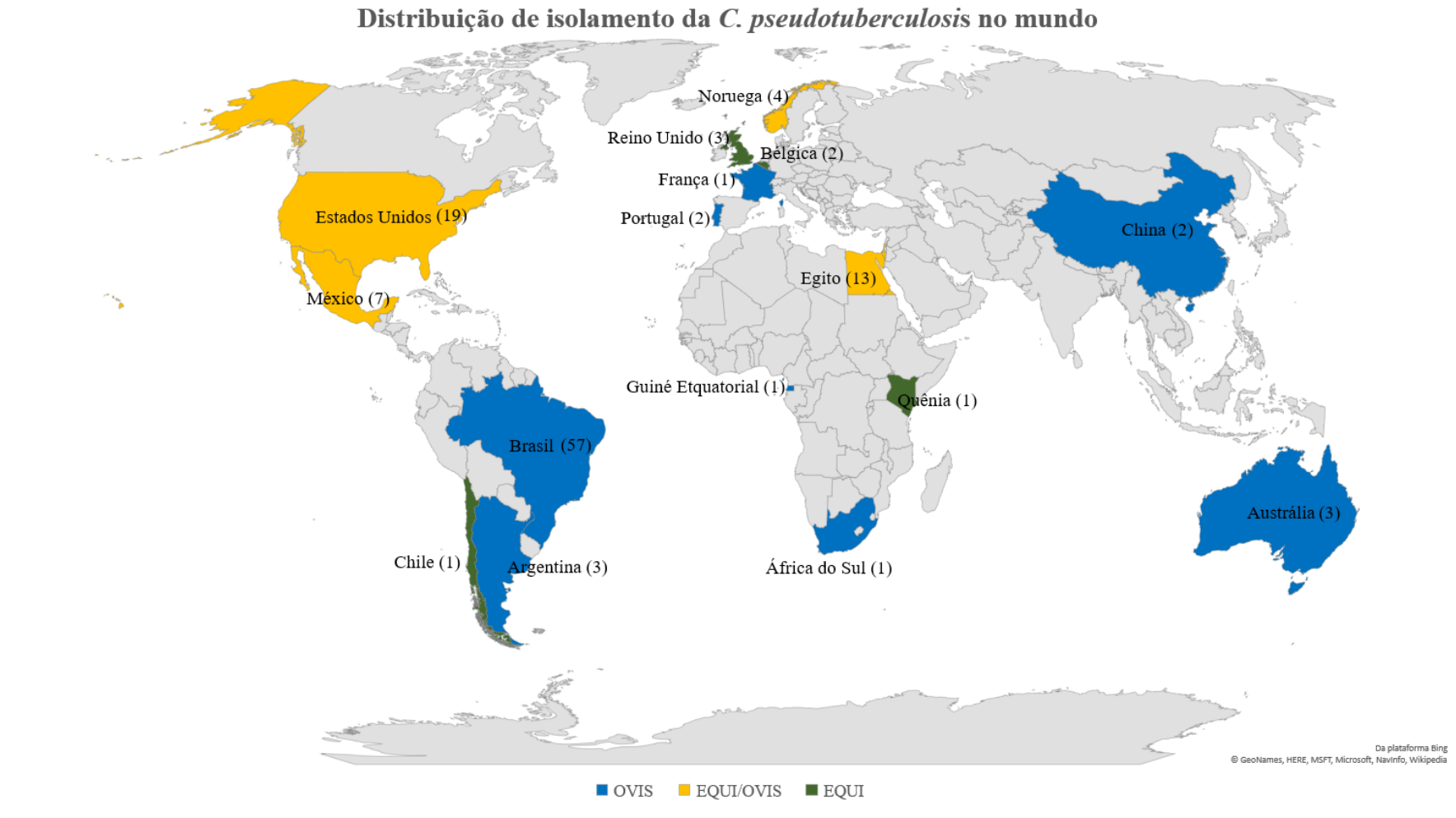
C. pseudotuberculosis é considerado um patógeno zoonótico de importância veterinária que causa perdas financeiras ao agronegócio em diversos países. Na Figura 6, podemos visualizar a distribuição mundial da ocorrência da infecção com a bactéria, com todos os países em que já foram isoladas amostras.

No Brasil é ainda mais preocupante, pois a atividade da pecuária se estende por todo território nacional, principalmente na região nordeste. Apesar de não haver registro na literatura de infecção pela *C. pseudotuberculosis* em bovinos e equinos no Brasil, a criação de bovinos, equinos, caprinos e ovinos é importante não apenas do ponto de vista econômico, mas também como estratégia de coexistência no semiárido.

De acordo com Farias e colaboradores (FARIAS *et al.*, 2014), a espécie *Capra hircus* se adaptou facilmente às condições climáticas dessa área, característica que faz do Nordeste o maior produtor do Brasil de pequenos ruminantes, com 93,2% do rebanho nacional (8.944.461) (MAGALHÃES *et al.*, 2018). Consequentemente, as perdas econômicas devido à redução da produção de leite e carne, bem como o aproveitamento da pele do animal, afetam diretamente a renda dos criadores.

.

Figura 6 - Distribuição global de isolados e sequenciados de *C. pseudotuberculosis* até 2020. Apenas isolados do biovar *ovis* (Azul); País com isolados do biovar *equi* e *ovis* (Amarelo); País com apenas isolados do biovar *equi* (Verde).



4.10 Genômica de *C. pseudotuberculosis*

Os primeiros registros de sequências gênicas completas de *C. pseudotuberculosis* começaram pela construção de bibliotecas genômicas através de vetores plasmidiais e em BACs (*Bacterial Artificial Chromossome*). Neste trabalho, foi utilizada a linhagem *C. pseudotuberculosis* 1002, que foi isolada de um caprino com granulomas em linfonodos superficiais, na cidade de Sobral, no estado do Ceará, Brasil, em 1971 (MEYER *et al.*, 2002), e obtidos 215 *Genomic Survey Sequences* (GSSs), os quais foram depositadas no G*enBank* (DORELLA, F. A. *et al.*, 2006). A construção dessas bibliotecas foi de grande importância para estabelecer relações filogenéticas com outras espécies do gênero e como parâmetro para os futuros sequenciamentos a partir das técnicas de *Whole genome sequencing*.

Então, no período de 2009 a 2010, formou-se a Rede Paraense de Genômica e Proteômica (RPGP), apoiada pela FAPESPA e a Rede Genoma de Minas Gerais (RGMG), apoiada pela FAPEMIG, com o intuito de desenvolver trabalhos de sequenciamento e genômica comparativa de bactérias as quais são agentes etiológicos importantes dentro do Brasil.

A linhagem sequenciada pela RGMG foi *C. pseudotuberculosis* 1002 (Cp1002, biovar *ovis*), selecionada devido aos vários trabalhos já desenvolvidos pelo grupo do Professor Roberto Meyer e do Professor Vasco Azevedo em experimentos de bancada, linhagem a qual, posteriormente, se tornou modelo para o biovar *ovis* em nosso laboratório. No Brasil, até o momento, não existem relatos de infecção natural pelo biovar *equi,* sendo assim, os casos e a relevância fazem relação ao biovar *ovis*.

A linhagem Cp1002 foi o primeiro genoma montado no estado de Minas Gerais, o que justifica sua importância como organismo modelo para nosso grupo. Seu primeiro sequenciamento foi pela plataforma 454 Roche (Pirosequenciamento) e Sanger, apresentando um genoma circular com ~2,35Mb. Nesse mesmo tempo, a *C. pseudotuberculosis* C231 isolada de um ovino e sequenciada pela plataforma 454 Roche no *Commonwealth Scientific and Industrial Research Organisation* (CSIRO) na Austrália, sendo montada e anotada através da RGMG.

Em paralelo, um grupo na Universidade de Bielefeld na Alemanha realizava o sequenciamento de *C. pseudotuberculosis* FRC41 com a abordagem de pirosequenciamento pela plataforma *Genome Sequencer FLX,* sendo que essa linhagem foi isolada de uma garota de 12 anos na França (TROST *et al.*, 2010). Os genomas dessas linhagens foram finalizados e depositados como sequências completas no NCBI, podendo ser obtido com o número de acesso

do *GenBank*: CP001809.1 (Cp1002); CP001829.1 (CpC231) e a CP002097.1 (CpFRC41) (RUIZ *et al.*, 2011; TROST *et al.*, 2010).

Entre maio de 2009 a 2011, na Universidade Federal do Pará, foram concluídos os primeiros sequenciamentos de nova geração na América Latina, utilizando o sistema SOLiD™ (*Supported Oligonucleotide Ligation and Detection*) da Applied Biosystems, que era uma das plataformas de sequenciamento genômico que utilizava tecnologia NGS.

Foram decodificados, naquele momento, os genomas de três novas linhagens de *C. pseudotuberculosis*: a linhagem *C. pseudotuberculosis* 258, isolada de cavalo com linfangite ulcerativa na Bélgica (SOARES, SIOMAR C. *et al.*, 2013); a *C. pseudotuberculosis* 162, isolada do abscesso externo no pescoço de um camelo no Reino Unido em 1999 (HASSAN *et al.*, 2012) e a *C. pseudotuberculosis* 31 isolada do edema da pele de um Búfalo no Egito, sendo que todas as linhagens sequenciadas neste dado momento foram do biovar *equi*. O genoma completo de cada uma também foi depositado no NCBI, com os códigos do *GenBank* CP003540.1 (Cp258); CP003652.1 (Cp162) e CP003421.1 (Cp31).

Essas foram as primeiras linhagens de *C. pseudotuberculosis* a serem sequenciadas, tanto do biovar *ovis* quanto do *equi.* Contudo, foi apenas o começo dessa jornada científica liderada pelo Professor Vasco Azevedo (UFMG) e o Professor Artur Silva (UFPA). Com várias publicações nacionais e internacionais, vários grupos de pesquisas entraram em contato para enviar amostras de isolados de *C. pseudotuberculosis* e assim, estabelecer parcerias e projetos de grande impacto. Vale ressaltar também que várias linhagens sequenciadas por Ion Torrent PGM™, que serão descritas, foram fruto da parceria do LGCM com o AQUACEN, coordenado pelo Prof. Dr. Henrique César Pereira Figueiredo, professor da Escola de Veterinária da UFMG.

Dentre estes isolados podemos destacar onze linhagens isoladas de Búfalos (biovar *equi*) em regiões do Cairo-Egito pelo Professor Salah A. K. Selim durante um surto de doença edematosa de pele, ou *Oedematous Skin Disease* (OSD), no verão de 2008. As onze linhagens foram sequenciadas por Ion Torrent PGM-318™ Hi-Q 400pb, e analisadas por Viana e colaboradores em 2016-2017 (VIANA *et al.*, 2017). Já nos EUA, doze linhagens foram isoladas de cavalos na região da Califórnia (biovar *equi*), sequenciadas pela plataforma Ion Torrent PGM-318™ e, após análises, foi possível identificar ilhas de patogenicidade com fatores de virulência importantes como um cluster de genes de permeação de oligopeptídeos. Além disso, independentemente do tipo de infecção, os isolados da Califórnia mostraram maior similaridade genômica e proximidade filogenética entre si do que com o restante das linhagens do biovar *equi* (BARAÚNA *et al.*, 2017).

Isolados do México, seis linhagens oriundas de caprinos, ovinos e equinos demonstraram pelas análises comparativas, uma possível evidência de um *cluster* de genes relacionados ao complexo CRISPR-Cas, que foi encontrado exclusivamente no biovar *equi,* e outro cluster de proteínas relacionados com genes dos sistemas de restrição-modificação (RM) do tipo III (PARISE *et al.*, 2018). Contudo, estudos para ampliar essas evidências serão necessários.

No Brasil, três linhagens provenientes da região do semiárido baiano foram isoladas de abscessos em linfonodos superficiais de caprinos. São a CpVD57 (*GenBank*: CP009927), a CpT1 (*GenBank*: CP015100) e a CpMIC-6 (*GenBank*: CP019769.1), as quais já foram utilizadas para testes vacinais *in vivo,* inclusive a CpMIC-6 (também chamada de Pus6) considerada uma linhagem virulenta por testes de infecções *in vitro* e *in vivo* pelo Laboratório de Imunologia e Biologia Molecular (LABIMUNO) (COELHO, 2007).

A CpT1 e VD57 foram isoladas de caprinos e usadas em vários trabalhos para caracterização de perfis microbiológicos e imunológicos pelo Laboratório de Imunologia da UFBA (MOURA-COSTA *et al.*, 2008). Foi construída uma linhagem mutante atenuada da CpT1, gerando a Cp13 (*GenBank*: CP014998) (DORELLA, FERNANDA A. *et al.*, 2009). Em 2016, foi realizado o primeiro sequenciamento da CpT1 (ALMEIDA; LOUREIRO; *et al.*, 2016). Além disso a própria CpT1 e a Cp13 também já foram apuradas em teste de resistência a estresses abióticos e analisadas por RNA-seq (IBRAIM *et al.*, 2019).

Ainda no continente americano, mais três linhagens foram isoladas na região da Patagônia, na Argentina. A CpPAT10, a CpPAT14 e a CpPAT16, todas isoladas de ovelhas com abcessos em consequência da LC, sendo que as duas últimas linhagens foram recentemente sequenciadas pela plataforma Illumina Hiseq 2500 com *paired-end*.

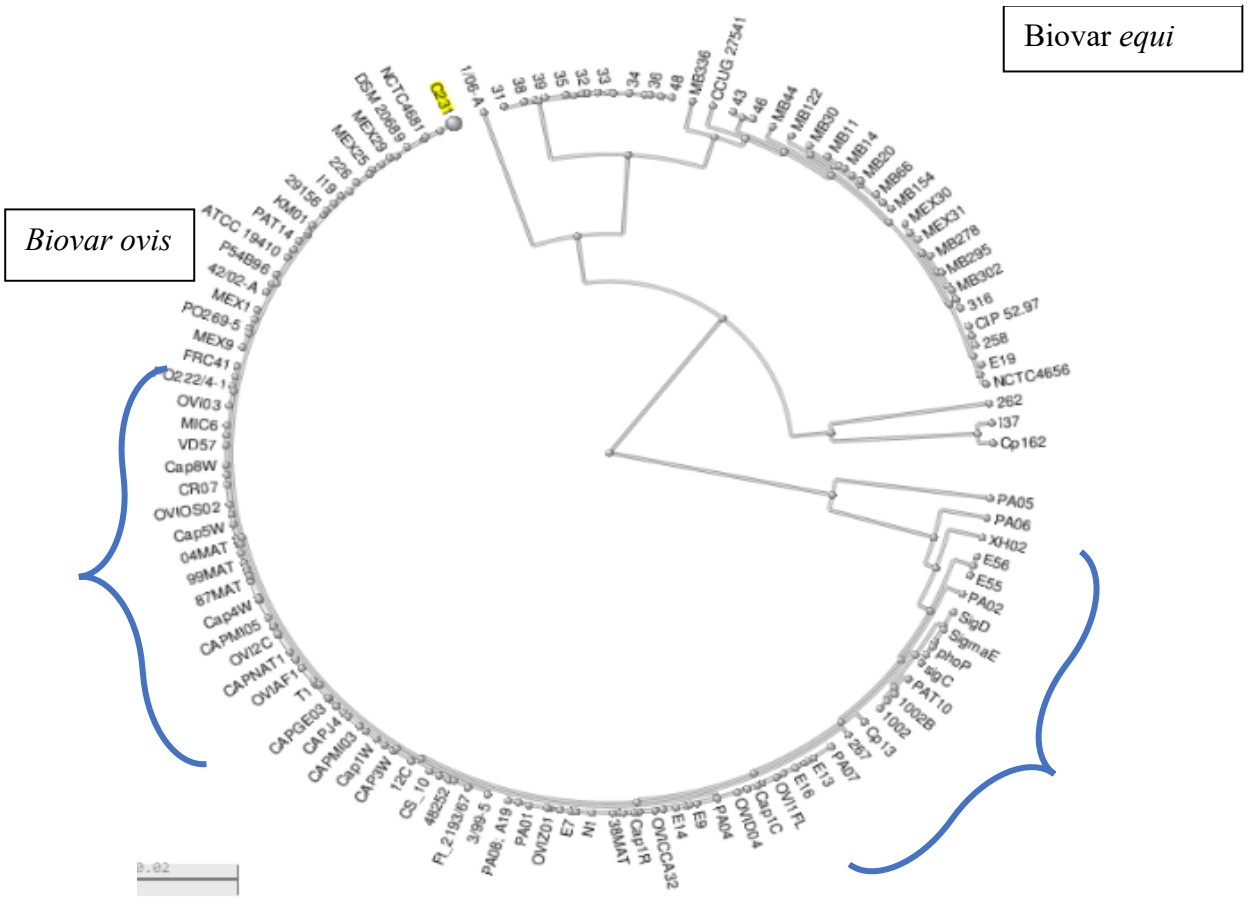
Várias outras linhagens de diferentes países e hospedeiros também foram selecionadas (Tabela 2 - Apêndice A). A construção desse repositório genômico foi fundamental para realização dos estudos de genômica comparativa e pan-genômica de *C. pseudotuberculosis* e demais espécies do gênero (ARAÚJO *et al.*, 2019; RUIZ *et al.*, 2011; SOARES, SIOMAR DE CASTRO, 2013; VIANA *et al.*, 2017).

Em 2013, Soares e colaboradores, realizaram as primeiras análises de pan-genômica com 15 genomas da espécie, caracterizando esta espécie com um pan-genoma aberto, no qual aproximadamente 19 novas sequências de codificação de proteínas seriam adicionadas para cada novo genoma. O genoma central é constituído de 1.504 sequências que codificam proteínas. Análises mais detalhadas do pan-genoma revelaram diferenças entre as estirpes *ovis, e* o biovar *ovis* mostrou um comportamento mais clonal do que as linhagens do biovar *equi*

(SOARES, SC *et al.*, 2013). No mesmo ano, Soares e colaboradores identificaram por vacinologia reversa 49 possíveis proteínas a serem candidatas a alvos vacinais usando o genoma de *C. pseudotuberculosis* 258 biovar *equi* (SOARES, SIOMAR C. *et al.*, 2013).

Todos estes trabalhos demonstram a importância de se disponibilizar para comunidade científica, genomas de qualidade para que trabalhos futuros possam ser reproduzidos com a maior confiabilidade possível. Na Figura 7, podemos observar a distribuição de todas as linhagens de *C. pseudotuberculosis* disponíveis no NCBI. São 115 genomas completos (95% dos genomas) e 8 incompletos, sendo que nosso grupo depositou 114 linhagens, das quais 71 linhagens tiveram a montagem ou remontagem neste trabalho de tese, e as mesmas estão marcadas em verde (Tabela 2 - Apêndice A).

Figura 7 - Representação em dendograma de todas as linhagens de *C. pseudotuberculosis* disponíveis na base de dados do NCBI. É possível notar a divisão em dois grandes clados, pertencentes ao biovar *equi* e *ovis*.



Fonte: National Center for Biotechnology Information, Genome Database (Junho de 2020).

4.11 Artigo I – Re-sequencing and optical mapping reveals misassemblies and real inversions on *Corynebacterium pseudotuberculosis* genomes

O artigo em destaque dessa seção foi publicado na revista - ***Scientific Report*s (*Nature Research*)** em novembro de 2019 (DOI: 10.1038/s41598-019-52695-4), fator de impacto: 4,122 (2018).

Em 2016, o trabalho intitulado (*Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002*) realizado por Mariano (MARIANO, DIEGO CÉSAR BATISTA *et al.*, 2016), do qual também sou coautor, esclareceu que mesmo a linhagem mais trabalhada do nosso grupo apresentava erros de ordenação gênica em seu genoma.

A *C. pseudotuberculosis* 1002 (biovar *ovis*) foi então re-sequenciada por Ion Torrent PGM^{TM} 400 pb *fragment* e agora depositada como Cp1002B. A essência desse manuscrito foi responder a perguntas que vinham nos acompanhando, como: Quais outros genomas de *C. pseudotuberculosis* depositados poderiam estar incompletos? E quais informações poderíamos estar perdendo? Para isso, re-sequenciamos 8 linhagens de *C. pseudotuberculosis* por ION Torrent PGM^{TM} que haviam sido sequenciadas pela tecnologia SoliD V2 e V3, Sanger e 454, além de três novas linhagens que são de interesse para o grupo.

A linhagem *C. pseudotuberculosis* 29156 (biovar *ovis*) foi isolada de bovino em Israel, o qual é o único hospedeiro, em que foram isolados ambos os biovares. Por fim, as linhagens CpMB302 e a CpMB11 pertencentes ao biovar *equi* e isoladas na região da California nos EUA, onde há uma grande incidência de infecção, como já descrito pelo grupo da Dra. Sharon J. Spier (HAAS *et al.*, 2017).

Além disso, ter mais linhagens de mais diferentes regiões e hospedeiros proporciona um melhor entendimento da biologia e a evolução da bactéria. Em resumo, ao utilizar a técnica de mapa óptico em genomas de *C. pseudotuberculosis* já depositados, para detecção e correção de erros de montagem e em novos genomas para evitar possíveis erros, obtém-se dados com maior confiabilidade e acurácia. O mapa óptico foi essencial para detectar e corrigir vários erros, como por exemplo a localização física errônea de alguns *loci* gênicos, a identificação de genes essenciais para a diferenciação do *biovar* que não estavam presentes nos dados depositados, e expôs evidências de rearranjos entre *clusters* de ribossomos.

www.nature.com/scientificreports

SCIENTIFIC REPORTS
nature research

OPEN

Re-sequencing and optical mapping reveals misassemblies and real inversions on *Corynebacterium pseudotuberculosis* genomes

Thiago de Jesus Sousa[1], Doglas Parise[1], Rodrigo Profeta[1], Mariana Teixeira Dornelles Parise[1], Anne Cybelle Pinto Gomide[1], Rodrigo Bentos Kato[1], Felipe Luiz Pereira[2], Henrique Cesar Pereira Figueiredo[2], Rommel Ramos[3], Bertram Brenig[4], Artur Luiz da Costa da Silva[3], Preetam Ghosh[5], Debmalya Barh[6], Aristóteles Góes-Neto[1] & Vasco Azevedo[1*]

The number of draft genomes deposited in Genbank from the National Center for Biotechnology Information (NCBI) is higher than the complete ones. Draft genomes are assemblies that contain fragments of misassembled regions (gaps). Such draft genomes present a hindrance to the complete understanding of the biology and evolution of the organism since they lack genomic information. To overcome this problem, strategies to improve the assembly process are developed continuously. Also, the greatest challenge to the assembly progress is the presence of repetitive DNA regions. This article highlights the use of optical mapping, to detect and correct assembly errors in *Corynebacterium pseudotuberculosis*. We also demonstrate that choosing a reference genome should be done with caution to avoid assembly errors and loss of genetic information.

Next Generation Sequencing (NGS) platforms provide an exponential increase in the amount of data produced in a single assay (high-throughput data). This approach provided the scientific community with the ability to sequence more genomes at reduced costs. The NGS platforms perform the sequencing through different technologies, which were developed by different companies, such as 454 GS FLX system (Roche)[1]; Hiseq paired-end (Illumina)[2]; Ion Torrent PGM (Life Technologies)[3]; PacBio sequel system(Pacific Biosciences); and MinION (Oxford Nanopore)[4]. From these, thousands of genomic projects were created to sequence Bacteria, Archaea, and Eukarya species, viruses, and metagenomes[5].

The main database of these sequences is GenBank maintained by the National Center for Biotechnology Information (NCBI), which in September 2018, contained 153,992 bacterial genomes, most of these being drafts, and only 11,103 sequences (7%) were complete genome sequences. Furthermore, the complete sequences still might have misassemblies due to the presence of repetitive regions, such as ribosomal RNA (rRNA), transposases, phage regions, and plasmids[6]. These errors bias future studies and inferences, such as in comparative genomic or structural genomic analyses, and even ordering of phylogenetically related genomes[7]. Thus, obtaining a more precise and accurate complete genome sequence of an organism is fundamental to understanding its biological and evolutionary characteristics[7].

The assembly problem persists even with the increase in the reads size, sequencing quality, and updates of *de novo* assembly algorithms. Another limiting factor to the increase of complete sequences is the lack of trained professionals. However, approaches to support this process have been gaining prominence[8]. For example, the use of SSPACE[9] software to use paired-end reads to create a consensus sequence and perform scaffolding of contigs. Similarly, MapRepeat[10] and riboSeed[11] try to solve the repetitive region's problem.

[1]Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [2]National Reference Laboratory for Aquatic Animal Diseases (AQUACEN) of Ministry of Agriculture, Livestock and Food Supply, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [3]Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil. [4]Institute of Veterinary Medicine, University Göttingen, Göttingen, Germany. [5]Department of Computer Science, Virginia Commonwealth University, Richmond, United States. [6]Institute of Integrative Omics and Applied Biotechnology, Nonakuri West Bengal, India. *email: vasco@icb.ufmg.br

Strains	Sequencing	Reads	Assembly Software	Length (Mb)	Mapped reads (%)	Accession number	Reference
1002B	Ion PGM 200 bp	739,755	Mira v. 3.9.18	2.33511	99.70	CP012837.1	Mariano *et al.*[54]
29156	Ion PGM 200 bp	1,400,026	Newbler v. 2.9	2.33865	98.02	CP010795.1	On this work
I19	Ion PGM 400 bp	1,255,111	Spades v. 3.6.0	2.33759	99.64	CP002251.2	On this work
31	Ion PGM 400 bp	1,394,211	SPAdes 3.6.0	2.40296	99.57	CP003421.3	Viana *et al.*[55]
162	Ion PGM 200 bp	2,050,404	Newbler v. 2.9.	2.36587	98.00	CP003652.2	On this work
258	Ion PGM 200 bp	260,169	Spades v. 3.6.0	2.36982	99.41	CP003540.2	Mariano *et al.*[56]
CIP52.97	Ion PGM 400 bp	1,427,084	Mira v. 3.9.18	2.36939	99.68	CP003061.2	On this work
MB302	Ion PGM 400 bp	1,832,580	Newbler v. 2.9	2.36881	99.59	CP021982.1	Baraúna *et al.*[57]
T1	Ion PGM 200 bp	1,118,022	Newbler v. 2.9	2.3372	95.93	CP015100.1	Almeida *et al.*[58]
MB11	Ion PGM 200 bp	6,753,458	Mira 4.0.2	2.36342	99.24	CP013260.1	Baraúna *et al.*[59]

Table 1. Information on sequencing and assembling of strains.

In order to solve this assembly problem and to improve the generated data, we have the strategy of optical mapping, or Whole Genome Mapping (WGM), which is an approach that uses high-resolution restriction maps to generate the actual orientation of the organism's genome. It is the main method of large-scale genome analysis that provides complete visualization of the structural genome through a single image[12]. Optical mapping is based on the distance of the restriction sites for high precision map construction. This is a strategy in which data are obtained with greater precision since it is a physical result of the genome evaluation. This method, combined with the application of *de novo* assembly methodology, assists in the orientation of contigs[13].

The technique of optical mapping was first developed by Schwartz and collaborators in 1993, with the purpose of studying the chromosomal gene ordering of *Saccharomyces cerevisiae*[13]. Samad *et al.*, 1995, describe optical mapping as the novel approach for single-molecule DNA analysis using flowering microscopy to identify and estimate its size by the generated images[14]. Since then, several improvements have been added to the technique, especially in the images and algorithms for fragment size estimation[15]. Hence, WGM gained notoriety in several applications, such as in lineage typing in epidemic cases for clinical microbiology[16,17]; ordering of contigs generated by *de novo* assembly[7]; and in the study of inversions, insertions, deletions, duplications, and instability of bacterial genomes[18,19]. WGM has been successfully performed on very different types of organisms such as bacteria[20–22], fungi[23], plants[24–26], and mammals[27].

Regarding the genomic assembly strategy, WGM is an additional method that allows the ordering of the contigs and thus provides a size estimation of the gaps and their positions. This combination of methods is called a hybrid approach to scaffolding assembly, and it is feasible to acquire a complete genome of high quality and accuracy[7].

The work carried out by Mariano and collaborators in 2016, updated the genome of *Corynebacterium pseudotuberculosis* 1002[28] (later deposited as Cp1002B) using the optical mapping technique. *C. pseudotuberculosis* 1002B was the first organism of the *Corynebacterium* genus that had optical mapping applied in the detection and correction of assembly errors. From the results obtained by Mariano and collaborators in 2016, we decided to investigate another 10 genomes with this strategy, namely: *C. pseudotuberculosis* 29156 (Cp29156), *C. pseudotuberculosis* I19 (CpI19), *C. pseudotuberculosis* FRC41 (CpFRC41), *C. pseudotuberculosis* T1 (CpT1), *C. pseudotuberculosis* 31 (Cp31), *C. pseudotuberculosis* Cp162 (Cp162), *C. pseudotuberculosis* MB302 (CpMB302), *C. pseudotuberculosis* CIP52.97 (CpCIP52.97), *C. pseudotuberculosis* MB1 (CpMB11) and *C. pseudotuberculosis* 258 (Cp258) (Table 1). Among these genomes, 4 are strains from the biovar *ovis* and six from the biovar *equi*[29].

The strategy is to make *C. pseudotuberculosis* the most used organism in genomic studies involving the *Corynebacterium* genus. Therefore, a total of 11 strains were selected from different hosts, isolation sites and distributed between *ovis* and *equi* biovars, so that complete genomes can be made available, well assembled, and updated by new sequencing. In this manner, this data can be explored with greater reliability by future comparative studies, intraspecific evolutionary relationship analyses.

Results

Sequencing and assembly. The strains deposited by our research group were either re-sequenced (i.e., Cp1002 (1002B), CpI19, Cp31, Cp162, Cp258, CpCIP52.97), or were first sequenced using the Ion Torrent PGM™ platform (i.e., Cp29156, CpMB302, CpMB11, CpT1) (Table 1). Different software packages were used for *de novo* assembly (Table 1).

Optical mapping analysis: biovar *ovis* strains. The strains Cp1002 (CP001809.2) and CpI19 (CP002251.1) (Fig. 1A,B) showed an inversion of approximately 1.6 Mb and 1.22 Mb, respectively. It is observed in the regions flanking the first and third clusters of Ribosomal RNA in the CpI19 strain; while in Cp1002, it occurs between the first and fourth clusters. Figure 1A,B show the starting and ending points of the inversion, labeled as R1 and R2, respectively. The central block in Fig. 1 corresponds to the physical restriction map, while the upper and lower blocks represent the *in silico* restriction map generated by MapSolver™ software. The red regions of the central block (Fig. 1A,B) indicate that the same region exists in both the first version (Upper Block) and the updated version (Lower Block). Thus, they show that there was no significant loss between the compared versions in that region.

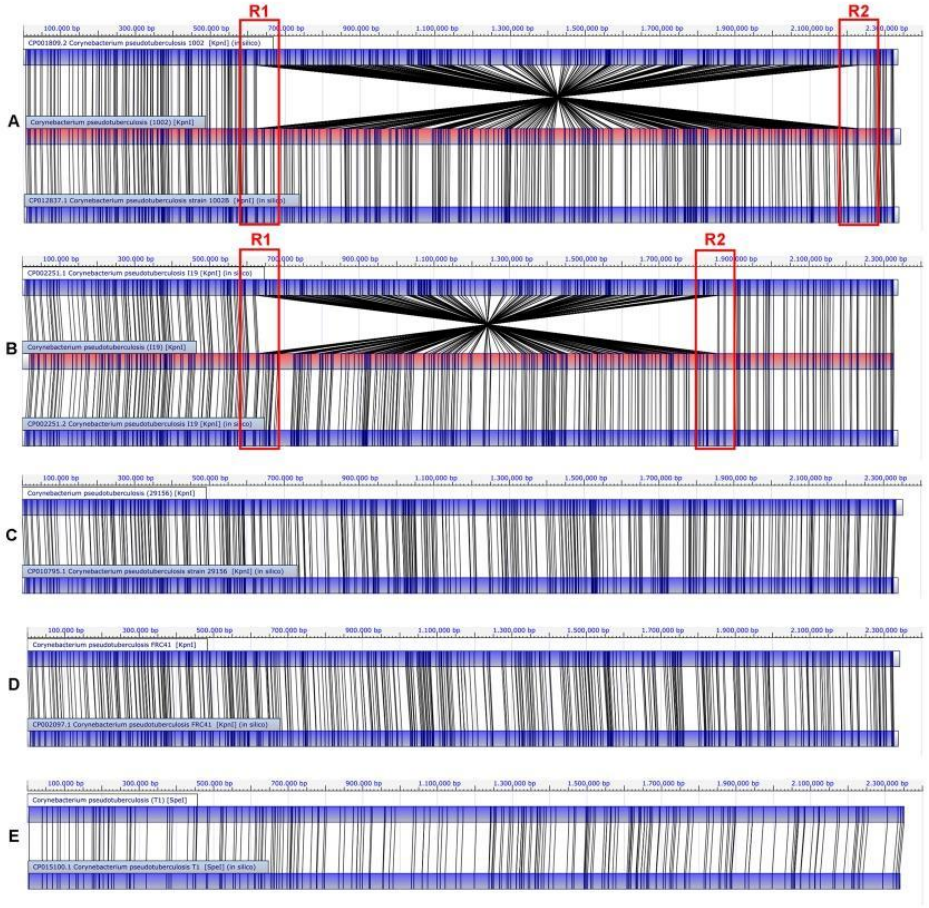


Figure 1. Optical map alignment of the selected *ovis* biovar strains. Comparisons between the first and the new version (when available), with *C. pseudotuberculosis* 1002 and 1002B (**A**); *C. pseudotuberculosis* I19 (**B**); *C. pseudotuberculosis* 29156 (**C**); *C. pseudotuberculosis* FRC41 (**D**); *C. pseudotuberculosis* T1 (**E**) are shown. R1 and R2 highlighted regions are events of inversion errors.

Strain CpFRC41 (CP002097.1) (Fig. 1D) showed a correct alignment of the restriction sites in the whole genome and thus, with no probable errors of assembly and ordering. Strains Cp29156 and CpT1 (Fig. 1C,E) were first sequenced in this work.

Optical mapping analysis: biovar *equi* strains. Figure 2A shows the alignment between the first (upper map) and the last (lower map) versions of the CpCp31 strain. The R1 region highlights the absence of corresponding restriction sites at the beginning of the genome. R2 region, in its turn, shows the absence of a chromosome region. This difference probably occurred due to errors in the assembly and gap closure process.

Worse problems were found in the previous version of Cp162 (Fig. 2B). The R1-labeled region, starting in 5′ end of the *dnaA* gene, shows no similarity between the restriction site patterns. In the R2 region, there is no linear alignment between the sites, probably due to the absence of genes. According to the optical map, an error may have occurred in the ordering of contigs in the R3 region, where the segment should be in another region of the genome. R4 region shows a ~0.85 Mb inversion in the middle of the genome, located explicitly between two clusters of ribosomal RNA.

The CpMB11 strain presented an error of choice of chromosome initiation site, in which a segment close to 5′ end of the *dnaA* gene should be situated at the end (Fig. 2C, regions R1 and R2). The Cp258 strain did not present chromosomal inversions in the deposited genome. However, a misalignment of the restriction sites (Fig. 2D, region R1) is shown next to the origin of replication at the 5′ end of the genome. Also, a region containing the genes *moaE*, *molB*, *molA*, *narI*, *narJ*, *narH*, *narG*, *narK*, *narT*, *moeY*, *moaC* is absent in the deposited chromosome (Fig. 2D, R2). The same situation occurred with the strain CpCIP5297 (Fig. 2E, regions R1, and R2). Finally, as for the Cp29156 and CpT1 strains, it is the first sequencing of the CpMB302 strain (Fig. 2F).

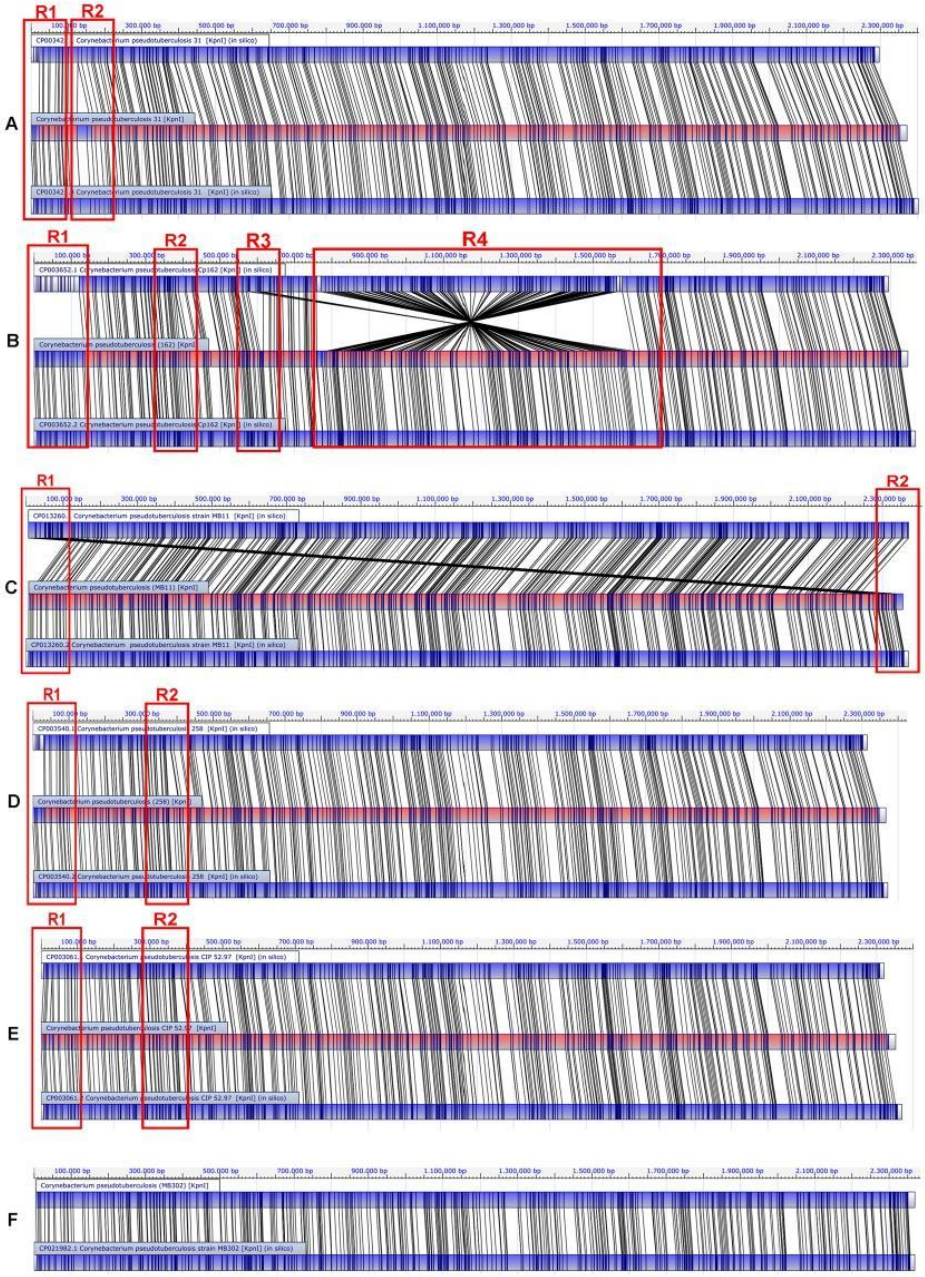


Figure 2. Optical map alignment of the selected *equi* biovar strains. Comparisons between the first and the new version (when available), with *C. pseudotuberculosis* 31 (**A**); *C. pseudotuberculosis* Cp162 (**B**); *C. pseudotuberculosis* MB11 (**C**); *C. pseudotuberculosis* 258 (**D**); *C. pseudotuberculosis* CIP52.97 (**E**); *C. pseudotuberculosis* 302 (**F**) are shown.

Content and genomes plasticity. This analysis showed a reduction of 136 bp in the total genome of the CpI19 strain, in addition to an increase of 34 CDSs and a reduction of 12 pseudogenes (Table 2). The updated version of strain Cp1002[28] showed a reduction of 6 bases in the total genome and a reduction in the number of annotated CDSs (n = 24) when compared with the older one. In this case, updating the annotation of genes identified as hypothetical protein might be the explanation. This comparison can be visualized in the map generated by BRIG, in which the last version is compared with the most updated one before the optical mapping. Strains

Isolates	Bases (bp)	CDS	Pseudogenes
I19^{1st}	2,337,730	2,095	57
I19^{2nd}	2,337,594	2,129	45
1002^{2nd}	2,335,113	2,095	47
1002B^{1st}	2,335,107	2,071	43
258^{1st}	2,314,404	2,088	46
258^{2nd}	2,369,817	2,129	34
162^{1st}	2,293,464	2,002	87
162^{2nd}	2,365,874	2,099	43
31^{1st}	2,297,010	2,063	46
31^{3rd}	2,402,956	2,173	4
CIP52.97^{1st}	2,320,595	2,060	75
CIP52.97^{2nd}	2,369,387	2,187	62

Table 2. Comparison between deposited and new version assembly of CpI19, Cp1002 (Cp1002B), Cp258, Cp162, Cp31, and CpCIP52.97 strains.

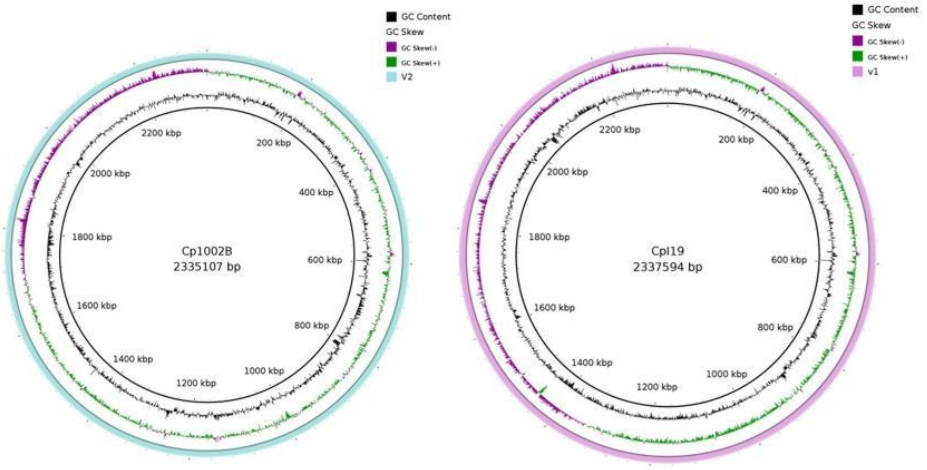


Figure 3. Comparative BRIG analysis of *ovis* biovar strains. Comparative genomic maps of the older versions (outermost circles) and their respective versions with optical map (inner black circles). (**A**) *C. pseudotuberculosis* 1002B. (**B**) *C. pseudotuberculosis* I19.

Cp1002B (Fig. 3A) and CpI19 (Fig. 3B) showed no gaps in the comparison, meaning that there were no relevant losses or gains among the versions.

The *equi* strains showed more changes among the genomes. In the Cp258 genome, obtained from the Ion Torrent PGM sequencing, an insert of ~55 kb was added. A total of 41 new CDSs and a reduction of 12 pseudogenes were included (Table 2). This had not been represented within the first genome, which was obtained by SOLiD v3 sequencing. Differences can also be visualized on the genomic map (Fig. 4A), where the highlighted red genes are important and essential genes for the strain classification e.g., the operon *nar*, with the *moaE*, *moaD*, *molB*, *molA*, *narI*, *narL*, *narG*, *narK*, *narT*, *moEY*, *mobA*, *moAc*, *moeA1* genes. Presence of genes coding for proteins such as Beta-lactamase, Vitamin K-dependent gamma-carboxylase, Heavy-metal-binding protein, Transposases, Type I restriction-modification system, and N-6 DNA Methylase is also important. Moreover, errors related to positioning and presence of CRISPR associated proteins were found among the assemblies.

The strain Cp162 also presented an increase of genomic content (~72 kb). Ninety-seven CDS and a reduction of 44 pseudogenes were found (Table 2). Figure 4B shows the absent regions of genes such as the complete cluster of the operon *nar*. Genes coding for Fe^{3+} dicitrate transport, ATP-binding protein FecE, Beta-lactamase, Vitamin K-dependent, gamma-carboxylase, Heavy-metal-binding protein, Phytoene dehydrogenase, CAAX protease self-immunity, Restriction endonuclease or methylase, Collagen-binding surface protein Cna-like, B-type domain protein, Membrane protein, ATP-dependent exonuclease, and several hypothetical proteins were also absent. A possible assembly error in the cluster of genes coding for CRISPR-associated proteins was found.

The new sequencing by Ion Torrent of the strain Cp31 resulted in the most significant increase in the gene content (~106 kb) among the selected strains. An increase of 110 CDSs and a reduction of 42 pseudogenes (Table 2) were detected. In Fig. 4C, we can highlight the inclusion of the corynephage with the tox gene of diphtheria toxin.

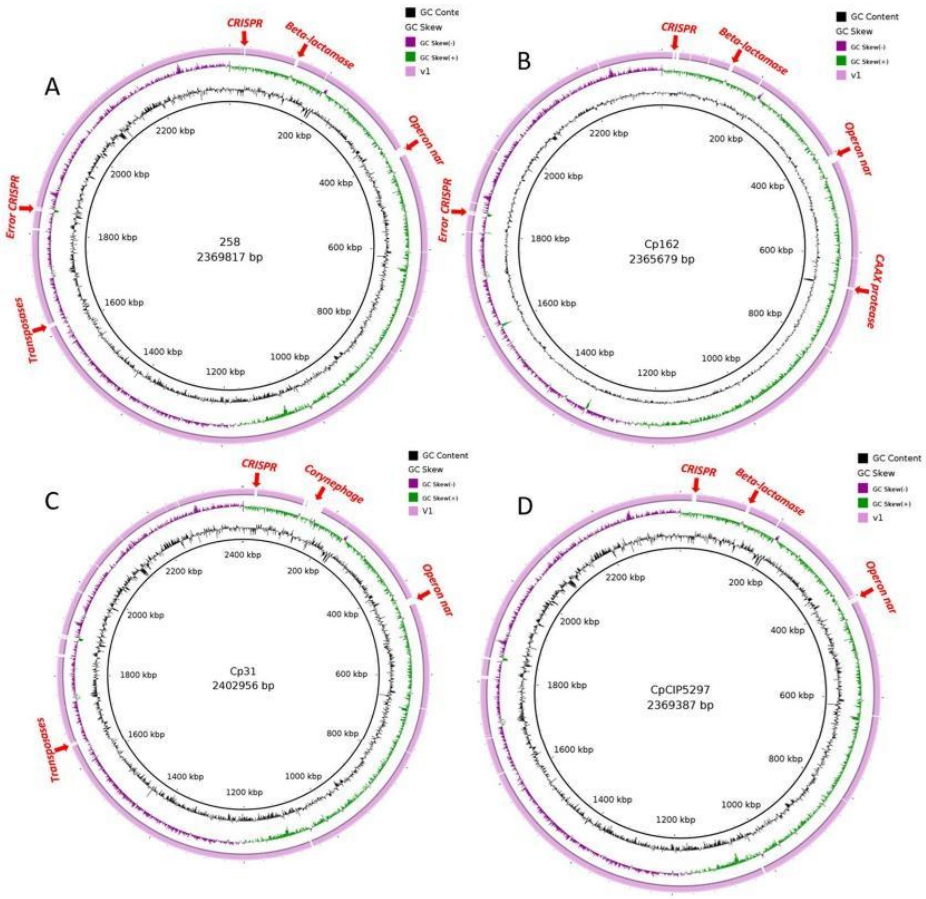


Figure 4. Comparative BRIG analysis of equi biovar strains. Comparative genomic maps of the older versions (outermost circles in purple) and their respective versions with optical map (inner black circles). (**A**) *C. pseudotuberculosis* 258. (**B**) *C. pseudotuberculosis* Cp162. (**C**) *C. pseudotuberculosis* 31. (**D**) *C. pseudotuberculosis* CIP52.97.

The CpCIP52.97 strain had an increase of ~49 kb, which represented a gain of 127 CDSs and the reduction of 13 pseudogenes (Table 2). In Fig. 4D, we can highlight the absence of essential genes, such as genes associated with CRISP (*cas2*, *cas1*, *cas3*, *cas4*, *cas5*, *cas6*, *cas7*). Once more, the operon *nar* was absent, as well as genes coding for Beta-lactamase, Phytoene dehydrogenase, integrins, transposases, and several hypothetical proteins.

With the complete and finalized genomes, a multiple alignment analysis of the genome was done using Mauve. Figure 5 shows the genomes of 5 *ovis* strains (i.e., Cp1002B, Cp29156, CpFRC41, CpI19, and CpT1). Blocks with the same color represent conserved regions, in which they share a high similarity. Inversions and rearrangement events were established as changes in the Cp1002B reference synteny (first genome of the figure). By analyzing the ends of the inverted blocks, the inversions are flanked by clusters of rRNAs. Only the CpFRC41 and CpT1 strains showed the same gene order (synteny) in their genome. Using the same strategy to compare the *equi* strains (e.g., Cp31, Cp258, Cp162, CpCIP52.97, CpMB302, and MB11), the strain Cp162 was the only one that showed an inversion and rearrangements (Fig. 6). However, these blocks are not flanked by the rRNA clusters.

Discussion

Although the optical map initially focused on typing and identification of strains without the need of sequencing, it is an efficient approach to order contigs generated by *de novo* assembly, allowing error detection and an accurate contig ordering in assemblies[30]. By definition, the optical mapping technique is a single molecule barcode using restriction enzymes, and the distances among these restriction sites, which are the basis for the alignments between optical maps and the *in-silico* maps (by contigs or genomes). These features are excellent for the scaffolding process with *de novo* assembled contigs[31]. Onmus-Leone and collaborators in 2013 applied this technique to successfully order and correct contigs generated in the *de novo* assembly step using pyrosequencing data of the bacteria *Providencia stuartii*. The mentioned authors established the strategy of using the optical map to order and correct contigs generated from short reads[32].

Figure 5. Comparative MAUVE analysis of *ovis* biovar strains. Comparison of Genome alignment of *C. pseudotuberculosis* 1002B, *C. pseudotuberculosis* 29156, *C. pseudotuberculosis* FRC41, *C. pseudotuberculosis* I19 and *C. pseudotuberculosis* T1 strains.

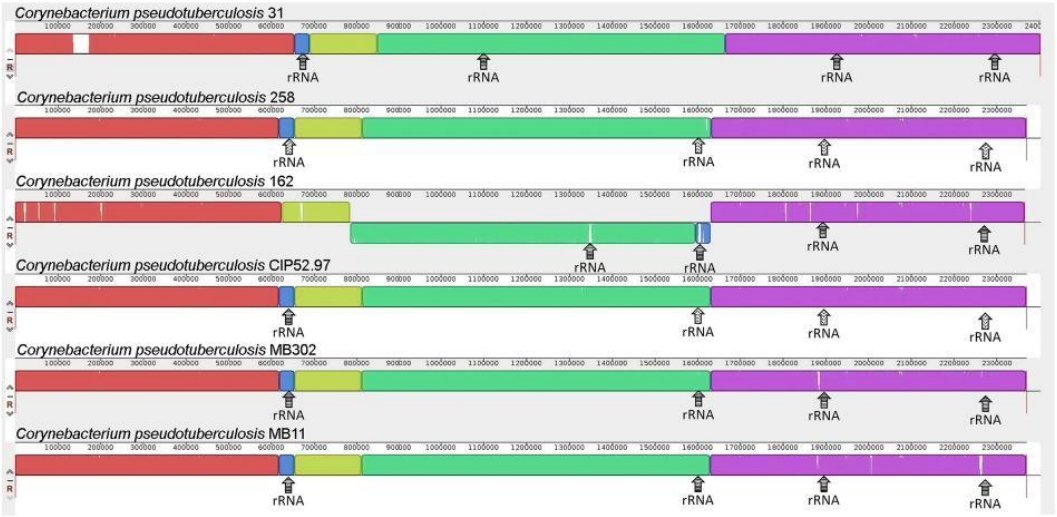


Figure 6. Comparative MAUVE analysis of *equi* biovar strains. Comparison of Genome alignment of *C. pseudotuberculosis* 31, *C. pseudotuberculosis* 258, *C. pseudotuberculosis* Cp162, *C. pseudotuberculosis* CIP52.97, *C. pseudotuberculosis* MB302 and *C. pseudotuberculosis* MB11.

Latreille and collaborators in 2007 highlighted the efficiency of the optical map suggesting it as a routine procedure in the assembly finishing process using *de novo* assembly[33]. According to these authors, it is possible to detect errors in the order and construction of the contigs by using this technique even when the organism presents several repeated regions. In the mentioned work, it was possible to finish the assembly by using optical map data even when cosmid libraries and overlapping restriction maps of BACs[33] have already been applied without success. Analyzing these results, it was possible to conclude that the optical maps are an excellent option for bacterial assembly finishing because the restriction sites cover these repetitive regions in most cases.

Repetitive regions are considered a major difficulty for *de novo* assembly algorithms, mostly in transposons and ribosomal RNA cluster regions[33]. The strain Cp1002 was sequenced by using 454 Genome Sequencer FLX (Roche), Sanger e PacBio technologies, but the error of inversion in rRNA clusters continued. Only when

sequencing with Ion Torrent PGM™ and ordering the contigs by using optical mapping data were done, the genome was correctly finished. The strain CpI19 was, at first, sequenced by using SOLiD v2 technology with 25 bp mate-paired reads and a coverage depth of 321-fold. The mate-paired technology may have contributed to the correct construction of the contigs, but due to its small size of reads sequenced, the inversion of the ribosomal RNA cluster regions occurred.

Trost and collaborators sequenced CpFRC41 strain in 2010[34] by using 454 Genome Sequencer FLX (Roche) sequencer; it was the first *C. pseudotuberculosis* strain deposited in GenBank by NCBI in 2010. The assembly was performed using *Corynebacterium diphtheriae* NCTC 13129 (BX248353.1) as the reference genome, and the gaps were closed by using Polymerase Chain Reaction (PCR) and the software r2cat[35]. Gap filling by using PCR probably contributed to no inversions found in the final genome (Fig. 1D). Schröder and collaborators in 2011 successfully used this approach in *Corynebacterium variabile* DSM 44702[36].

In CpMB11 strain (Fig. 1E) a standardization problem regarding the region of the 5′ end of the *dnaA* gene was found. Several works have shown that the most common origin of replication in bacteria is *oriC* and the first gene is the chromosomal replication initiator protein DnaA (*dnaA*)[37]. Thus, it is used as a standard pattern to linearize bacterial genomes.

Cp31 have been sequenced using several platforms: Solid v2 (CP003421.1), Ion Torrent PGM™ (CP003421.3) and PacBio. This fact confers more reliability to the assembly of this strain, which leads us to consider it as the reference strain in *equi* biovar. In its last version, published by Viana and collaborators in 2017, it was assembled using optical mapping technology. The strains Cp258, Cp162, Cp31 e CpCIP52.97 were initially sequenced in the SOLiD platform; these strains also belong to *equi* biovar and were characterized by using biochemical tests. Those were re-sequenced using Ion Torrent PGM™ platform and novel genomic regions were added to the genome sequence. Essential genes in regions, such as *nar* cluster and clusters associated with CRISPR proteins, which are only present in *equi* biovar, were added. The missing regions may be caused by an error propagation due to reference contig ordering, because of the first *equi* genome available, the Cp258[38] strain, was assembled using an *ovis* biovar strain as the reference, which does not completely present the referred clusters. The same issue happened to Cp31[39] isolate.

Similarly, the Cp162[40] strain presented the same issue because its contigs have been ordered according to *C. pseudotuberculosis* 316 (Cp316) (CP003077.1)[41]. Cp316 strain belongs to *equi* biovar and does not present *nar* operon in its former genomic sequence; it was assembled using the *ovis* strain CpFRC41 as a reference genome for contigs ordering. Presumably, the same problem happened in CpCIP52.97[42], but the genome used as a reference in the assembly process is not described in the article.

Using a complete genome deposited in public databases as a reference to assemble novel genomes is a risky strategy because even if it generates complete genomes more efficiently, it may disseminate assembly errors from one strain to other[8]. In this article, the optical map is used to validate contigs, and it is shown that extension and gap filling using read mapping or *de novo* assembly may generate assembly errors. We highlighted this technique because it does not present sequencing bias. Assembly statistics such as N50, coverage and depth coverage may generate false positive answers. Another strategy suggested by Lehri and collaborators is to use non-paired reads together with paired or long reads. A genomic region presenting assembly errors caused by insertions, deletions, inversions or rearrangements may hide significant biological variations or produce false interpretations, mostly in genomic analysis[43]. Even before the NGS platforms boomed, Schmutz and collaborators (2004) were concerned about the quality of the human genome[43], mostly because of the possible assembly errors.

The inversions caused by RNA ribosomal clusters in *ovis* biovar strains may occur due to the high similarity of these clusters. This kind of inversion has already been shown in literature in bacteria as *Salmonella paratyphi* A using restriction enzymes and pulsed-field electrophoresis gel (PFGE)[44]. The inversions among strains of the same species may be comprehended as homologous recombination events[45]. It can be highlighted that these inversions do not occur in strains belonging to *equi* biovar, except for Cp162. The inferred data concerning the genomic order of *C. pseudotuberculosis* strains were only achieved because optical mapping technology provides an accurate *in vitro* evidence.

Methods

Strain and DNA isolation and Genome sequencing. The methodology described below was applied to the novel sequencings of the strains *C. pseudotuberculosis* 1002(1002B), *C. pseudotuberculosis* 29156, *C. pseudotuberculosis* I19, *C. pseudotuberculosis* Cp162, *C. pseudotuberculosis* 258, *C. pseudotuberculosis* 31, *C. pseudotuberculosis* MB302, *C. pseudotuberculosis* MB11, *C. pseudotuberculosis* T1 and *C. pseudotuberculosis* CIP5297. The strains were cultivated in solid media with 1.5% of bacteriological agar. Subsequently, an isolated colony was used to grow in liquid media with brain-heart-infusion media (BHI-Hi Media Laboratories Pvt. Ltd, India) supplemented with 0.5% of Tween 80, at 37 °C for 20 hours under rotation. Genomic DNA was extracted following the protocol of Pacheco in 2006[46]. After the extraction step, the libraries were constructed with IonXpress™ Plus DNA Fragment Library Preparation Kit. The DNA samples were fragmented using Ion Shear™ Plus for five minutes at 37°. Then, adaptors from Ion Xpress™ Barcode Adapters kit were ligated for library quantification. Subsequently, the fragments were amplified using Ion PGM™ 200 bp or 400 bp kits. These reactions were transferred to the semiconductor chip (ION 318™ Chip v2), and it was put into Ion PGM™. During all the steps described above, all the manufacturer's instructions were strictly followed. No novel sequencing was performed for *C. pseudotuberculosis* FRC41.

Genome assembly and annotation. The analysis of the reads quality was performed by using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). No trimming was performed on reads with Phred score above 20, which were the majority. For contigs construction, we applied the *de novo* assembly strategy (no reference used) by using Mira v. 3.9.18[46], MIRA 4.0.2[46], Spades v. 3.6.0[47] e Newbler v. 2.9[48] (Table 2)

Strains	Enzyme	Length (bp)	Number of fragments	Average fragment size (bp)	Maximum fragment size (bp)	Minimum fragment size (bp)	Whole genome coverage
1002B	*Kpn1*	2,335,144	353	6,615.139	38,715	903	99.998%
29156	*Kpn1*	2,351,288	368	6,389.37	38,839	1,275	99.462%
I19	*Kpn1*	2,326,586	333	6,986.745	38,215	1,460	100.473%
FRC41	*Kpn1*	2,341,893	369	6,346.593	38,942	1,394	99.830%
31	*Kpn1*	2,372,071	346	6,855.697	35,806	1,544	101.302%
162	*Kpn1*	2,345,656	362	6,479.713	28,013	1,509	100.861%
258	*Kpn1*	2,366,195	346	6,838.714	36,249	1,667	100.153%
CIP52.97	*Kpn1*	2,352,141	347	6,778.504	36,145	1,567	100.733%
MB302	*Kpn1*	2,363,709	362	6,529.583	36,517	1,471	100.215%
T1	*Spe1*	2,350,532	193	12,178.922	54,138	1,787	99.432%
MB11	*Kpn1*	2,347,572	366	6,414.131	36,155	1,333	100.679%

Table 3. Information about the quality metrics of the optical maps used.

software. Scaffolds construction was manually performed in CLC Genomics Workbench (CLC-gw) version 7.0 (Qiagen) software using the visualization of the contigs mapped and ordered according to the restriction sites of the strains in MapSolver™ (OpGen). Then, *dnaA* gene was fixed in the probable *oriC* position in the chromosome by using an in-house python script. In order to fill the gaps and finish the assembly, GapBlaster[49] and FGAP[50] software was used and subsequently, the contigs were mapped to the scaffold or a reference genome by using CLC Genomics (Qiagen). The annotation was performed by using in-house scripts for annotation transference from *C. pseudotuberculosis* strains, which were manually curated in the UniProt database (http://uniprot.org). Finally, pseudogenes were manually curated by using Artemis software[51] and CLC Genomics (Qiagen).

Optical mapping. The optical maps were acquired from Opgen, Inc. The MapSolver™ (OpGen Inc.) software was used for the comparison of the physical restriction map and the restriction sites present in the assembled genome. Several pieces of information about metrics for the quality of each optical map are available in Table 3.

Genome plasticity and genetics content. This analysis was performed using genomic sequences before and after assembly assisted by optical mapping data. The maps comparing different versions of the studied strains were generated by using Blast Ring Image Generator (BRIG) v0.95[52]. For inversion, deletion and rearrangement analysis, the Mauve v. 2.3.1[53] software was used with progressiveMauve[53] option set.

Conclusion

The results obtained from optical mapping data analysis pointed errors in the assemblies of *C. pseudotuberculosis* genomes deposited in Genbank. Thus, the optical map was efficient in the assembly error detection of the strains Cp1002, CpI19, Cp31, Cp162, CpMB11, Cp258, and CpCIP52.97. Regarding the novel genomes, such as Cp29156, CpT1, and MB302, the optical map data contributed in the contigs ordering step, which contributed to a more efficient assembly finishing considering there are no assembly errors in the final version of the genomes. Furthermore, the update of the genomic sequences done by re-sequencing the genomes with Ion Torrent PGM™ platform was essential to the relevant genomic content increase, which happened in the strains previously sequenced using the SOLiD platform. We also pointed out several inversions caused by ribosomal RNA gene clusters in strains of the *ovis* biovar. Thus, we can suggest that the genomes deposited after applying this strategy made these strains more reliable for novel studies.

Received: 4 January 2019; Accepted: 18 October 2019;
Published online: 08 November 2019

References

1. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–80 (2005).
2. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
3. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
4. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 256 (2016).
5. Bertsch, J. *et al.* GOLD:Genomes Online Database. GOLD Statistics (2018). Available at, https://gold.jgi.doe.gov/statistics. (Accessed: 8th August 2018).
6. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* **10**, 599–606 (2012).
7. Lehri, B., Seddon, A. M. & Karlyshev, A. V. The hidden perils of read mapping as a quality assessment tool in genome sequencing. *Sci. Rep.* **7**, 43149 (2017).
8. Narzisi, G. & Mishra, B. Comparing De Novo genome assembly: The long and short of it. *PLoS One* **6** (2011).
9. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–9 (2011).
10. Mariano, D. C. *et al.* MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes. *Bioinformation* **11**(6), 276–279 (2015).
11. Waters, N. R., Abram, F., Brennan, F., Holmes, A. & Pritchard, L. riboSeed: leveraging prokaryotic genomic architecture to assemble across ribosomal regions. *Nucleic Acids Res.* **46**, e68–e68 (2018).

12. Wu, C., Schramm, T. M., Zhou, S., Schwartz, D. C. & Talaat, A. M. Optical mapping of the *Mycobacterium avium* subspecies *paratuberculosis* genome. *BMC Genomics* **10**, 25 (2009).
13. Schwartz, D. C. *et al.* Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110–4 (1993).
14. Samad, A., Huff, E. F., Cai, W. & Schwartz, D. C. Optical mapping: a novel, single-molecule approach to genomic analysis. *Genome Res.* **5**, 1–4 (2007).
15. Reslewic, S. *et al.* Whole-Genome Shotgun Optical Mapping of *Rhodospirillum rubrum* Whole-Genome Shotgun Optical Mapping of Rhodospirillum rubrum. *Appl. Environ. Microbiol. 2005* **71**, 5511 (2005).
16. Kotewicz, M. L., Mammel, M. K., LeClerc, J. E. & Cebula, T. A. Optical mapping and 454 sequencing of *Escherichia coli* O157: H7 isolates linked to the US 2006 spinach-associated outbreak. *Microbiology* **154**, 3518–3528 (2008).
17. Petersen, R. F. *et al.* Molecular Characterization of *Salmonella Typhimurium* Highly Successful Outbreak Strains. *Foodborne Pathog. Dis.* **8**, 655–661 (2011).
18. Sabirova, J. S., Xavier, B. B., Ieven, M., Goossens, H. & Malhotra-Kumar, S. Whole genome mapping as a fast-track tool to assess genomic stability of sequenced *Staphylococcus aureus* strains. *BMC Res. Notes* **7**, 1–6 (2014).
19. Shukla, S. K. *et al.* Comparative whole-genome mapping to determine *Staphylococcus aureus* genome size, virulence motifs, and clonality. *J. Clin. Microbiol.* **50**, 3526–3533 (2012).
20. Zhou, S. *et al.* A whole-genome shotgun optical map of *Yersinia pestis* strain KIM. *Appl. Environ. Microbiol.* **68**, 6321–6331 (2002).
21. Zhou, S. Whole-Genome Shotgun Optical Mapping of *Rhodobacter sphaeroides* strain 2.4.1 and Its Use for Whole-Genome Shotgun Sequence Assembly. *Genome Res.* **13**, 2142–2151 (2003).
22. Lin, J. Whole-Genome Shotgun Optical Mapping of *Deinococcus radiodurans*. *Science.* **285**, 1558–1562 (1999).
23. Olsen, R. A. *et al.* De novo assembly of *Dekkera bruxellensis*: a multi technology approach using short and long-read sequencing and optical mapping. *Gigascience* **4**, 56 (2015).
24. Chamala, S. *et al.* Assembly and validation of the genome of the nonmodel basal *angiosperm Amborella*. *Science.* **342**, 1516–1517 (2013).
25. Zhou, S. *et al.* Validation of rice genome sequence by optical mapping. *BMC Genomics* **8**, 278 (2007).
26. Tang, H., Lyons, E. & Town, C. D. Optical mapping in plant comparative genomics. *Gigascience* **4**, 1–6 (2015).
27. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* **31**, 135–41 (2013).
28. Mariano, D. C. B. *et al.* Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of *Corynebacterium pseudotuberculosis* strain 1002. *BMC Genomics* **17**, 1–7 (2016).
29. Oliveira, A. *et al.* Insight of Genus *Corynebacterium*: Ascertaining the Role of Pathogenic and Non-pathogenic Species. *Front. Microbiol.* **8**, 1937 (2017).
30. Bogas, D. *et al.* Applications of optical DNA mapping in microbiology. *Biotechniques* **62**, 255–267 (2017).
31. Mendelowitz, L. & Pop, M. Computational methods for optical mapping. *Gigascience* **3**, 1–7 (2014).
32. Onmus-Leone, F. *et al.* Enhanced De Novo Assembly of High Throughput Pyrosequencing Data Using Whole Genome Mapping. *PLoS One* **8**, 2–10 (2013).
33. Latreille, P. *et al.* Optical mapping as a routine tool for bacterial genome sequence finishing. *BMC Genomics* **8**, 321 (2007).
34. Trost, E. *et al.* The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC Genomics* **11**, 728 (2010).
35. Husemann, P. & Stoye, J. r2cat: Synteny plots and comparative assembly. *Bioinformatics* **26**, 570–571 (2009).
36. Schröder, J., Maus, I., Trost, E. & Tauch, A. Complete genome sequence of *Corynebacterium variabile* DSM 44702 isolated from the surface of smear-ripened cheeses and insights into cheese ripening and flavor generation. *BMC Genomics* **12**, 545 (2011).
37. Eisen, J. A., Heidelberg, J. F., White, O. & Salzberg, S. L. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* **1**, RESEARCH0011 (2000).
38. Soares, S. C. *et al.* Genome sequence of *Corynebacterium pseudotuberculosis* biovar *equi* strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. *J. Biotechnol.* **167**, 135–41 (2013).
39. Silva, A. *et al.* Complete genome sequence of *Corynebacterium pseudotuberculosis* Cp31, isolated from an Egyptian buffalo. *J. Bacteriol.* **194**, 6663–6664 (2012).
40. Hassan, S. S. *et al.* Whole-genome sequence of *Corynebacterium pseudotuberculosis* strain Cp162, isolated from camel. *J. Bacteriol.* **194**, 5718–5719 (2012).
41. Ramos, R. T. J. *et al.* Genome Sequence of the *Corynebacterium pseudotuberculosis* Cp316 Strain, Isolated from the Abscess of a Californian Horse. *J. Bacteriol.* **194**, 6620–6621 (2012).
42. Cerdeira, L. T. *et al.* Complete genome sequence of *Corynebacterium pseudotuberculosis* strain CIP 52.97, isolated from a horse in Kenya. *J. Bacteriol.* **193**, 7025–7026 (2011).
43. Salzberg, S. L. & Yorke, J. A. Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320–4321 (2005).
44. Liu, S. L. & Sanderson, K. E. The chromosome of *Salmonella paratyphi* A is inverted by recombination between rrnH and rrnG. *J. Bacteriol.* **177**, 6585–6592 (1995).
45. Raeside, C. *et al.* Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with *Escherichia coli*. *MBio* **5**, e01377–14 (2014).
46. Chevreux, B., Wetter, T. & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma, 45–56, 10.1.1.23/7465 (1999).
47. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
48. 454 Life Sciences Corp. Local Newbler 2.9 documentation, https://hpc.wm.edu/software/docs/newbler/index.html (2013).
49. de Sá, P. H. C. G. *et al.* GapBlaster—A Graphical Gap Filler for Prokaryote Genomes. *PLoS One* **11**, e0155327 (2016).
50. Piro, V. C. *et al.* FGAP: an automated gap closing tool. *BMC Res. Notes* **7**, 371 (2014).
51. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464–469 (2012).
52. Alikhan, N. F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics* **12**, 402 (2011).
53. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS One* **5**, e11147 (2010).
54. Mariano, D. C. B., Ramos, R. T. J. & Azevedo, V. A. D. C. Montagem e finalização de genomas procariotos com mapeamento óptico. *Novass* **76** (2016).
55. Viana, M. V. C. *et al.* Comparative genomic analysis between Corynebacterium pseudotuberculosis strains isolated from buffalo. *PLoS One* **12**, e0176347 (2017).
56. Mariano, D. C. B. *et al.* SIMBA: A web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology. *BMC Bioinformatics* **17**, 456 (2016).
57. Baraúna, R. A. *et al.* Assessing the Genotypic Differences between Strains of *Corynebacterium pseudotuberculosis* biovar *equi* through Comparative Genomics. *PLoS One* **12**, e0170676 (2017).

58. Almeida, S. *et al*. Complete Genome Sequence of the Attenuated *Corynebacterium pseudotuberculosis* Strain T1. *Genome Announc.* **4**, e00947–16 (2016).
59. Barauna, R. A. *et al*. Genomic analysis of four strains of *Corynebacterium pseudotuberculosis* bv. *equi* isolated from horses showing distinct signs of infection. *Stand. Genomic Sci*. **12**, 16 (2017).

Acknowledgements
The author's thanks CNPq, FAPEMIG, CAPES, and PRPq by the support. This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais (PRPq).

Author contributions
T.d.J.S. performed genome assemblies, optical mapping analysis, interpreted all data regarding and was a major contributor in writing the manuscript. D.P. and M.T.D.P. performed genome assemblies, optical mapping analysis, and were contributors in writing the manuscript. R.P. performed content and genomes plasticity analysis and was a contributor in writing the manuscript. F.L.P., H.C.P.F., and B.B. performed genome sequencing and were contributors in writing the manuscript. A.C.P.G., R.B.K., R.R., A.S., P.G., D.B., A.G.N., V.A. contributed to data interpretation and were major contributors in writing the manuscript.

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

Additional information
Correspondence and requests for materials should be addressed to V.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.11.1 Discussão ampliada do Artigo I

Ao comparar as linhagens do biovar *equi* (Cp31, Cp258, Cp162, CpCIP52.97, CpMB302 e MB11), a linhagem Cp162 foi a única que apresentou inversão e rearranjo (Figura 6 do artigo I da tese). No entanto, esses blocos não são flanqueados pelos *clusters* de rRNA, pois é comum encontrar esse tipo de inversão em bactérias e é difícil separar esses eventos de rearranjos, levando a equívocos nas montagens. A posição dos blocos submetidos à inversão e rearranjo foi confirmada pelo mapa óptico e pela análise de sintenia, pois de acordo com o padrão dos sítios de restrição o contig poderia ser alinhado somente nesta região.

Com o intuito de descrever melhor esse evento de rearranjo cromossômico, enviamos as 6 linhagens do biovar *equi* do artigo I (*C. pseudotuberculosis* 162, 31, CIP52.97, 258, MB11 e MB302) para sequenciamento na plataforma Illumina Hiseq 2500, com inserto de 450 pb (*paired-end)*. A qualidade do sequenciamento foi excelente, com valores de *Phred score* entre 38 e 40. Após essa etapa, as leituras foram submetidas ao processo de montagem utilizando o programa Edena V3.131028 e os *contigs* foram alinhados com o mapa óptico de cada uma. Com os *contigs* ordenados, o fechamento de *gaps* foi realizado pelo Gfinisher (GUIZELINI *et al.*, 2016) e GapBlaster (DE SÁ *et al.*, 2016).

Para investigar o possível evento de rearranjo ocorrido na linhagem Cp162, foi realizada uma análise de sintenia comparativa através do programa Mauve (DARLING; MAU; PERNA, 2010) com as linhagens *C. pseudotuberculosis* 162, 31, CIP52.97, 258, MB11, MB302, I37, 262 e 1002B. As linhagens I37 e 262 (ambas do biovar *equi*) não possuem mapa óptico, mas foram selecionadas por apresentarem maior similaridade com a Cp162. Já a linhagem 1002B foi selecionada por ser o modelo do biovar *ovis* e assim pode-se apurar se o evento de rearranjo também ocorre na mesma. Na Figura 8, observa-se que, na região R1, há um *cluster* com 32 genes (descritos na Tabela 1), flanqueado por dois genes codificadores de proteína transposase. Esse evento de transposição gênica foi identificado apenas na linhagem Cp162 (Figura 8B), sendo visualizado por meio da análise de sintenia gerada pelo Mauve. Na mesma região nos demais genomas, o gene para transcrição de transposase não foi detectado.

Por meio desses elementos, as bactérias mantêm um equilíbrio entre integridade e instabilidade do próprio genoma. Para isso, os cromossomos bacterianos são complexos e dinâmicos, características que dão flexibilidade e plasticidade a esses organismos. A instabilidade do genoma pode ser resultante de mutações pontuais ou de rearranjos, como deleções, duplicações, inserções, inversões ou translocações (KRAWIEC; RILEY, 1990). Existem vários tipos de elementos genéticos especializados que desempenham um papel na

instabilidade genômica bacteriana. Podemos citar como exemplo de elementos móveis: sequências de inserção [SIs], transposons, ilhas genômicas e integrons (DARMON; LEACH, 2014).

No evento apontado na Figura 8, podemos descartar a chance de presença de uma ilha genômica, segundo o resultado predito pelo programa GIPSy, ao comparar a Cp162 com a *C. glutamicum* linhagem ATCC 13032 (NZ_CP025533.1). É possível, que devido ao flanqueamento pelas transposases nas extremidades do *cluster*, esse rearranjo tenha sido mediado por elementos transponíveis que possuem repetições terminais invertidas (RIs) curtas e usam transposases para reconhecer e processar as extremidades desses elementos(DARMON; LEACH, 2014).

Os elementos transponíveis geralmente duplicam a sequência de destino na qual eles se integram, criando uma duplicação no sítio de destino. A seleção dessa região de destino é uma função da transposase, sendo que os locais de destino de alguns elementos são bem específicos como é o caso do transposon bacteriano Tn7, por exemplo (CRAIG, 1991). Esse evento de rearranjo não havia sido relatado anteriormente, pois se encontrava em outra região devido a um erro de montagem (Figura 8A), identificado somente a partir da análise do mapa óptico.

Vale ainda ressaltar que existe uma pequena região (R2) no começo desse *cluster* em análise (Figura 8B), que possui uma transposição na linhagem Cp1002B. Nesta região R2, temos os seguintes genes codificantes: *hypothetical protein*, DEAD/DEAH *box helicase*, DUF3239 *domain-containing protein*, *pyridoxal phosphate-dependent aminotransferase*, 16S Ribosomal RNA, 23S Ribosomal RNA e 5S ribosomal RNA (Tabela 1, abaixo). Como foi demonstrado no artigo I da tese, as linhagens do biovar *ovis* possuem rearranjos cromossômicos causados por *clusters* de rRNAs ribossomais.

Desta forma, este foi o primeiro trabalho dentro do gênero *Corynebacterium* em que foi possível distinguir rearranjos cromossômicos naturais de erros de montagens através do mapa óptico, para obtenção de uma montagem completa precisa e acurada. Assim, chegamos à evidência, a partir da genômica estrutural, de um *cluster* único na linhagem Cp162, o que pode nos levar a discutir eventos evolutivos dentro da mesma espécie em próximos trabalhos.

Figura 8 - Análise da sintenia entre linhagens de *C. pseudotuberculosis* do biovar *ovis* e *equi*, a partir do Mapsolver (A) e Mauve (B).

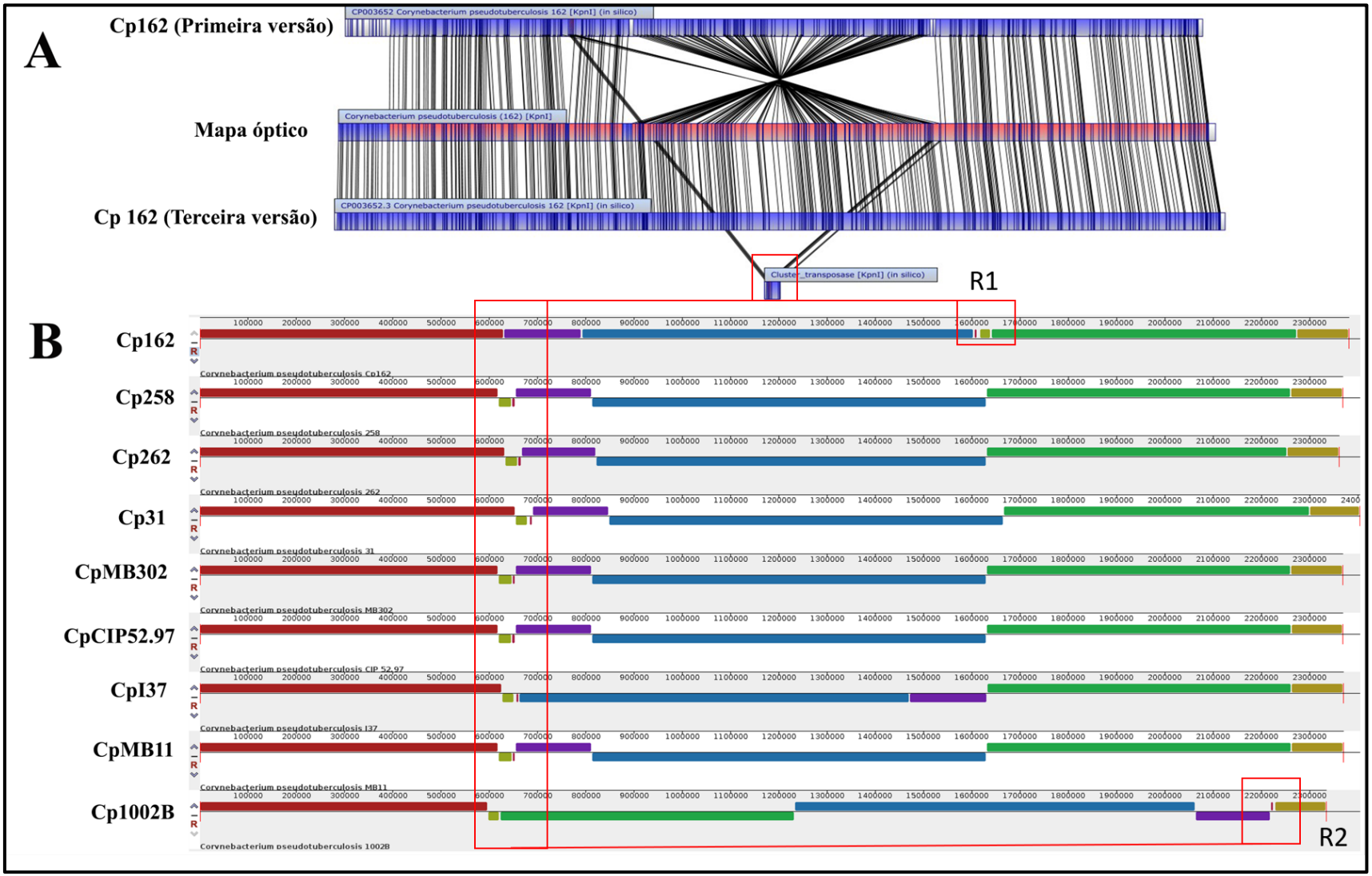


Tabela 1- Predição de todos os genes presente no *cluster* que ocorre à inversão na linhagem *C. pseudotuberculosis* 162.

Gene	Protein_id	Produto gênico	Posição do gene
	WP_048653436.1	*IS110 family transposase*	1603613..1604824
	WP_080577978.1	*hypothetical protein*	1605096..1607594
	WP_014800171.1	*DEAD/DEAH box helicase*	1607709..1609358
	WP_013241448.1	*DUF3239 domain-containing protein*	1609425..1610066
	WP_014800170.1	*pyridoxal phosphate-dependent aminotransferase*	1610066..1611238
rRNA	-	*16S ribosomal RNA*	1611456..1612976
rRNA	-	*23S ribosomal RNA*	1613341..1616417
rRNA	-	*5S ribosomal RNA*	1616526..1616642
	WP_014801140.1	*hypothetical protein*	1616737..1617051
	WP_013241445.1	*ATP-binding cassette domain-containing protein*	c(1617423..1618178)
	WP_014401050.1	*iron chelate uptake ABC transporter family permease subunit*	c(1618175..1619233)
	WP_013241443.1	*iron chelate uptake ABC transporter family permease subunit*	c(1619226..1620191)
	WP_041481198.1	*ABC transporter substrate-binding protein*	c(1620307..1621299)
	WP_014800166.1	*carbonic anhydrase*	c(1621907..1622629)
tmRNA	-	*transfer-messenger RNA*	c(1623996..1624377)
smpB	WP_014800164.1	*SsrA-binding protein SmpB*	c(1624514..1625005)
	WP_014366662.1	*ABC transporter permease*	c(1625113..1626015)
ftsE	WP_014800163.1	*Cell division ATP-binding protein FtsE*	c(1626037..1626726)
	WP_014366660.1	*hypothetical protein*	1626719..1626898
	WP_014800162.1	*AbgT family transporter*	1626985..1628613
	WP_014800161.1	*topology modulation protein*	1628674..1629147
prfB	WP_014800160.1	*peptide chain release factor 2*	1629181..1630281
	WP_014800159.1	*inositol monophosphatase family protein*	1630395..1631279
hisN	WP_072577785.1	*histidinol-phosphatase*	1631321..1632127
	WP_014800157.1	*hypothetical protein*	c(1632202..1633386)
	WP_013241430.1	*biotin/lipoyl-binding protein*	c(1633482..1633844)
	WP_013241429.1	*hypothetical protein*	c(1633869..1634132)
	WP_014800156.1	*acyl-CoA carboxylase subunit beta*	c(1634142..1635698)
	WP_041481197.1	*methylmalonyl-CoA carboxytransferase subunit 5S*	c(1635714..1637195)
	WP_014800154.1	*amino acid permease*	c(1637501..1638940)
	WP_013241424.1	*hypothetical protein (frameshifted)*	c(1639321..1639610)
	WP_048653436.1	*IS110 family transposase*	1640073..1641284

CAPÍTULO II

5. CAPÍTULO II

5.1 Objetivos específicos do capítulo II

a) Realizar o sequenciamento e a montagem dos genomas de 45 novas linhagens de *C. pseudotuberculosis* isoladas de caprinos e ovinos provenientes de locais diferentes, com a tecnologia Illumina Hiseq 2500 *paired-end*;
b) Realizar o re-sequenciamento e a montagem dos genomas de 11 linhagens de *C. pseudotuberculosis,* sendo 5 exemplares do biovar *ovis* com a tecnologia Illumina Hiseq 2500 *paired-end* para avaliar a influência da plataforma de sequenciamento no genoma final;
c) Analisar todos genomas das linhagens de *C. pseudotuberculosis* do biovar *ovis* sequenciados pela plataforma Illumina Hiseq 2500 *paired-end*, para identificar marcadores moleculares únicos relacionados ao hospedeiro e/ou habitat de isolamento.
d) Obter genomas completos e de alta acurácia para que sejam utilizados como referência na ordenação durante o pipeline de montagem de novas linhagens de *C. pseudotuberculosis*, aumentando assim a eficiência dessa etapa;
e) Realizar novas análises comparativas para identificar possíveis *clusters* de genes que não foram evidenciados anteriormente com as primeiras versões dos genomas.

5.2 Estrutura do Capítulo II

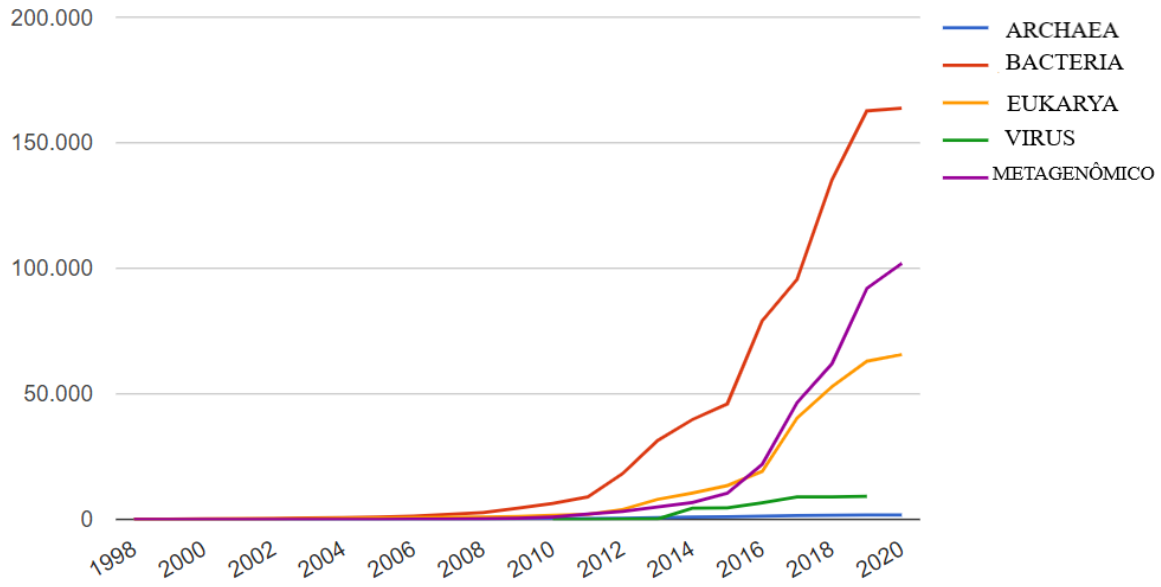
O capítulo II é focado em análises de genômica comparativa. Inicialmente, teremos a introdução aos conceitos de genômica comparativa e pan-genômica, além do capítulo de livro: **Chapter 5 - Pan-genomics of veterinary bacteria and its applications,** que faz parte do livro: **Pan-genomics: Applications, Challenges, and Future Prospects**. Publicado pela revista **Elsevier book** em fevereiro de 2020.

Em seguida, um segundo artigo da tese intitulado: **New genomics discoveries through re-sequencing and correction of assembly in *Corynebacterium pseudotuberculosis* genomes,** que será submetido à revista **Frontiers in Genetics section livestock Genomics,** fator de impacto de 3.517 (2018).

5.3 – Introdução à genômica comparativa e pan-genômica.

Após o desenvolvimento da NGS em 2005, o número de genomas sequenciados aumentou exponencialmente. Compondo essa crescente, o domínio Bactéria é o maior representante dos projetos, mesmo após 15 anos, mostrando a importância do mesmo para a comunidade científica mundial (Figura 9) (METZKER, 2010). Assim, projetos voltados ao estudo de grupos de organismos tornaram-se viáveis. A revolução proporcionada pela NGS trouxe consigo um imenso volume de genomas, sendo a maioria constituída de versões *drafts* ou incompletas (LEHRI; SEDDON; KARLYSHEV, 2017).

Figura 9 - Gráfico do Genomes Online Database (GOLD) em relação aos projetos genômicos por ano e domínio.



Fonte: *Genomes OnLine Database, GOLD statistics* (2020).

Análises comparativas entre genomas ganham robustez à medida que mais organismos são sequenciados. Dessa forma novas hipóteses sobre evolução, bioquímica, genética, metabolismo e vias fisiológicas dos organismos podem ser sugeridas (SIVASHANKARI; SHANMUGHAVEL, 2007).

A genômica comparativa é fundamentada na comparação, geralmente por alinhamento entre sequências, de um conjunto de genes ou proteínas entre organismos procurando compreender particularidades genéticas relacionadas, como por exemplo, divergências evolutivas. A partir desses dados são identificadas e caracterizadas regiões conservadas e não conservadas, seja em organismos procarióticos ou eucarióticos (EDWARDS; HOLT, 2009).

A genômica comparativa aplicada a dados bacterianos recebe destaque devido à diversidade fenotípica e genética desses organismos, seja intraespecífica ou interespecífica. Além disso, as bactérias habitam ambientes diversos, possuem amplo espectro de hospedeiros e exibem estilos de vida distintos, o que gera mais opções de comparações (MERHEJ *et al.*, 2009).

Por meio da genômica comparativa, podemos explorar evidências do processo evolutivo, tais como: deleções, inserções, ganho e perda de material genético por transferência horizontal ou vertical, mutações pontuais e inversões cromossômicas. Também é possível identificar características padrão com o objetivo de caracterizar grupos, como: número de exons e íntrons, distribuição de nucleotídeos (conteúdo GC), uso de códons preferenciais em genes transcritos, frequência de nucleotídeos e conservação ou compartilhamento de domínios funcionais.

A contemplação de todas essas análises de forma conjunta possibilita a identificação de ilhas genômicas (patogênicas, simbióticas, metabólicas ou de resistência), genes homólogos, e/ou identificação de mutações específicas. A partir da identificação desses elementos, podemos chegar ao desenvolvimento de métodos profiláticos ou de controle como drogas para doenças infecciosas de ação rápida ou crônica (PROSDÓCIMI; MOREIRA, 2015).

5.4 Pan-genômica

Com uma visão mais abrangente e envolvendo conceitos de genômica comparativa, tem-se a estratégia de estudo do pan-genoma. O primeiro a aplicar essa metodologia foi Tettelin e colaboradores (2005), em um estudo com oito linhagens de *Streptococcus agalactiae* (TETTELIN *et al.*, 2005). Desde então, essa área vem crescendo com estudos de diversas bactérias, inclusive Rouli e colaboradores (2015) a caracterizaram como uma das principais ferramentas para estudos de patogenicidade em bactérias (ROULI *et al.*, 2015), além de fornecer uma base sólida de dados para estudos de evolução baseados em filogenômica e reconstruções de árvores filogenéticas. Por exemplo, é possível identificar eventos de transferência horizontal de genes, os quais podem modificar o perfil genético de uma bactéria, conferindo-lhe resistência a antibióticos ou habilidade para infectar novos hospedeiros (SOUCY; HUANG; GOGARTEN, 2015).

As bactérias, de modo geral, conseguem modificar seu material genético de forma bastante eficiente, sendo esta capacidade definida como plasticidade genômica, como: aquisição de plasmídeos, por transposons e as ilhas genômicas. Contudo, essas regiões apresentam características que podem ser identificadas por meio da genômica comparativa, que são: conteúdo de GC distinto do organismo original, uso de códon dos genes presentes nas ilhas genômicas,

presença de sequências de inserção, regiões flanqueadas por tRNAs ou transposases (BELLANGER *et al.*, 2014).

Dentro do resultado de uma análise de pan-genômica, os genes codificadores de proteínas são divididos em três partes: genoma central (*core* genes); genoma acessório ou dispensável (*shared* genes); e genes específicos da espécie ou da linhagem (*singleton* genes). Além disso, o pan-genoma é considerado "aberto" desde que novos genes sejam adicionados significativamente ao repertório total para cada novo genoma adicional e "fechado" quando os genomas recém-adicionados não influenciam para aumentar significativamente o repertório total dos genes (GUIMARÃES *et al.*, 2015; TETTELIN *et al.*, 2005).

Para caracterizar o pan-genoma como fechado ou aberto é aplicada a lei de Heap (TETTELIN *et al.*, 2008), que usa o valor de α. Quando α é menor que 1, diz-se que o pan-genoma está aberto, já quando o α é maior que 1, o pan-genoma está fechado. Contudo, esse conceito não é consensual, pois, devido a eventos de transferência horizontal, bactérias podem incorporar material genético por meio de plasmídeos ou transposons. Com esses elementos móveis, ilhas genômicas podem ser adquiridas e, consequentemente, genes de resistência e virulência. Assim, esse conceito pode variar de acordo com o ambiente ao qual essas bactérias estão expostas. Além disso, patógenos intracelulares podem perder genes no processo de adaptação ao organismo hospedeiro, mantendo apenas os genes essenciais (DE BARSY *et al.*, 2016).

Para analisar os genomas de forma que todos sejam comparados contra todos, métodos computacionais são desenvolvidos com o intuito de identificar genes ortólogos, levando em consideração as mutações sítio-específicas e identificar os genes compartilhados com base na homologia. Programas como PGAP – *Pan-Genomes Analysis Pipeline* (ZHAO *et al.*, 2012), PanWeb - *A web interface for pan-genomic analysis* (PANTOJA *et al.*, 2017), BPGA - *An ultra-fast pan-genome analysis pipeline* (CHAUDHARI; GUPTA; DUTTA, 2016), PanOCT (FOUTS *et al.*, 2012) e Roary - *The pan genome pipeline* (PAGE *et al.*, 2015) são exemplos de estratégias utilizadas para as análises pan-genômicas.

O Roary se destaca por ser otimizado para genomas procariotos, com execução rápida, além de poder ser usado em pequenos conjuntos de dados em um computador pessoal ou grandes conjuntos de dados em *workstations* com grande capacidade computacional, trabalhando sempre em multicores tirando o máximo da capacidade de cada dispositivo. Isso faz com que o mesmo seja mais rápido e consuma menos memória RAM que os demais citados anteriormente. O *software* é construído na linguagem *perl* e tem vários *scripts* em *perl* e *python* para personalizar o resultado final, fornecendo ao usuário uma excelente versatilidade nas análises. O Roary é um programa livre e está disponível no link: http://sanger-pathogens.github.io/Roary.

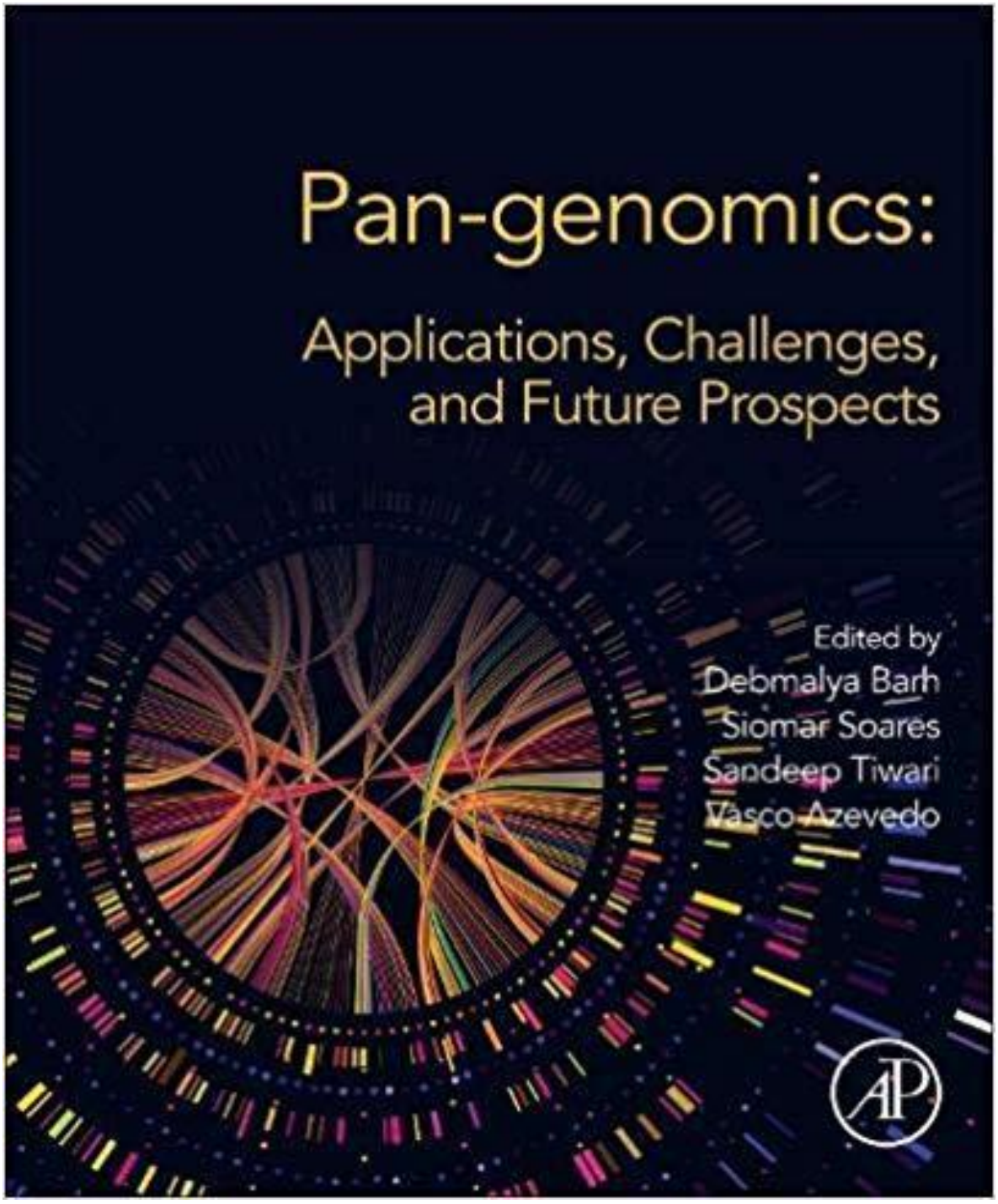
Outro programa com destaque dentro da comunidade científica é o BPGA (CHAUDHARI; GUPTA; DUTTA, 2016), pois apresenta um pipeline de execução rápida. O BPGA oferece uma série de recursos inovadores para análises de genômica comparativa, como filogenia com o genoma central, MultiLocus Sequence Typing (MLST), presença e/ou ausência de genes, análises de KEGG – *Kyoto* Encyclopedia of Genes and Genomes e COG database – *A tool for genome-scale analysis of protein functions and Evolution*, além da identificação de genes essenciais, acessórios e únicos.. O BPGA pode ser aplicado em dados de metagenomas, vírus, plantas e outros, e está disponível pelo link: https://iicb.res.in/bpga/index.html.

Devido à grande importância da genômica comparativa como ferramenta de análise, abordaremos na próxima seção uma revisão de conceitos e estudos relevantes já realizados no contexto de bactérias patogênicas de interesse médico-veterinário, incluindo *C. pseudotuberculosis,* que é o modelo adotado neste manuscrito.

5.5 Book - Pan-genomics: Applications, Challenges, and Future Prospects:

Com o intuito de disseminar conceitos e relatar estudos relevantes em genômica comparativa, apresenta-se neste capítulo de livro uma revisão da literatura sobre pesquisas de pan-genômica e genômica comparativa de bactérias patogênicas que causam doenças veterinárias, incluindo algumas responsáveis por zoonoses, como: *Corynebacterium pseudotuberculosis*; *Corynebacterium ulcerans*; *Streptococcus suis*; *Brachyspira hyodysenteriae*; *Moraxella bovoculi*; *Pasteurella multocida*; *Mannheimia haemolytica*; *Clostridium botulinum*; *Campylobacter*; *Streptococcus agalactiae*; *Francisella tularensis*; *Corynebacterium diphtheriae*; *Brucella spp*. Essas pesquisas têm contribuído para o desenvolvimento de métodos profiláticos e diagnósticos mais rápidos e econômicos, avanços em estudos taxonômicos, compreensão de variações genéticas e elucidação de mecanismos de patogênese.

Book - Pan-genomics: Applications, Challenges, and Future Prospects:



Contents

Contributors

Talita Emile Ribeiro Adelino
Laboratório de Genética Celular e Molecular, ICB, Universidade Federal de Minas Gerais; Fundação Ezequeil Dias (Funed), Belo Horizonte, Brazil

Jamil Ahmad
Research Center for Modeling & Simulation (RCMS), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Shahbaz Ahmed
Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology, Islamabad, Pakistan

Luiz Carlos Junior Alcantara
PG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte; Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro; Laboratório de Genética Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Amjad Ali
Department of Plant Biotechnology, Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Rabia Amir
Department of Plant Biotechnology, Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Fabricio Araujo
Institute of Biological Sciences, Federal University of Pará (UFPA), Belém, Brazil

Muneeba Arveen
Department of Plant Biotechnology, Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Vasco Azevedo
PG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

Jahanzaib Azhar
Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology, Islamabad, Pakistan

Luciana Balbo
State University of Londrina, Londrina, Brazil

Li Bao
National Clinical Research center for Cancer, Tianjin Medical University Cancer Institute and Hospital; Key Laboratory of Cancer Prevention and Therapy, Tianjin, China

Debmalya Barh
Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Purba Medinipur, India

Fernanda Khouri Barreto
Laboratório de Patologia Experimental, Instituto Gonçalo Moniz, Fiocruz Bahia; Instituto Multidisciplinar em Saúde—IMS, Universidade Federal da Bahia (UFBA), Salvador, Brazil

Attya Bhatti
Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology, Islamabad, Pakistan

Andreas Burkovski
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Roberta Torres Chideroli
State University of Londrina, Londrina, Brazil

Mauricio Corredor
GEBIOMIC Group, FCEN, University of Antioquia, Medellin, Colombia

Kenny da Costa Pinheiro
Institute of Biological Sciences, Federal University of Pará (UFPA), Belém, Brazil

Artur Luiz da Costa Silva
Institute of Biological Sciences, Federal University of Pará (UFPA), Belém, Brazil

Hamza Arshad Dar
Department of Plant Biotechnology, Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Letícia de Castro Oliveira
Department of Immunology, Microbiology and Parasitology, Institute of Biological Science and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, Brazil

Siomar de Castro Soares
Department of Immunology, Microbiology and Parasitology, Institute of Biological Science and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, Brazil

Jaqueline Goes de Jesus
Laboratório de Flavivírus, IOC, Fundação Oswaldo Cruz, Rio de Janeiro; Laboratório de Patologia Experimental, Instituto Gonçalo Moniz, Fiocruz Bahia, Salvador, Brazil

Thiago de Jesus Sousa
PG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

Tulio de Oliveira
KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa

Stephane Fraga de Oliveira Tosta
PG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG); Laboratório de Genética Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Ulisses de Pádua Pereira
State University of Londrina, Londrina, Brazil

B978-0-12-817076-2.09988-1, 09988

5.6 Chapter 5 - Pan-genomics of veterinary bacteria and its applications.

CHAPTER 5

Pan-genomics of veterinary pathogens and its applications

Thiago de Jesus Sousa[a], Arun Kumar Jaiswal[a,b], Raquel Enma Hurtado[a], Stephane Fraga de Oliveira Tosta[a], Siomar de Castro Soares[b], Anne Cybelle Pinto Gomide[a], Luiz Carlos Junior Alcantara[d], Debmalya Barh[c], Vasco Azevedo[a], Sandeep Tiwari[a]

[a]PG Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil
[b]Department of Immunology, Microbiology and Parasitology, Institute of Biological Science and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, Brazil
[c]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Purba Medinipur, India
[d]Laboratório de Flavivírus, Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, Brazil

1 Introduction

Pan-genome is an approach that contributes to the research of bacterial pathogenesis. This terminology was proposed in 2005 in research with the bacterium *Streptococcus agalactiae*, by the researcher Tettelin and collaborators [1]. In this work, they define the pan-genome as a set of genes in a given study group, considering core genome, the genes present in all strains in the group of study; dispensable genes as absent genes in one or more strains; and, genes that are considered unique in each lineage of the study group. Pan-genome can be considered open or closed, depending on the bacterial ability to acquire exogenous regions (DNA) [1] and the lifestyle that will determine this issue [2]. From the sequencing, one can thoroughly study each region of the genome, contributing with unpublished information. Since 2005, with the era of new sequencers, the speed, ease, and reliability of data have been increasing and with them the number of bacterial genomes deposited in public databases [3]. Pan-genome studies can be applied with different goals, such as taxonomy, reverse vaccinology, gene variation, pathogenesis [4], among others. This chapter is foucused on the pan-genomics studies carried out on pathogenic bacteria that cause veterinary diseases, including the ones responsible for zoonotic diseases. From the genetic repertoire studies, the key points (genes) supposedly involved in the spread of disease, bacterial resistance, infection, adhesion, can be detected, leading to practical solutions against the disease being studied. An important fact is an identification, from taxonomic studies among the lineages, of horizontal gene transfer, which in addition to contributing to evolutionary information, may be used to infer possibly emerging pathogens, once the previously harmless pathogen may become pathogenic. Horizontal gene transfer causes a considerable impact on genomic plasticity,

Pan-genomics: Applications, Challenges, and Future Prospects
https://doi.org/10.1016/B978-0-12-817076-2.00005-6

Table 1 Pathogenic bacterial species of veterinary and human importance

Bacterial species causing animal infection	Bacterial species causing animal and human infection
Corynebacterium pseudotuberculosis	*Brucella*
Corynebacterium ulcerans	*Corynebacterium diphtheriae*
Streptococcus suis	*Francisella tularensis*
Brachyspira hyodysenteriae	*Campylobacter*
Moraxella bovoculi	*Clostridium botulinum*
Mannheimia haemolytica	*Streptococcus agalactiae*
Pasteurella multocida	

bacterial evolution, and adaptation, and leads to an inquiry into species determination [5]. The strains of the same species can differ considerably in the gene repertoire, which confers a versatile adaptation to a wide range of environments [6]. From the Pan-genome studies, one can perform this thorough analysis between different genomes leading to an understanding of evolutionary strategies, acquisition of resistance, hereditary variation leading to its evolutionary adaptation and, in some situations, the results can lead to even a proposal of species redefinition [2, 3].

Table 1 shows the list of pathogenic bacterial species of veterinary and human importance that already have Pan-genome studies. The studies have a high impact in the diagnosis, prophylaxis, and verification of the genetic variation among the strains. Thus, these studies could provide effective solutions to fight against the diseases that cause significant damage to the agribusiness or a constant public health problem.

2 Pan-genomics studies of pathogenic bacteria causing veterinary and zoonotic diseases

2.1 *Corynebacterium pseudotuberculosis*

Corynebacterium pseudotuberculosis is the agent of Caseous lymphadenitis (CLA), but may also cause other chronic diseases such as ulcerative lymphangitis. *C. pseudotuberculosis* has as host small and large ruminants, causing significant economic losses, and there are already some cases in the literature of transmission in humans as well. One way to contribute to the health of these host is the study of the genomes of *C. pseudotuberculosis* strains, which brings good discussions about the evolutionary understanding of the species, adaptation, and interaction with the host. Thus, it is possible to elucidate inferences about genes or virulence factors, and consequently more efficient and cheaper vaccines, drugs, and diagnostics. An example of this is reverse vaccinology, which aims to identify targets for vaccines and/or drugs by computational means and thereby reducing in vivo and in vitro tests [7].

In 2011, Ruiz et al. compared two genomes of *C. pseudotuberculosis*, strains 1002 and C231, which were first complete genomes deposited in National Center for Biotechnology Information (NCBI). These two strains are very similar, with approximately 95% similarity from the amino acid sequences of the predicted protein pool. The two strains are also very similar concerning genomic composition, G+C content values, gene size, operon composition, and gene density. However, significant differences are observed for genome size, a number of pseudogenes and lineage-specific genes. As expected, the strains including *C. pseudotuberculosis* 1002 and C231 showed high conservation in the genus, with approximately 97% of their genes presenting conservation in the gene order [8].

In 2013, Soares et al. did an antigenic target prediction study with the *C. pseudotuberculosis* 258 strain genome for the prediction of biotechnology vaccines. Then, by reverse vaccinology, 49 possible proteins were identified as vaccine target candidates, where one target was present on a pathogenicity island [9]. In the same year, Soares et al. made a pan-genomic analysis with 15 genomes of *C. pseudotuberculosis*, characterizing this species with an open pan-genome, in which approximately 19 new protein coding sequences were to be added for each new genome. The core genome consists of 1504 sequences encoding proteins. More detailed analyses of the pan-genome revealed differences between *ovis* and *equi* biovar strains, where the biovar *ovis* showed a more clonal behavior than the biovar *equi* strains [10].

But in the last quarter of 2018, the number of complete genomes has increased to 72. This genome data was made with new sequencing platforms and methodologies. In addition, many other works that corrected errors present in the assembly of the deposited genomes were published, suggesting an update in these pan-genomic studies [11].

2.2 *Corynebacterium ulcerans*

Corynebacterium ulcerans has emerged as a relevant zoonotic pathogen. An increasing number of cases of *C. ulcerans* infection have been reported from many countries including Brazil. *C. ulcerans* has a wide range of animal hosts [12].

Pan-genomic studies of *C. ulcerans* showed that the main virulence factor in this species is the *tox* gene, mainly present in *Corynebacterium diphtheriae*. The *tox* gene is found in lysogenic corynephages, but also on a pathogenicity island. In some strains the function of *tox* gene found inactivated due to frameshift mutation. However, several other genes encoding virulence-associated proteins, such as phospholipase D (*Pld*), neuraminidase H (*NanH*), corynebacterial protease (*CP40*), venom serine protease (*Vsp1* and *Vsp2*), ribosomal-binding protein (*Rbp*, similar to Shiga-like toxin), and adhesive surface pili are present in different *C. ulcerans* strains [13].

Pan-genomic studies have identified the presence of multiple prophages that are an important source of genomic plasticity. Surface pili are responsible for adhesion and

invasion of host cells, which play an essential role in the virulence of pathogenic bacteria. A study with 19 strains of *C. ulcerans* published in 2018 by Subedi et al. identified 4120 genes, including 1405 core genes and 2715 accessory genes. Among the proteins of the core genome, there were 351 proteins with transmembrane domains, 3 with additional signal peptides, 2 cell wall-anchored proteins, and 82 secreted proteins, of which 46 were identified as putative lipoproteins. The accessory genome included 611 membrane-associated proteins, 65 with additional signal peptide features and 46 with an LPXTG motif. A total of 116 accessory proteins were secreted via sec-dependent secretory pathways. Membrane-associated and secreted proteins are essential for host-pathogen interactions and virulence [13]. Therefore, in addition to the variation in the virulence genes, the number of transmembrane, lipoprotein, and secreted proteins may be responsible for the variation in their virulence characteristics. Indeed, a variation in the ability to cause arthritis in a mice model by different *C. ulcerans* strains was previously reported. As mentioned earlier, prophages are the primary source of diversity among these strains.

2.3 *Streptococcus suis*

Streptococcus suis is a Gram-positive bacterium considered one of the essential bacterial pathogens in the swine industry in the world, mainly in China. In addition, *S. suis* is also an emerging zoonotic pathogen. It is classified into 33 serotypes, where serotypes 1, 2, 3, 7, 9, and 1/2 are the most prevalent in swine, and strains that cause human infections were also found among these serotypes. In 2018, there are 42 complete genomes deposited in the NCBI, with a single chromosome of approximately 2 Mb [14].

A study in 2011 by Zhang et al. [14], with 13 complete genomes found 2374 orthologous genes and 1211 unique genes, a core genome with 1343 genes, and the observed pan-genome shared by the 13 strains consisted of 3585 genes. In this pan-genomic analysis, they estimated that for each newly sequenced genome, 82 genes are added, characterizing that the species has an open pan-genome. This is consistent with an earlier study on the core and pan-genome of *Streptococcus*, which indicated that *S. suis* was the ancestor with the highest number of genetic gains and losses [14].

2.4 *Brachyspira hyodysenteriae*

Brachyspira spp. is found colonizing intestines of some species of mammals and birds, and shows different degrees of enteropatogenicity. *B. hyodysenteriae* is an important swine pathogen, which causes dysentery in these animals. It has three complete genomes deposited in NCBI, and its genome size consists of a ~3-Mb chromosome and a ~36-kb plasmid. This plasmid is conserved among several strains, but it is not found in any non-virulent isolated strain in the field, suggesting that it may be an essential virulence factor for the species [15].

Genomic studies between *Brachyspira pilosicoli*, *Brachyspira intermedia*, *Brachyspira hyodysenteriae*, and *Brachyspira murdochii*, suggest *B. pilosicoli* lost many transport-related proteins, which might reflect its adaptation to a more specialized ecological niche. The highest level of reductive evolution in *B. pilosicoli* suggests that it is a pathogen older than *B. hyodysenteriae*. The pathogenicity of the younger *B. hyodysenteriae* may be related to the acquisition of the 32 kb plasmid [15]. In general, recent studies suggest that *B. hyodysenteriae* and *B. pilosicoli* are more specialized pathogens and have less genetic material and diversity. These strains have undergone specialization process independently, which is suggested by the little genetic material that is shared only between them. In addition, studies suggest that there was a reductive evolution with *B. hyodysenteriae* and *B. pilosicoli* since they have the two smaller genomes. Reductive evolution may be involved in the loss of genes, especially transport proteins [15]

2.5 *Moraxella bovoculi*

Infectious bovine keratoconjunctivitis (IBK) affects cattle, causing pain, blindness in severe cases, and reduced weight gain in animals. In addition to concern about animal health and welfare, IBK's economic impact may be significant, with estimates exceeding US$ 150 million in direct and indirect economic losses. As microbiological characteristics, they are coccobacillus and Gram-negative. *Moraxella bovoculi* has been extensively associated with IBK in the absence of *Moraxella bovis* since its initial description in 2007 [16].

Genomic studies with this species are scarce. Studies in the literature have shown that the diversity of single nucleotide polymorphisms (SNPs) in *M. bovoculi* is high, with 81,284 SNPs identified in eight genomes (being seven complete genomes). Two distinct genotypes are represented, isolated from IBK (genotype 1) and the nasopharynx of cattle without clinical IBK signs (genotype 2). Only in genotype 1, it found repeats-in-toxin (RTX) putative pathogenesis factor and 10 putative antibiotic resistance genes carried within a genomic island (GI). Due to very high recombination, genotype 1 subtypes cannot be distinguished at the SNPs level, although these subtypes may vary in their virulence potential. Interspecific recombination with *M. bovis* indicates that, for at least two loci, these species share a common genetic set. Because of this, future work as the development of IBK vaccines may benefit from the identification and characterization of conserved outer membrane proteins shared by both *Moraxella* species [16].

2.6 *Pasteurella multocida*

Pasteurella multocida is a Gram-negative commensal and bacterial pathogen causing economically important diseases of veterinarian interest as hemorrhagic septicemia, fowl cholera, atrophic rhinitis, and pneumonia in a broad range of animal species, likewise it is a zoonotic agent to humans through bites infections [17]. A last pangenomic study

on 109 *P. multocida* isolates describes a pan-genome with 4256 repertoire genes, 1806 core genes (42.43%), 1841 dispensable genes (43.25%), and 609 strain-specific genes (14.3%) [18]. Similar results describe the accessory genome with 52.91% and dispensable genes with 33.47%, showing an open pangenome to species [19]. The dispensable genes content assigned to COG categories belong to carbohydrate transport and metabolism (9.54%), transcription (4.85%), replication, recombination and repair (3.08%), inorganic ion transport, and metabolism (4.6%) [19]. The presence of these highlighted functional categories could be associated with its environmental fitness [20, 21], whereas 46.35% and 49% of unique and dispensable genes are assigned to unknown function, revealing a large number of noncharacterized proteins involved in diversification process [19]. Association studies of the accessory genome would show the presence of specific genes in a specific disease [19, 22, 23] but not a predilection to a host [18]. Complementary comparative genomic analysis show the accessory genome belonged to prophages, ICE, GI and plasmids, as well as the presence of a unique large integrative conjugative element, ICEPmu1, containing 88 genes of which 12 genes encoding resistance to antibiotics [24]. Likewise, pathogenomics analysis among virulent avian *P. multocida* strains (P1059 and/or X73) against an avirulent strain Pm70 identified 336 genes of which 61 genes present unknown function [22]. Other studies corroborated the presence of a cluster of genes involved in the transport and modification of citrate, galactitol-specific phosphotransferases, transport and utilization of L-fucose shared by at least two fowl cholera strains X73, F216, P1059, and F218 [19, 22]. The presence of these cluster of genes related to metabolism and adhesion could provide the capacity of adaptation and virulence to avian host [22]. Also, the genomic comparison among Hemorragic Septicemia-associated strains and strains not associated with the disease show two unique intact prophages present on all HS strains [23]. Additionally, phylogenomic and comparative genomics analysis based on the accessory genome shows the clustering of some *P. multocida* strains by disease [19, 22, 23], which supports the SNPs phylogenetic clustering [19]. Population phylogenies based on core genes show a relationship with the predilection to a host and geographical association [25] or MLST distribution [19]. These studies showed a great diversity at the gene level; likewise, this reflects the associations of genetic groups that present determinate mobile genetic element that could be involved with the capacity to infect. All the studies so far allow us to show the importance of accessory genome in the genetic diversification process and evolutionary adaptation of *P. multocida* species [19, 25, 26].

2.7 *Mannheimia haemolytica*

Mannheimia haemolytica is a hemolytic, Gram-negative coccobacillus, commensal of the upper respiratory tract and nasopharynx, and causal agent of respiratory disease on ruminants, mainly associated with the bovine respiratory disease with economic losses to the

cattle industry worldwide [27, 28]. Pan-genome analysis of 21 *M. haemolytica* isolates identified 9507 orthologous groups of genes, 1333 core genes (14%), and 6350 dispensable genes (66.8%) [29]. The pan-genome of all 21 *M. haemolytica* strains is open and the accessory genome is composed of 66.8% and 81.8% of dispensable and unique genes, respectively, containing uncharacterized or hypothetical proteins [29]. The virulence and etiology of *M. haemolytica* is strongly associated with serotypes, being serotype 1 and 6 responsible for pneumonia in bovine and serotype 2 responsible for pneumonia in sheep and prevalent as commensal among healthy cattle [29, 30]. Comparative pathogenomic studies found differences between S1, S6 bovine strains with the presence of more integrative conjugative elements and prophages than S2 strain and also differences of spacer sequences on CRISPR arrays. Likewise, the presence of antimicrobial-resistant (AMR) contained in conjugable element (ICE) is more prevalent in S2 than S1 and S6 strains. The AMR may be removed in SA and S6 through effective antimicrobial therapies in diseases animal compared with healthy animals. However, little is known about how genetic differences among serotypes contribute to pathogenesis in this species [29, 30]. The identification of variable mobile genetic elements as prophages and ICEs would be implied in the genetic diversification process, pathogenicity, and evolutionary adaptation [29–31]. First comparative genomic analysis between three strains of *M. haemolytica* from bovines and ovines found a high percentage of hypothetical proteins in the content of unique genes (57%) and phage related genes (20% and 29% from A1 and B strain, respectively), where the authors correlated the variable gene pool with specific phenotypes (strain virulence, species specificity, etc.) [30]. From the analysis of 11 bovine isolates, 14 prophage clusters were identified, which contain toxin-antitoxin systems and multiple virulence-associated genes involved in virulence and antimicrobial resistance [29]. It was detected a CRISPR-Cas that play a role in immune evasion or adhesion during infection [29]. Integrative conjugative elements were found in nine strains, playing a role in the survival through the multidrug resistances [32, 33], and regulating their dissemination through toxin-antitoxin and entry exclusion systems [29]. Comparative genomic analyses of pathogenic strains would allow a better comprehension of the pathogenicity and the prediction of resistance mechanisms. Likewise, pan-genome analysis allows the discovery of all spectrum of genes represented, which are implicated in the genetic diversity and evolution of the species (Table 2).

2.8 *Clostridium botulinum*

Clostridium botulinum is an anaerobic, Gram-positive, and spore-forming pathogen in charge of the rising of food contamination cases over the world. The transmission of the disease from *C. botulinum* is resonating, by the unexpected hospital outbreaks and expanded obstruction against multiple drugs [38]. *C. botulinum* is able to produce botulinum toxins and these toxins (BoNT) are considered to be the most toxic substances

Table 2 An overview of Pan-genome studies in veterinary infection related bacteria

Name of the bacteria	Disease	Host	Genome size (Mb)	Pan genome analysis	References
Brucella spp.	Brucellosis	Human, Bovine and small ruminants	3.3	To get insights of the survival mechanism	[34–36]
Brachyspira hyodysenteriae	Swine dysentery	Mammals and birds	3.052	Reduction of many transport-related proteins	[15]
Corynebacterium ulcerans	Diphtheria-like infection and extrapharyngeal infections	Animal/ Human	2.497	In core genome, 351 were transmembrane domains, 3 with additional signal peptides, and 2 were cell wall-anchored proteins, 82 were predicted to be secreted, of which 46 were identified as putative lipoproteins.	[13]
Corynebacterium diphtheriaediphtheria	Diphtheria	Humans and animal	2.444	57 genomics islands, most of them pathogenicity islands and associated with adhesive pili, responsible for the adhesion	[37]
Corynebacterium pseudotuberculosis	Caseous lymphadenitis/ ulcerative lymphangitis	Animal	2.337	Revealed differences between ovis and equi biovar strains	[10]
Clostridium botulinum	Botulism	Human and animal	3.917	Open pangenome, the study was to study symptoms related to this	[38]

				bacteria with respect to the wide range of hosts	
Campylobacter spp.	Campylobacteriosis	Human and animal	1.818		
Francisella tularensis	Tularaemia	Lagomorphs and humans	1.825	The presence of point mutations, insertion elements and small indels resulting in gene deactivation in the process of differentiation from the nonpathogenic strain into the human pathogenic strains	[39]
Moraxella bovoculi	Infectious Bovine Keratoconjunctivitis (IBK)	Cattle	2.214	81,284 SNPs identified in eight genomes	[16]
Mannheimia haemolytica	Respiratory disease	Cattle	2.635	Open and the accessory genome is composed of 66.8% and 81.8% of dispensable and unique genes, respectively	[29]
Pasteurella multocida	Hemorrhagic septicemia, fowl cholera, atrophic rhinitis and pneumonia	Animals	2.305	The importance of accessory genome in the genetic diversification process and evolutionary adaptation	[19, 25, 26]
Streptococcus agalactiae	Meningoencephalitis, Septicemia, Meningitis, Neonatal sepsis and pneumonia	Cattle, Fish and Human	2.081	Vaccine targets identification, 36 antigenic proteins as possible vaccine targets	[40]
Streptococcus suis	Meningitis, septicaemia	Swine and Human	2.096	Each newly sequenced genome, 82 genes were added, Open pan-genome	[14]

occurring in nature [41]. Botulism is a perilous flaccid paralytic disease caused by eight different neuroparalytic toxin subtypes (A–H) [42]. Toxin subtypes A, B, E, and F are rarely and recently discovered, and serotype H is mainly responsible for human botulism, whereas toxin types C and D are involved in animal botulism around the world [42, 43]. The instances of Botulism infection are exceptionally normal in wild and local creatures and happen sporadically just as hugely everywhere throughout the world. The cattle and birds are extremely affected species of animals, despite the fact that botulism cases likewise are typically found among horses, sheep, and goats. The bacteria produce botulinum neurotoxins that act on the nerve endings, blocking acetylcholine discharge [44–46]. *C. botulinum* is the third most infectious agent worldwide to human and animal health. Botulism cases are exceptionally critical in ruminants, common in birds and dogs, and have additionally been reported in other species, specifically dogs, pigs, horses, and wild mammals in Brazil [47]. The first Botulism disease was reported in Brazil in 1960s in the state of Piauí in cattle, and was later identified in other species, such as sheep, goats, and buffaloes in all Brazilian regions [47]. The strain A2 of *C. botulinum* was recognized as resistant to metronidazole and penicillin [48]. A pan-genome work was published by Bhardwaj et al. [38], to comprehend the symptoms related to this bacteria with respect to the wide range of hosts. The successive calculation and characterization of the core and pan-genome subset disclosed the identification of more specific targets for drug designing and vaccine development [38]. In this study, 13 genomes of *C. botulinum* were used for pan-genome analysis and they identified 889 genes as core genome and 287 strain-specific genes. The reported open pan-genome in their analysis, which indicates unique genes, suggests that new genes could be added with every newly added genome sequence. Core, unique, and accessory genes were further categorized, in which most of core genes belong to metabolism and genetic information processing. Core-genome calculation exposes high level of genomic similarity among the genomes with low variation in GC content. The persistence of singleton genes shows the capacity to get novel virulence traits. The identification and analysis of GIs helped characterize potential drugs and vaccine targets [38].

2.9 *Campylobacter*

The *Campylobacter* species constitute a highly biological diverse group of organisms, some of which are widely known causative agents of clinical illness in animals and humans [49]. The disease Campylobacteriosis is an aggregate depiction for infectious diseases, caused by members of the bacterial genus *Campylobacter*. The infection is present in animals such as poultry, cattle, pigs, wild birds, and wild mammals. *Campylobacter* bacterium is one of the greatest agents of foodborne diarrheal illness in humans, and in addition, commonly causes gastroenteritis worldwide [50–52] and affects 9 million people each year, costing around €2.4 billion [53, 54]. Generally, infections are not extreme, being the most critical

symptom the gastroenteritis; however, they can also cause extraintestinal manifestations such as reactive arthritis, inflammatory bowel disease (IBD), Guillain-Barré syndrome (GBS), and in some cases, infection lead to death. Infections in Human are fundamentally connected with taking care of and additionally devouring poultry meat [54, 55]. The related subspecies *C. fetus* subsp. fetus and *C. fetus* subsp. *venerealis* of *Campylobacter fetus* are well-known pathogens of reproductive failures in ruminants [56]. The *C. fetus* subsp. *fetus* shows a wide ranging of host distribution, colonizes the gastrointestinal tract, and is generally linked with sheep and cattle abortion, while *C. fetus* subsp. *venerealis* has low host range, is restricted to the bovine genital tract, and the primary cause of venereally transmitted infectious, infertility, and embryonic mortality in cattle [49, 57]. In addition to *C. fetus* subsp. *fetus*, *Campylobacter jejuni* subsp. *jejuni* is also a major pathogen of *Campylobacter* species related with sheep abortion [49, 57, 58]. *C. fetus* subsp. *venerealis*, infections is also known as bovine genital campylobacteriosis (BGC), bovine venereal campylobacteriosis, or vibriosis, which is characterized by infertility and early embryonic deaths [57, 59, 60]. Rather than its public health importance, the ecological and evolutionary aspects of the *Campylobacter* are still poorly understood. Nevertheless, they could have an intense effect on transmission and human infection and it is not explained properly how *Campylobacter coli* and *C. jejuni*, which have similar host niches and frequently exchange genetic material, show differences in their disease epidemiology [61]. Throughout the decades, antibiotics have been arbitrarily used in animal production to control, prevent, and treat infections and to increase animal growth [62]. The primary cause of rise and spread of antibiotic resistance among *Campylobacter* spp. is the use of unregulated antimicrobial agents in food animal production, which has led to the development of antibiotic resistance in campylobacter subspecies [63–65]. *Campylobacter* antibiotics resistance is emerging globally and has already been described by several authors earlier and also acknowledged by the WHO, as a problem of public health importance [63, 65–68]. Antibiotics, generally tetracycline, macrolides, and (fluoro) quinolones, are used for more severe cases. Nevertheless, the growth of resistance to tetracycline, erythromycin, and (fluoro) quinolones of *C. coli* and *C. jejuni* strains might compromise the efficacy of this treatment [65]. Work published by Lefébure et al., in 2010, used 42 strains of *C. coli* and 43 strains of *C. jejuni*, where the pan-genome of both species combined reaches approximately 3000 genes [69]. In another study published in 2014 by Méric et al., seven strains of *C. jejuni* and *C. coli* genomes were used for pan-genome analysis. They identified 3933 genes as pan-genome, a core genome of 1035, and the accessory genome contained 2792 genes [61].

2.10 *Streptococcus agalactiae*

S. agalactiae is a bacterium that causes illnesses in cattle, fish, and human [40]. In human, it is frequently associated with meningitis, neonatal sepsis, pneumonia, and pregnant

women [40, 70]. This bacterium is associated with typical gut flora and genital tract, moreover, it is also found colonizing 10%–40% of pregnant women [71, 72]. A notable number of newborn infant infections from *S. agalactiae* have been identified, making it necessary to investigate it in view of its substantial morbidity and mortality [73, 74]. In dairy cattle, *S. agalactiae* (Lancefield group B; GBS) is additionally a noteworthy pathogen of clinical and subclinical mastitis, which affects quality and production of milk [70]. *S. agalactiae* is an evolving pathogen in fish, which causes meningoencephalitis and septicemia. The pathogen has been accounted with high mortality in wild and cultured species worldwide [40, 75, 76]. *S. agalactiae* developed phenotypic and genotypic antibiotic resistance patterns in China, being isolated from cows with mastitis [77]. Bolukaoto et al. [71] isolated an antibiotic resistant strain of *S. agalactiae* from pregnant women in Garankuwa, South Africa. *In silico* techniques like Pan-genome, Pan-modelome, Subtractive genomics, and Reverse vaccinology are playing a key role in quick and rapid identification of new therapeutic targets in the post-genomic era [78]. In 2013 Pereira et al., published research article for vaccine targets against *S. agalactiae* where they used 15 genomic strains from different isolates (10 from human isolates, 4 from fish and 1 from cow). Their pan-genome analysis identified 5143 genes on the pan-genome and 1111 genes as part of the core-genome, shared by all genomes. They identified 36 antigenic proteins as possible vaccine targets, which were conserved in all 15 strains and, in future, will be used as vaccine candidates [40].

2.11 *Francisella tularensis*

F. tularensis is a highly infectious, Gram-negative, facultative, and intracellular bacterium, which presents rod-shaped or coccoid cells and is also aerobic and nonmotile [79]. *F. tularensis* is the etiological agent of tularaemia—a zoonotic disease that has been described in animals, predominantly in rodents, lagomorphs, and humans [80]. In this group, six clinical manifestations are characterized by the form of entrance of bacteria: ulceroglandular, glandular, oropharyngeal, oculoglandular, pneumonic tularaemia, and typhoidal tularaemia forms [81]. The occurrence of tularemia is equally influenced by the host and the different subspecies [80] as the four proposed subspecies of *F. tularensis* subspecies *tularensis*, *holarctica*, *novicida*, and *mediasiatica* differ in virulence and geographical range.

Rohmer and collaborators (2007) compared two pathogenic subspecies in humans; *F. tularensis* subspecies *tularensis* and *holarctica* against *F. tularensis* subspecies *novicida* U112, described as nonpathogenic in humans but reproducing in mice a tularaemia-like disease [82]. The comparison revealed the presence of point mutations, insertion elements, and small indels resulting in gene deactivation in the process of differentiation from the nonpathogenic strain into the human pathogenic strains [39].

In order to investigate adaptations within the genus *Francisella*, in 2009, Larsson and collaborators compared 13 *F. tularensis* isolates from different subspecies to the genomes

of 3 isolates of *Francisella novicida* and 1 isolate of *Francisella philomiragia*. Although *F. novicida* and *F. tularensis* present an average nucleotide identity of >97%, *F. novicida* is less virulent in mammals, with rare descriptions of human infections and seems to have a less specialized cycle. This increased host association were related to events of random insertions like the duplication of the *Francisella* Pathogenicity Island [83].

2.12 *Corynebacterium diphtheriae*

C. diphtheriae is the etiological agent of diphtheria, an acute disease localized in the upper respiratory tract leading to ulceras at the mucosa, and formation of an inflammatory pseudomembrane [84]. *C. diphtheriae* strains can be divided into toxigenic strains, which carry the *tox* structural gene and nontoxigenic strains, which do not carry the tox gene. Nontoxigenic *C. diphtheriae* strains are related with severe pharyngitis and tonsillitis, endocarditis, osteomyelitis, splenic abscesses, and septic arthritis [85]. In order to explore the genetic basis of different interactions with host tissues and clinical manifestations of infection by a variety of *C. diphtheriae* strains, several studies have been developed to provide information if a group of genes can be related with a clinical manifestation of *C. diphtheriae* infection.

Trost and collaborators performed a pan-genome study of *C. diphtheriae* comparing 13 genomes of strains isolated from patients with classical diphtheria, pneumonia, endocarditis, and the strain *C. diphtheriae* NCTC 13129 as a reference. It was demonstrated a high synteny level and a core genome consisting of 1632 conserved genes and, on average, 65 unique genes per strain. The analysis of genome-wide motif searches of *tox*-controlling regulator DtxR showed that the DtxR regulons presented differences due to gene variation on those sites responsible for interactions with DtxR. One important finding was the identification of 57 genomics islands, most of them are pathogenicity islands and associated with adhesive pili, responsible for the adhesion of *C. diphtheriae* to different host tissues [37]. Other study performed with 48 *C. diphtheriae* isolates from Australia over a 12-year period. The pan-genome analysis revealed 22 genes from gene group I significantly associated with respiratory infection [86].

Although the detection and isolation of *C. diphtheriae* in animals is poorly described at the literature, it is extremely relevant to try to understand the role of animals at transmission of *C. diphtheriae*, once the majority of isolated strains from animals had direct contact with humans [87].

C. diphtheriae had been characterized from four different animal species (dog, cat, cow, and horse) showing nontoxigenic, toxigenic, and nontoxigenic tox-bearing (NTTB) *C. diphtheriae* strains. All these reports had different clinical manifestations from pharyngitis, parotitis, otitis, chronic active dermatitis, draining wound infection to nonhealing pyogenic stake wound [88–93], which may contribute to the poor investigations of injuries as a result of *C. diphtheriae* infection in other animal species.

As described by Sing et al. [87], in the first finding of a nontoxigenic *C. diphtheriae* biovar *belfanti* in a wild red fox with no human contact, *C. diphtheriae* was accompanied by *Streptococcus canis*, an opportunistic pathogen of this species. Even though the contributions of lesions cannot be attributed just to *C. diphtheriae* [87], this case brings forward the possibility of *C. diphtheriae* infections being not detected as pathogenic bacteria in humans and animal infections [87].

2.13 *Brucella* spp.

The genus *Brucella* is composed by seven species, being them *Brucella neotomae*, *Brucella melitensis*, *Brucella abortus*, *Brucella suis*, *Brucella ovis*, *Brucella canis*, and *Brucella maris*. They are Gram-negative facultative intracellular and coccobacilli nonmotile bacteria [34]. Brucellosis is a zoonotic disease caused by *Brucella* spp. affecting mainly mammals, such as cattle, goats, camels, sheeps, pigs, dogs, which can lead to sterility or abortion, and humans, which causes serious, debilitating illness [34].

More than one species as the etiological agent of brucellosis has fomented several pan-genomic studies in order to identify different contributions of each agent.

Yang and collaborators performed pan-genomics analysis with 42 *Brucella* complete genomes in order to get insights on the survival mechanism of *Brucella* spp. in vivo. From the genes analyzed, the core genome contains 1710 clusters, 1182 clusters were strain-specific genes, and 2477 clusters were accessory genome. The core functions were mainly related with conservation, amino acid metabolism, and energy [36].

Although many studies look for genomic characteristics that can be distinguishable as a host adaptation, a comparative genomics study identified clonal isolates of *B. melitensis* Biovar 3 with no signature of host adaptation, investigating strains of a same outbreak from three different species (human, bovine, small ruminants) [35].

3 Conclusions

The infectious diseases that can be naturally transmitted between animals and humans are known as zoonoses. The causative agent of zoonoses includes wide range of pathogens such as viruses, bacteria, fungi, and parasites. Due to the advancement of the sequencing technology, there are multiple genome data of these pathogens available. Using bioinformatics and comparative genomics approaches can help in better understanding the dynamics of the pathogenies. Such as in the identification of common virulence factors in pathogenicity islands, which have a direct impact in the shared and singletons genes. Also, they may help in finding new vaccine and drug targets through the use of core genome information. Other omics analyses may also be performed like, pan-transcriptomics and pan-proteomics to discover the different patterns of gene expression of these organisms in different hosts, shedding a light on their adaptability. Finally, pan-genomics may contribute in the search for efficient new solutions against these diseases that cause several animal and human losses worldwide in agriculture and health systems.

References

[1] H. Tettelin, V. Masignani, M.J. Cieslewicz, C. Donati, D. Medini, N.L. Ward, et al., Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome", Proc. Natl. Acad. Sci. U. S. A. 102 (39) (2005) 13950–13955.
[2] L. Rouli, V. Merhej, P.E. Fournier, D. Raoult, The bacterial pangenome as a new tool for analysing pathogenic bacteria, New Microbes New Infect. 7 (2015) 72–85.
[3] X. Zhang, X. Liu, F. Yang, L. Chen, Pan-genome analysis links the hereditary variation of Leptospirillum ferriphilum with its evolutionary adaptation, Front. Microbiol. 9 (2018) 577.
[4] P.-G.C. Computational, Computational pan-genomics: status, promises and challenges, Brief. Bioinform. 19 (1) (2018) 118–135.
[5] V. Daubin, G.J. Szollosi, Horizontal gene transfer and the history of life, Cold Spring Harb. Perspect. Biol. 8 (4) (2016).
[6] A. Mira, A.B. Martin-Cuadrado, G. D'Auria, F. Rodriguez-Valera, The bacterial pan-genome:a new paradigm in microbiology, Int. Microbiol. 13 (2) (2010) 45–57.
[7] L.C. Guimaraes, J. Florczak-Wyspianska, L.B. de Jesus, M.V. Viana, A. Silva, R.T. Ramos, et al., Inside the pan-genome—methods and software overview, Curr. Genomics 16 (4) (2015) 245–252.
[8] J.C. Ruiz, V. D'Afonseca, A. Silva, A. Ali, A.C. Pinto, A.R. Santos, et al., Evidence for reductive genome evolution and lateral acquisition of virulence functions in two Corynebacterium pseudotuberculosis strains, PLoS One 6 (4) (2011).
[9] S.C. Soares, E. Trost, R.T. Ramos, A.R. Carneiro, A.R. Santos, A.C. Pinto, et al., Genome sequence of Corynebacterium pseudotuberculosis biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production, J. Biotechnol. 167 (2) (2013) 135–141.
[10] S.C. Soares, A. Silva, E. Trost, J. Blom, R. Ramos, A. Carneiro, et al., The pan-genome of the animal pathogen Corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains, PLoS One 8 (1) (2013).
[11] D.C. Mariano, J. Sousa Tde, F.L. Pereira, F. Aburjaile, D. Barh, F. Rocha, et al., Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002, BMC Genomics 17 (2016) 315.
[12] W.B. Whitman, F. Rainey, P. Kämpfer, M. Trujillo, J. Chun, P. DeVos, et al., Bergey's Manual of Systematics of Archaea and Bacteria, 2015.
[13] R. Subedi, V. Kolodkina, I.C. Sutcliffe, L. Simpson-Louredo, R. Hirata Jr., L. Titov, et al., Genomic analyses reveal two distinct lineages of Corynebacterium ulcerans strains, New Microbes New Infect. 25 (2018) 7–13.
[14] A. Zhang, M. Yang, P. Hu, J. Wu, B. Chen, Y. Hua, et al., Comparative genomic analysis of Streptococcus suis reveals significant genomic diversity among different serotypes, BMC Genomics 12 (2011) 523.
[15] T. Hafstrom, D.S. Jansson, B. Segerman, Complete genome sequence of Brachyspira intermedia reveals unique genomic features in Brachyspira species and phage-mediated horizontal gene transfer, BMC Genomics 12 (2011) 395.
[16] A.M. Dickey, G. Schuller, J.D. Loy, M.L. Clawson, Whole genome sequencing of Moraxella bovoculi reveals high genetic diversity and evidence for interspecies recombination at multiple loci, PLoS One 13 (12) (2018).
[17] B.A. Wilson, M. Ho, Pasteurella multocida: from zoonosis to cellular microbiology, Clin. Microbiol. Rev. 26 (3) (2013) 631–655.
[18] Z. Peng, W. Liang, F. Wang, Z. Xu, Z. Xie, Z. Lian, et al., Genetic and phylogenetic characteristics of Pasteurella multocida isolates from different host species, Front. Microbiol. 9 (2018) 1408.
[19] R. Hurtado, D. Carhuaricra, S. Soares, M.V.C. Viana, V. Azevedo, L. Maturrano, et al., Pan-genomic approach shows insight of genetic divergence and pathogenic-adaptation of Pasteurella multocida, Gene 670 (2018) 193–206.
[20] A.N. Brooks, S. Turkarslan, K.D. Beer, F.Y. Lo, N.S. Baliga, Adaptation of cells to new environments, Wiley Interdiscip. Rev. Syst. Biol. Med. 3 (5) (2011) 544–561.
[21] C. Simon, A. Wiezer, A.W. Strittmatter, R. Daniel, Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome, Appl. Environ. Microbiol. 75 (23) (2009) 7519–7526.

[22] T.J. Johnson, J.E. Abrahante, S.S. Hunter, M. Hauglund, F.M. Tatum, S.K. Maheswaran, et al., Comparative genome analysis of an avirulent and two virulent strains of avian Pasteurella multocida reveals candidate genes involved in fitness and pathogenicity, BMC Microbiol. 13 (2013) 106.
[23] A.M. Moustafa, T. Seemann, S. Gladman, B. Adler, M. Harper, J.D. Boyce, et al., Comparative genomic analysis of asian haemorrhagic septicaemia-associated strains of Pasteurella multocida identifies more than 90 haemorrhagic septicaemia-specific genes, PLoS One 10 (7) (2015).
[24] G.B. Michael, K. Kadlec, M.T. Sweeney, E. Brzuszkiewicz, H. Liesegang, R. Daniel, et al., ICEPmu1, an integrative conjugative element (ICE) of Pasteurella multocida: structure and transfer, J. Antimicrob. Chemother. 67 (1) (2012) 91–100.
[25] D. Zhu, J. He, Z. Yang, M. Wang, R. Jia, S. Chen, et al., Comparative analysis reveals the Genomic Islands in Pasteurella multocida population genetics: on symbiosis and adaptability, BMC Genomics 20 (1) (2019).
[26] J.D. Boyce, T. Seemann, B. Adler, M. Harper, Pathogenomics of Pasteurella multocida, Curr. Top. Microbiol. Immunol. 361 (2012) 23–38.
[27] G.H. Frank, Pasteurellosis of cattle, in: C. Adlam, J.M. Rutter (Eds.), Pasteurella and Pasteurellosis, Academic Press, New York, 1989, pp. 197–221.
[28] M.R. Ackermann, K.A. Brogden, Response of the ruminant respiratory tract to Mannheimia (Pasteurella) haemolytica, Microbes Infect. 2 (9) (2000) 1079–1088.
[29] C.L. Klima, S.R. Cook, R. Zaheer, C. Laing, V.P. Gannon, Y. Xu, et al., Comparative genomic analysis of Mannheimia haemolytica from bovine sources, PLoS One 11 (2) (2016).
[30] P.K. Lawrence, W. Kittichotirat, J.E. McDermott, R.E. Bumgarner, A three-way comparative genomic analysis of Mannheimia haemolytica isolates, BMC Genomics 11 (2010) 535.
[31] E.C. Keen, Paradigms of pathogenesis: targeting the mobile genetic elements of disease, Front. Cell. Infect. Microbiol. 2 (2012) 161.
[32] C. Eidam, A. Poehlein, A. Leimbach, G.B. Michael, K. Kadlec, H. Liesegang, et al., Analysis and comparative genomics of ICEMh1, a novel integrative and conjugative element (ICE) of Mannheimia haemolytica, J. Antimicrob. Chemother. 70 (1) (2015) 93–97.
[33] M.L. Clawson, R.W. Murray, M.T. Sweeney, M.D. Apley, K.D. DeDonder, S.F. Capik, et al., Genomic signatures of Mannheimia haemolytica that associate with the lungs of cattle with respiratory disease, an integrative conjugative element, and antibiotic resistance genes, BMC Genomics 17 (1) (2016) 982.
[34] K.L. Cosford, Brucella canis: an update on research and clinical management, Can. Vet. J. 59 (1) (2018) 74–81.
[35] M. Holzapfel, G. Girault, A. Keriel, C. Ponsart, D. O'Callaghan, V. Mick, Comparative genomics and in vitro infection of field clonal isolates of Brucella melitensis biovar 3 did not identify signature of host adaptation, Front. Microbiol. 9 (2018) 2505.
[36] X. Yang, Y. Li, J. Zang, Y. Li, P. Bie, Y. Lu, et al., Analysis of pan-genome to identify the core genes and essential genes of Brucella spp, Mol. Gen. Genomics. 291 (2) (2016) 905–912.
[37] E. Trost, J. Blom, S. de Castro Soares, I.H. Huang, A. Al-Dilaimi, J. Schroder, et al., Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia, J. Bacteriol. 194 (12) (2012) 3199–3215.
[38] T. Bhardwaj, P. Somvanshi, Pan-genome analysis of Clostridium botulinum reveals unique targets for drug development, Gene 623 (2017) 48–62.
[39] L. Rohmer, C. Fong, S. Abmayr, M. Wasnick, T. Larson Freeman, M. Radey, et al., Comparison of Francisella tularensis genomes reveals evolutionary events associated with the emergence of human pathogenic strains, Genome Biol. 8 (6) (2007).
[40] U.P. Pereira, S.C. Soares, J. Blom, C.A.G. Leal, R.T.J. Ramos, L.C. Guimarães, et al., In silico prediction of conserved vaccine targets in Streptococcus agalactiae strains isolated from fish, cattle, and human samples, Genet. Mol. Res. 12 (3) (2013) 2902–2912.
[41] C. Rasetti-Escargueil, E. Lemichez, M. Popoff, Variability of botulinum toxins: challenges and opportunities for the future, Toxins 10 (9) (2018).
[42] M.R. Popoff, Ecology of neurotoxigenic strains of clostridia, Curr. Top. Microbiol. Immunol. 195 (1995) 1–29.

[43] J.R. Barash, S.S. Arnon, A novel strain of Clostridium botulinum that produces type B and type H botulinum toxins, J. Infect. Dis. 209 (2) (2014) 183–191.
[44] M. Kruger, M. Skau, A.A. Shehata, W. Schrodl, Efficacy of Clostridium botulinum types C and D toxoid vaccination in Danish cows, Anaerobe 23 (2013) 97–101.
[45] K. Oguma, T. Yamaguchi, K. Sudou, N. Yokosawa, Y. Fujikawa, Biochemical classification of Clostridium botulinum type C and D strains and their nontoxigenic derivatives, Appl. Environ. Microbiol. 51 (2) (1986) 256–260.
[46] E.L. Ortolani, L.A. Brito, C.S. Mori, U. Schalch, J. Pacheco, L. Baldacci, Botulism outbreak associated with poultry litter consumption in three Brazilian cattle herds, Vet. Hum. Toxicol. 39 (2) (1997) 89–92.
[47] R.O.S. Silva, C. Oliveira, L.A. Gonçalves, F.C.F. Lobato, Botulism in ruminants in Brazil, Ciência Rural 46 (8) (2016).
[48] C. Mazuet, E.J. Yoon, S. Boyer, S. Pignier, T. Blanc, I. Doehring, et al., A penicillin- and metronidazole-resistant Clostridium botulinum strain responsible for an infant botulism case, Clin. Microbiol. Infect. 22 (7) (2016). 644.e7-e12.
[49] O. Sahin, M. Yaeger, Z. Wu, Q. Zhang, Campylobacter-associated diseases in animals, Annu Rev Anim Biosci. 5 (2017) 21–42.
[50] R. Jain, S. Singh, V. SK, A. Jain, Genome-wide prediction of potential vaccine candidates for Campylobacter jejuni using reverse vaccinology, Interdiscip. Sci. 11 (2019) 337–347.
[51] A.H.M. van Vliet, J.M. Ketley, Pathogenesis of enteric Campylobacter infection, J. Appl. Microbiol. 90 (S6) (2001) 45S–56S.
[52] J.I. Dasti, A.M. Tareen, R. Lugert, A.E. Zautner, U. Groß, Campylobacter jejuni: a brief overview on pathogenicity-associated factors and disease-mediating mechanisms, Int. J. Med. Microbiol. 300 (4) (2010) 205–211.
[53] I.A. Gillespie, S.J. O'Brien, J.A. Frost, G.K. Adak, P. Horby, A.V. Swan, et al., A case-case comparison of Campylobacter coli and Campylobacter jejuni infection: a tool for generating hypotheses, Emerg. Infect. Dis. 8 (9) (2002) 937–942.
[54] M. Meunier, M. Guyard-Nicodème, E. Hirchaud, A. Parra, M. Chemaly, D. Dory, Identification of novel vaccine candidates against campylobacterthrough reverse vaccinology, J Immunol Res 2016 (2016) 1–9.
[55] R. Janssen, K.A. Krogfelt, S.A. Cawthraw, W. van Pelt, J.A. Wagenaar, R.J. Owen, Host-pathogen interactions in Campylobacter infections: the host perspective, Clin. Microbiol. Rev. 21 (3) (2008) 505–518.
[56] M.A. van Bergen, K.E. Dingle, M.C. Maiden, D.G. Newell, L. van der Graaf-Van Bloois, J.P. van Putten, et al., Clonal nature of Campylobacter fetus as defined by multilocus sequence typing, J. Clin. Microbiol. 43 (12) (2005) 5888–5898.
[57] M.B. Skirrow, Diseases due to Campylobacter, Helicobacter and related bacteria, J. Comp. Pathol. 111 (2) (1994) 113–149.
[58] O. Sahin, C. Fitzgerald, S. Stroika, S. Zhao, R.J. Sippy, P. Kwan, et al., Molecular evidence for zoonotic transmission of an emergent, highly pathogenic Campylobacter jejuni clone in the United States, J. Clin. Microbiol. 50 (3) (2012) 680–687.
[59] S. Hum, Bovine abortion due to Campylobacter fetus, Aust. Vet. J. 64 (10) (1987) 319–320.
[60] C.A. Kirkbride, Etiologic agents detected in a 10-year study of bovine abortions and stillbirths, J. Vet. Diagn. Investig. 4 (2) (1992) 175–180.
[61] G. Méric, K. Yahara, L. Mageiros, B. Pascoe, M.C.J. Maiden, K.A. Jolley, et al., A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic Campylobacter, PLoS One 9 (3) (2014).
[62] E. Rozynek, K. Dzierzanowska-Fangrat, B. Szczepanska, S. Wardak, J. Szych, P. Konieczny, et al., Trends in antimicrobial susceptibility of Campylobacter isolates in Poland (2000-2007), Pol. J. Microbiol. 58 (2) (2009) 111–115.
[63] J. Takkinen, A. Ammon, O. Robstad, T. Breuer, Campylobacter Working Group, European survey on Campylobacter surveillance and diagnosis 2001, Euro Surveill. 8 (11) (2003) 207–213.
[64] J.L. Smith, P.M. Fratamico, Fluoroquinolone resistance in campylobacter, J. Food Prot. 73 (6) (2010) 1141–1152.

[65] J. Silva, D. Leite, M. Fernandes, C. Mena, P.A. Gibbs, P. Teixeira, Campylobacter spp. as a foodborne pathogen: a review, Front. Microbiol. 2 (2011) 200.
[66] J.R. Greig, Quinolone resistance in Campylobacter, J. Antimicrob. Chemother. 51 (3) (2003) 740–742.
[67] P.F. McDermott, S.M. Bodeis-Jones, T.R. Fritsche, R.N. Jones, R.D. Walker, Broth microdilution susceptibility testing of Campylobacter jejuni and the determination of quality control ranges for fourteen antimicrobial agents, J. Clin. Microbiol. 43 (12) (2005) 6136–6138.
[68] J.E. Moore, M.D. Barton, I.S. Blair, D. Corcoran, J.S. Dooley, S. Fanning, et al., The epidemiology of antibiotic resistance in Campylobacter, Microbes Infect. 8 (7) (2006) 1955–1966.
[69] T. Lefébure, P.D. Pavinski Bitar, H. Suzuki, M.J. Stanhope, Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept, Genome Biol. Evol. 2 (2010) 646–655.
[70] V.P. Richards, P. Lang, P.D. Bitar, T. Lefebure, Y.H. Schukken, R.N. Zadoks, et al., Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted Streptococcus agalactiae, Infect. Genet. Evol. 11 (6) (2011) 1263–1275.
[71] J.Y. Bolukaoto, C.M. Monyama, M.O. Chukwu, S.M. Lekala, M. Nchabeleng, M.R. Maloba, et al., Antibiotic resistance of Streptococcus agalactiae isolated from pregnant women in Garankuwa, South Africa, BMC Res. Notes 8 (2015) 364.
[72] S.D. Manning, Molecular epidemiology of Streptococcus agalactiae (group B Streptococcus), Front. Biosci. 8 (2003) s1–18.
[73] C.J. Baker, Group B streptococcal infections, Clin. Perinatol. 24 (1) (1997) 59–70.
[74] A. Schuchat, Epidemiology of group B streptococcal disease in the United States: shifting paradigms, Clin. Microbiol. Rev. 11 (3) (1998) 497–513.
[75] G.F. Mian, D.T. Godoy, C.A. Leal, T.Y. Yuhara, G.M. Costa, H.C. Figueiredo, Aspects of the natural history and virulence of S. agalactiae infection in Nile tilapia, Vet. Microbiol. 136 (1-2) (2009) 180–183.
[76] M. Chen, R. Wang, L.P. Li, W.W. Liang, J. Li, Y. Huang, et al., Screening vaccine candidate strains against Streptococcus agalactiae of tilapia based on PFGE genotype, Vaccine 30 (42) (2012) 6088–6092.
[77] J. Gao, F.Q. Yu, L.P. Luo, J.Z. He, R.G. Hou, H.Q. Zhang, et al., Antibiotic resistance of Streptococcus agalactiae from cows with mastitis, Vet. J. 194 (3) (2012) 423–424.
[78] S.B. Jamal, S.S. Hassan, S. Tiwari, M.V. Viana, L.J. Benevides, A. Ullah, et al., An integrative in-silico approach for therapeutic target identification in the human pathogen Corynebacterium diphtheriae, PLoS One 12 (10) (2017).
[79] A.B. Sjöstedt, Francisella, Bergey's Manual of Systematics of Archaea and Bacteria, John Wiley & Sons, Ltd, 2015.
[80] D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity, D.R. Boone, P. De Vos, et al., Bergey's Manual® of Systematic Bacteriology, Springer-Verlag, 2005.
[81] M. Maurin, M. Gyuranecz, Tularaemia: clinical aspects in Europe, Lancet Infect. Dis. 16 (1) (2016) 113–124.
[82] M. Santic, M. Molmeret, Y. Abu Kwaik, Modulation of biogenesis of the Francisella tularensis subsp. novicida-containing phagosome in quiescent human macrophages and its maturation into a phagolysosome upon activation by IFN-gamma, Cell. Microbiol. 7 (7) (2005) 957–967.
[83] P. Larsson, D. Elfsmark, K. Svensson, P. Wikström, M. Forsman, T. Brettin, et al., Molecular evolutionary consequences of niche restriction in Francisella tularensis, a facultative intracellular pathogen, PLoS Pathog. 5 (6) (2009).
[84] L. Hadfield Ted, P. McEvoy, Y. Polotsky, V.A. Tzinserling, A.A. Yakovlev, The pathology of diphtheria, J. Infect. Dis. 181 (s1) (2000) S116–S120.
[85] V. Sangal, P.A. Hoskisson, Evolution, epidemiology and diversity of Corynebacterium diphtheriae: new perspectives on an old foe, Infect. Genet. Evol. 43 (2016) 364–370.
[86] V.J. Timms, T. Nguyen, T. Crighton, M. Yuen, V. Sintchenko, Genome-wide comparison of Corynebacterium diphtheriae isolates from Australia identifies differences in the Pan-genomes between respiratory and cutaneous strains, BMC Genomics 19 (1) (2018).
[87] A. Sing, R. Konrad, D.M. Meinel, N. Mauder, I. Schwabe, R. Sting, Corynebacterium diphtheriae in a free-roaming red fox: case report and historical review on diphtheria in animals, Infection 44 (4) (2016) 441–445.

[88] L. Corboz, R. Thoma, U. Braun, R. Zbinden, Isolation of Corynebacterium diphtheriae subsp. belfanti from a cow with chronic active dermatitis, Schweiz. Arch. Tierheilkd. 138 (12) (1996) 596–599.
[89] A. Kraszewska, Z. Anusz, Appearance in domestic animals of Corynebacterium diphtheriae and other Corynebacterium strains pathogenic for man, Przegl. Epidemiol. 33 (2) (1979) 269–276.
[90] L. Detemmerman, D. Rousseaux, A. Efstratiou, C. Schirvel, K. Emmerechts, I. Wybo, et al., Toxigenic Corynebacterium ulcerans in human and non-toxigenic Corynebacterium diphtheriae in cat, New Microbes New Infect. 1 (1) (2013) 18–19.
[91] B.A. Leggett, A. De Zoysa, Y.E. Abbott, N. Leonard, B. Markey, A. Efstratiou, Toxigenic Corynebacterium diphtheriae isolated from a wound in a horse, Vet. Rec. 166 (21) (2010) 656–657.
[92] B. Henricson, M. Segarra, J. Garvin, J. Burns, S. Jenkins, C. Kim, et al., Toxigenic Corynebacterium diphtheriae associated with an equine wound infection, J. Vet. Diagn. Investig. 12 (3) (2000) 253–257.
[93] K. Zakikhany, S. Neal, A. Efstratiou, Emergence and molecular characterisation of non-toxigenic tox gene-bearing Corynebacterium diphtheriae biovar mitis in the United Kingdom, 2003-2012, Euro Surveill. 19 (22) (2014).

5.6 Artigo II - New insights through re-sequencing and correction of assembly in *Corynebacterium pseudotuberculosis* genomes

*Thiago Jesus Sousa1, Anne Cybelle Pinto Gomide[1], Diego Neres1, Rodrigo Profeta[1], Arun Kumar Jaiswal[1], Rodrigo Dias[1], Flávia Aburjaile[1], Núbia Seyffert[2], Thiago Castro[2], Mateus Matiuzzi[3], Renata Faria[3], Bertram Brenig[4], Siomar Soares[5], Vasco Azevedo[1]**

[1]Laboratory of Cellular and Molecular Genetics, Institute of Biological Science, Department of Genetics and Evolution, Universidade Federal de MinaS Gerais, Belo Horizonte-Minas Gerais, Brazil

[2] Institute of Biology, Universidade Federal da Bahia, Salvador-Bahia, Brazil.

[3]Laboratory of Microbiology and animal immunology, Animal Science Department, Universidade Federal do Vale do São Francisco, Petrolina - Pernambuco, Brazil.

[4] Institute of Veterinary Medicine, University Göttingen, Göttingen, Germany.

[5]Department of Microbiology, Immunology and Parasitology. Institute of Biological and Natural Sciences. Universidade Federal do Triangulo MIneiro, Uberaba-Minas Gerais, Brazil.

*** Correspondence:**
Vasco Ariston de Carvalho Azevedo
vascoariston@gmail.com

Keywords: Pan-genomic, Illumina Hiseq, Mis-assemblies, SNPs, Cutinase. (Min.5-Max. 8)

Abstract

Due to the continuous increase of the accessibility of genomes sequenced on NGS platforms, it has been possible to cover different platforms and perform quality assembly, achieving better accuracy in the genomic data. However, a combination of high-performance assembly strategies, such as the optical map, ordering by reference, and validation by the paired library, is still necessary for robust results. From high quality assembled genomes, it is possible to perform comparative genomics to investigate the organism's biology based on the gene sequences and genome plasticity analyses. In this way, this work contemplates the genomic

analysis of 50 strains of *Corynebacterium pseudotuberculosis* biovar *ovis*, a bacterium distributed worldwide, which causes different diseases in different hosts. The genomes were sequenced, assembled, and validated for structural analysis. From this strategy, it was possible to update pan genomic studies, where new genes were identified and the absence of others that had not been previously described. Also, it was possible to identify SNPs and non-synonymous mutations. The new genome versions, including strains from different countries and hosts, with high quality, may contribute to discoveries on new targets for vaccines and drugs against this bacterium, which causes significant damage to agribusiness.

1 Introduction

Corynebacterium pseudotuberculosis is a pathogenic agent of Caseous Lymphadenitis in goats and sheep, but in other small ruminants too (Fontaine and Baird, 2008). Within the same genus, it can be highlighted other pathogenic bacteria, such as *Corynebacterium diphtheriae*, responsible for diphtheria in humans (Hou et al., 1997); *Corynebacterium glutamicum* and *Corynebacterium efficiens*, bacteria used in the production of amino acids, such as L-aspartate and L-lysine (Moore et al., 2010); and *Corynebacterium ulcerans*, also of medical-veterinary interest for animal production, such as pigs (Contzen et al., 2011).

C. pseudotuberculosis presents a worldwide distribution, but the largest epicentre of its infections is found in the northeast region of Brazil. This fact is due to the dry climate and the vegetation of that area, such as the caatinga, in which the host *Capra hircus* is fully adapted (Magalhães et al., 2018). Moreover, cactus specimens' remarkable presence in caatinga flora facilitates the development of wounds on the animal skin through scratches, making it more prone to bacterium installation and a further infection process (Dorella et al., 2006).

In Brazil's northeast region, the largest production of goats can be found, which represents 93.2% of the national territory, with 8,944,461 animals (Magalhães et al., 2018). The estimate of infection by this disease is 78.9% (Seyffert et al., 2010), which impacts goat farming activity. It also reduces milk production, weight loss in beef goats, and full use of the skin due to the wounds caused by granulomas' formation in superficial lymph nodes. Although it is not considered lethal, the economic losses are significant due to the reduction of milk, meat, and leather production, directly affecting the farmer's income (Barnabé et al., 2019).

Within the *C. pseudotuberculosis* species, there is a subdivision into two biovars based on reducing nitrate. Those are called *ovis* (negative nitrate reduction) and *equi* (positive nitrate reduction) (Chaudhari et al., 2016). In Brazil, there are no reports in the literature on cases of

infection caused by biovar *equi*; however, it is highly prevalent in horses in other parts of the world, mainly in the United States (Spier et al., 2012).

The current control measures of *C. pseudotuberculosis* are based on the isolation of infected animals and cleaning with aseptic solutions after surgical removal of granulomas present on the surface between the epidermis and dermis of these animals (ALVES et al., 1997). However, the best strategy would be through immunoprophylaxis using vaccines, allowing the herd protection against this bacterial infection and not allowing the infection to remain (Fontaine et al., 2006).

Obtaining the complete DNA sequence of a pathogenic organism is extremely useful for molecular biology studies with emphasis on the search for therapeutic targets, gene completeness, evolutionary processes, mutations, among other characteristics (Lee et al., 2016). In this context, the pan-genomic analysis of *C. pseudotuberculosis* strains may clarify the presence and absence of virulence, resistance to antibiotics, and essential metabolic processes related to different aspects of the infectious process. This sequencing must be conducted comprehensively, especially when the genome in question has excellent sequencing and assembly quality. In this context, errors in a genome assembly are detrimental to the analyses since they generate dubious interpretations that may not reflect reality. Therefore, the improvement of deposited assemblies, even for complete genomes, is important and essential to avoid error propagation in biological databases.

In 2013, Soares and colleagues carried out a pan-genome study with 15 *C. pseudotuberculosis* strains. In summary, they found 16 pathogenicity islands and confirmed the presence of virulence genes such as *pld*, *fag* operon, *spaD* and *spaA* and pilus in all strains, and the diphtheria toxin (tox) exclusively in *C. pseudotuberculosis* 31(biovar *equi*)**.** However, the new genomes and the corrected assemblies (Sousa et al., 2019) lead to new investigations on the biovar *ovis* pan-genome and are useful in detecting new prophylactic targets. Thus, our goal is to investigate whether the new techniques and approaches used to improve and correct genomes may provide new structural and functional genomic information for *C. pseudotuberculosis*, emphasizing biovar *ovis*. For this purpose, we sequenced 50 strains with the Illumina Hiseq 2500 platform and compare it with the old versions to propose new genomic evidence not yet explored.

2 Materials and Methods

Sequencing: To support the pan-genome study of *C. pseudotuberculosis* isolates belonging to the biovar *ovis*, a total of 50 strains from different regions of the world were selected. Among these 50 strains, 44 new isolates have not yet been used in other studies, and among these 44 strains, 31 were isolated in the region around Petrolina, in the state of Pernambuco – Brazil. More details about the isolates can be seen in Supplementary Table 1. All samples were cultured in a solid medium containing 1.5% bacteriological agar. Subsequently, isolated colonies were grown in brain-heart infusion broth medium (BHI-Hi Media Laboratories Pvt. Ltd, India) supplemented with 0.5% Tween 80 at 37°C for 20 hours under agitation. The genomic DNA extraction followed a protocol previously established (Pacheco et al., 2007). The DNA was then sent for sequencing in the Illumina HiSeq 2500 platform using a 450 bp paired-end library (2x151 bp each read and 298-300 bp of insert), with an insert size of ~ 450 bp, at the Institute of Veterinary Medicine at the University of Göttingen in Germany.

Genome assembly and annotation: The analysis of the raw reads' quality was performed using the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), with all data presenting an average Phred score of more than 35. We applied an ab initio assembly strategy for the construction of the contigs using Spades v. 3.14.0 (Bankevich et al., 2012) and Edena version 3.131028 (Hernandez et al., 2014). The best assembly was then selected via QUAST software (Gurevich et al., 2013). For *C. pseudotuberculosis* CpC231, Cp1002, CpI19, CpFRC41, Cp29156, and CpT1, contigs were ordered and oriented using the MapSolver™ (OpGen) optical map based on enzyme restriction sites. The other 44 strains (Table 1) were mapped using *C. pseudotuberculosis* strains as a reference, presenting 99 % of gene content similarity. Afterwards, the *dnaA* gene position was corrected based on the origin of replication C (*oriC*) present in the scaffolds using an *in-house* script. The gap closure step was carried out through a pipeline in Shell to organize and optimize the steps, and through a pipeline to organize and optimize the steps, named as manage contigs v.1.3 (available in https://github.com/Thiagojsousa/Genome_Scripts). This pipeline includes the software CONTIGuatorF (Galardini et al., 2011), GapBlaster v. 1.1.2 (de Sá et al., 2016), Gfinisher v. 1.4 (Guizelini et al., 2016) and CLC Genomics Workbench (CLC-genomics Workbench) version 7.0 (Qiagen). Annotation was performed via PROKKA v. 1.4.6 (Seemann, 2014). The 50 complete genomes were deposited in the GenBank - NCBI database under the accession numbers described in supplementary table 1.

Pan-genomics analysis: Pan-genomics analysis was performed with the ROARY software (Page et al., 2015) for the *C. pseudotuberculosis* biovar *ovis* 50 strains. Moreover, OrthoFinder software (Emms and Kelly, 2015) was also used for the 15 strains previously investigated by

Soares and colleagues in 2013 (Soares, 2013) to maintain the same methodology used in this previous study.

SNPs analysis: To detect SNPs in all the 50 strains and perform cluster visualization, Parsnp v.1.2, and Gingr version v.1.2 were used (Treangen et al., 2014). Python scripts were built to manipulate the files and analyze the synonymous mutations.

3 Result and discussions

To initiate our proposal, sequencing and assembly of the 50 genomes of biovar *ovis* by *ab initio* assembly method were necessary. This sequencing technology was selected (Illumina Hiseq 2500) due to the high quality of reads with an average quality score Phred score of more than 35 and coverage of ~900-fold. The *ab initio* assembly method was efficient, which made it possible to generate assemblies with an average of 8 contigs per genome. Then, after gap closure, all 50 genomes were deposited as complete genomes at GenBank (Table 1).

The main problem in the assembly of *C. pseudotuberculosis* genome, which limits the construction of a single contig by the ab initio method, are the regions of Ribosomal RNA (rRNA). There are four rRNA clusters in *C. pseudotuberculosis*, so the minimum expected number is five contigs or four gaps per genome. These rRNA clusters are repetitive regions, which encodes the 5S, 16S, and 23s gene structures, with approximately 6000 bp and similarity ranging from 98.04% to 99.94%. The heuristic analysis applied by *the ab initio* assembly software cannot differentiate the reads of each cluster on the genome chromosome, and finally concentrate in a single contig. Even with this obstacle, the optical map strategy represents an efficient approach to order these contigs and thus avoid chromosomal rearrangements but only in CpC231, Cp1002, CpI19, CpFRC41, Cp29156 e CpT1, as demonstrated on methodology.

3.1 Comparison between the genomes of the first version of the first pan-genome of *C. pseudotuberculosis* with the new genomes.

Soares *et al.* (Soares et al., 2013) described the first pan-genome of *C. pseudotuberculosis* with fifteen strains, where nine were *ovis* strains and six *equi* strains. We focus on *ovis* strains for this current work, those strains have been re-sequenced and re-assembled (Tabel 2). In previous work, as presented by Sousa *et al.* (Sousa et al., 2019), using optical map data that were not available previously, evidence of assembly errors were detected in the genomes used for the first pan-genome analysis. Thus, it was essential to reassess them to identify the influence of these errors.

The current pan-genome, considering the 9 strains of *C. pseudotuberculosis* of biovar (*ovis* and *equi* strains), showed a total of 2635 genes, and in the previous study, 2782 total genes were represented. This small difference may be due to updates on the GenDB platform since 2012 or associated with genes encoding to small proteins. The pan-genome of *C. pseudotuberculosis* increased had an α of 0.912 under alpha value, indicating that it has an open pan-genome. The core genes were previously represented as 1504 genes and now 1820. The tg (θ) value for all genomes (15 strains) was nine, which means that each sequenced genome added approximately nine genes as singletons to the total gene pool for the species. This evidence in the core genome's value among the 15 strains demonstrates the new assemblies' contribution to the new pan genomic analysis and probably new genes not represented until now.

Considering only the biovar *ovis* strains, a current analysis of a total of 2358 genes was obtained as a pan-genome --- antes era 2,403. The biovar *ovis* pan-genome increased by 0.948, showing that it is still growing more gradually than the pan-genome of all species. According to the exponential regression decay, the core genome genes of the biovar *ovis* are represented by1945 genes and tend to stabilize in approximately 1748 genes. In the previous study, it was shown that this value was 1818 genes. The tg (θ) for the biovar *ovis* was 2.5, which means that each sequenced genome added approximately 2.5 genes as singletons to the full gene set of the *ovis* subset.

Finally, within the biovar *equi*, we obtained a pan-genome with 2366 genes in the current analysis – antes era 2,521. The biovar *equi* pan-genome has an α of 0.923, indicating that it has an open pan-genome. According to the exponential regression decay, the core genome genes are 1891 genes and tend to stabilize at approximately 1545 genes. Before, it was reported a 1599 genes core genome. The tg θ for biovar *equi* was 18.67, which means that each sequenced genome added approximately 18 genes as singletons to the complete set of genes in the *equi* subset.

When comparing the result presented by Soares (Soares et al., 2013) with the new versions of the re-sequenced genomes in this work (Table 2), it is noted that there was an increase for the core genome and singletons genes. This first comparison was carried out to assess whether there was an influence of the quality of the assembly in the genome's genetic diversity. This proves that the new genome versions contributed to the addition of new genes, which were not present in the first assemblies.

3.2 Pan-genomics of biovar *ovis* comparing the 50 strains sequenced by Illumina Hiseq 2500.

Our intention in proposing a new pan-genome for the biovar *ovis* was based on the previous analysis results, where an increase in the number of genes in the re-sequenced genomes was observed. Thus, discussing the results with these new genomes characterized by excellent quality and accuracy could contribute to new biological information related to the microrganism-host interaction and prophylaxis. The analysis was performed considering 50 strains obtained from different hosts and geographic regions, and all of them sequenced by the Illumina Hiseq 2500 platform.

3.2.1 Analysis of genomic similarity and prediction of genomic islands.

Comparative analysis allows us to acquire knowledge about the genome's structure in question and to suppose evolutionary processes of losses or gains of genes. Thus, through a genomic similarity analysis carried out by BRIG, it was possible to observe that all genomes have high similarity compared to the reference strain (*C. pseudotuberculosis* 414). This strain was used as a reference because it showed the ability to form biofilm at the site of adhesion (personal communication).

All 50 strains shared the 17 genomic islands predicted through GIPSy (Figure 1). Nine pathogenicity islands (PI) were identified, with emphasis on GI7, GI8, which have the virulence genes *ureB*, *srtA*, *spaI*, *spaC*, *spaC2*, *spaC3,* and GI9 with the genes *srtB*, *spaD*, *srtC*, *spaA*.

The *ureB* gene is responsible for the synthesis of the beta subunit of the urease enzyme. This enzyme is a metalloenzyme that r*equi*res nickel and acts in the hydrolysis of urea in ammonia and carbon dioxide in the gastric mucus layer to facilitate its initial interaction in this acidic environment. In addition to being able to induce the production of antibodies and T lymphocytes, as shown in *Helicobacter pylori,* where it is considered a virulence factor (Chmiela and Kupcinskas, 2019) *srtA*, *spaI*, *spaC*, *spaC2*, *spaC3, srtB*, *spaD*, *srtC*, s*paA* are part of the pili complex, which are structures responsible for adhesion, the onset of infection and proliferation of bacteria, and act in the communication between the extracellular and intracellular environment in *Corynebacteria species* (Ott, 2018).

Regarding resistance islands (RI), four were predicted, with emphasis on the GI12 that has the genes *ciuA*, *ciuB*, *ciuC*, *ciuD* e *ciuE*, which are genes of the ciuABCDEF operon, a siderophore cluster that acts on the uptake and transport of iron by bacteria (Ling et al., 2013; Ibraim et al., 2019). Virulence genes have also been predicted in other regions, such as *pld*, *fagC*, *fagB*, *fagA*, *fagD*, *rmlB*, *hmuT*, *hmuU*, *hmuV*, *sodA*, *dtxR*, *glnA1*. The *pld* gene is translated into the phospholipase D toxin, which presents an exotoxin action, being one of the main virulence factors in *C. pseudotuberculosis* already described (Tachedjian et al., 1995;

Menzies et al., 2004; Leal et al., 2018). The *rmlB* gene encodes the enzyme dTDP-D- glucose-4,6-dehydratase. In *Mycobacteria,* it has already been tested using mutants, along with *rmlC,* and demonstrated to be essential for the bacteria's growth and survival (Li et al., 2006). The *hmuT (HBP /substrate-binding protein)*, *hmuU (permease)*, *hmuV (ATPase)* genes have already been described in *C. diphtheriae* in the operon *hmuTUV. The three genes, described before, probably act as a second means of acquiring iron, now through heme transport proteins in the Corynebacterium genus (Draganova et al., 2015). The sodA gene has already been studied in C. glutamicum and encodes a superoxide dismutase* (El Shafey and Ghanem, 2015). Finally, the *glnA1* gene is related to the synthesis of glutamine synthetase, involved in nitrogen metabolism via ammonium assimilation. In *M. tuberculosis, glnA1*knockout mutants resulted in a decrease in cellular appearance's rigidity and strength (Tripathi et al., 2015). Among the genes mentioned above, *ureB*, *rmlB*, *sodA*, *glnA1* had not yet been explored and probably, not predicted and identified in the first pan-genomics work proposed by Soares in 2013 and later works with *C. pseudotuberculosis*.

Still, regarding the similarity between the strains, the only difference observed was in the Cp1002 strain and other mutants (SigK, SigM, SigC, SigD, and SigB), as highlighted in the BRIG map (Figure 1), where the gene for the cutinase family protein appears. In Figure 2 it is possible to observe the cutinase protein flanked by regions of tRNA-Glu (on the left) and two more for tRNA-Glu and tRNA-Gln (on the right).

To certify this gene's presence in the *Corynebacterium* genus, a search by BLAST was performed on the complete genomes of all species belonging to the genus *Corynebacterium* deposited in the NCBI, (58 genomes), and the homologous cutinase gene was present in only eight *C. pseudotuberculosis* species. Figure 3 shows the phylogenetic reconstruction using the Maximum Likelihood method, where it is possible to perceive a clade with 100% bootstraps aggregating *C. pseudotuberculosis*, *C. ulcerans*, *C. diphtheriae*, *C. vitaeruminis, C. sphenisci*. *C. xerosis* shares other clade*s* with Mycobacterium, Rhodococcus hoagi, and Nocardia farcinis as an outside group *matruchotii* and *Corynebacterium imitans* were more distant from the genus *Corynebacterium* than expected, however, with no statistical support for the bootstraps in these clades.

Pan-genome analysis of *C. pseudotuberculosis* biovar *ovis* identified two proteins predicted as cutinase family protein, the first one with 308 amino acids (Locus tag: CpCAP414_03265) (named *cut1* gene) and the second with 469 amino acids (Locus tag: CpCAP414_08590) (named *cut2* gene). Despite having homologous domains, both are distinct, presenting only 25% identity in 40% coverage. Interestingly, *cut2*, the second cutinase gene (CpCAP414_08590) (Figure 2), is absent in the Cp1002 strain. This gene's absence was not due to an assembly error because when analyzing the content generated by Edena and SPAdes, this region was also not represented, and no

content was lost. To investigate the genetic loss, we analyzed 115 complete genomes of *C. pseudotuberculosis* deposited in the NCBI, and after comparing through BLAST the similarity of the genomes with *cut2* sequence, we observed a lack of the *cut2* gene in the Cp1002 and CpPAT10 strains, both flanked by tRNA and tmRNA genes. As already demonstrated in the literature, regions that are flanked by tRNA and tmRNA genes are hotspots for insertion elements (Hou, 1999).

Cutinases are enzymes that perform the hydrolysis of insoluble biopolyesters, and cutinase is a structural component of the plant cell wall (Chen et al., 2008). Thus, studies have already shown the action of cutinase in the fungal infection process in plants, as in *Arabidopsis thaliana* infection by *Phytophthora brassicae*, being described as a virulence factor for the genus *Phytophthora* (Belbahri et al., 2008). Studies with cutinase in the genus *Corynebacterium* are limited; however, there are reports of its expression as a heterologous protein in *C. glutamicum* for industrial purposes (Hemmerich et al., 2019). The presence of cutinase in the genus *Mycobacterium* has already been reported, which is interesting since it is not a plant pathogen. In *M. tuberculosis*, West, and collaborators (West et al., 2009b) described seven proteins predicted as cutinase-like proteins, being that 6 of them have a secretion signal. The study went deeper into the six proteins, showing that they are homologous to the fungi' cutinases, such as *Fusarium solani*. These cutinases are alpha/beta hydrolases with three amino acid residues in the active site, namely serine, aspartic acid, and histidine. The reference article tested these enzymes' action on the cutin substrate and did not obtain degradation results. The proteins predicted as cutinases in *M. tuberculosis* showed a fatty acid degradation activity along with esterases or lipases (West et al., 2009a).

Among the six supposed cutinases, *Culp1* and *Culp2* demonstrated the potential to act as a virulence factor in *M. tuberculosis*. *Culp1* demonstrated to be a potential vaccine target, as it is secreted in the extracellular environment and interacts with host molecules, contributing to the virulence and intracellular survival of *M. tuberculosis*. *Culp1* is unlikely to be essential for the viability of *M. tuberculosis*, as it is absent in most *M. bovis* (BCG) vaccinal strains *(Mahairas et al., 1996)*.

The cut2 present in *C. pseudotuberculosis* has ~ 25% similarity and ~ 46% coverage with the *M. tuberculosis* proteins, and it still retains the GxSxGA protein domain (Figure 4). It is possible to be the same prediction error, as demonstrated in *M. tuberculosis*. As reported above, in addition to a small part of the bacteria of the genus *Corynebacterium* having at least one copy, it is absent in non-pathogenic species such as *C. glutamicum* and *C. efficiens* and present in more pathogenic bacteria, such as *C. pseudotuberculosis*, *C. ulcerans*, *C. diphtheriae*, *C. vitaeruminis*. Thus, *cut2* may not be an essential virulence factor, but it may affect the process of interaction with the host and assist in the survival of bacteria within the intracellular environment. *Cut2* may have been lost

by the *C. pseudotuberculosis* 1002 strain in the evolutionary process through two tRNA regions in the extremities.

Pan-genomic analyses made with 50 *C. pseudotuberculosis* strains were performed using Roary. Due to the clonal characteristics of *C. pseudotuberculosis*, 92.06% of the genes, of a total of 2088 genes, belong to the central or core genome. Regarding single genes or singletons, 50 genes were found. Finally, there are 130 genes shared between 2 to 49 strains, and these values are better described in Figure 4, as Roary splits them into soft-core, shell, and cloud.

3.2.2 Analysis of nucleotide mutations.

When comparing the genome sequence of individuals of the same species, a similarity value above 99% is expected since changes of one or two bases may represent differences. These events are characterized as Single Nucleotide Polymorphism (SNP), and its occurrence can significantly impact significantly when changes the amino acid transcription codon. When these nucleotide variations occur in at least 1% of a restricted population, they are considered SNPs; if not, the nucleotide variation is a simple mutation. Considering the gene density found in prokaryotes, SNPs' incidence can be high and drastically influences transcription, protein structures, phenotype modification, and survival (Brookes, 1999).

In *C. pseudotuberculosis,* there are two short reports of SNPs, first in 2016 by Almeida and collaborators (Almeida et al., 2016), when comparing the CpVD57 strain (biovar *ovis*) with 18 other strains (biovar *ovis* and *equi*). In the Almeida article, we have a discrepancy of 35 SNPs in Cp1002 and 2.404 SNPs for Cp267, both from the biovar *ovis*. For the strains of biovar *equi*, we have 25.905 SNPs in Cp316 (biovar *equi*) and 24,274 in Cp162. Another study in 2017 by Baraúna and collaborators (Baraúna et al., 2017), now analyzing 18 strains of the biovar *equi*, did not identify the number of SNPs and just used them phylogenetic reconstruction. The authors considered that the SNPs induce a similar topography of the phylogenetic tree compared to the reconstruction using the core genome. However, both previously cited studies used sequencing by the IonTorrent PGM platform, which has indels (insertion or deletion of nucleotides), mainly in regions of homopolymers. When generating this data, the author does not use the Ion Torrent PGM Hi-Q Sequencing Kit. This Kit reduces the insertion of homopolymers, assists in reading quality, and improves the final genome (Pereira et al., 2016). Considering this situation, we use in this work sequencing data obtained through the Illumina Hiseq 2500 paired-end platform (Quail et al., 2012).

Thus, with an entire set of 50 strains of the biovar *ovis* and using the Cp1002 strain (GCA_000144935.3) as a reference, a total of 3410 SNPs were detected. After comparisons and gene predictions, 2836 mutations were detected in protein-coding regions. The codon changes

represented 1693 non-synonymous mutations, which would lead to an amino acid change in the protein sequence.

From the SNPs' identification, it was possible to perform a clustering based on these changes (Figure 7). Despite the high level of similarity, it was possible to isolate the strains according to the country and/or region of isolation, although the strains isolated in Brazil were in different clades. It was possible to group some strains according to the place of isolation, such as Cp99MAT, Cp04MAT, Cp87MAT, all of them isolated on Fazenda Matadouro - Petrolina; the Cp1002 (and its six mutants) and CpT1 were isolated in Bahia state – Brazil. The CpE13, CpE7, CpE9, CpE16, CpE14, Cp414 were isolated in Fazenda POÇO V (Embrapa) - Petrolina; The CpPAT16 and CpPAT14 were isolated in the region of Patagonia – Argentina. The CpI19 and Cp29156 were isolated in the region of Rehovot (Israel). This may be an indication of how the environment can influence the selective pressure on *C. pseudotuberculosis* genes.

4 Conclusion

The new versions of the genomes herein presented show unpublished results for the *C. pseudotuberculosis* pan-genome analysis. Thus, the investment in re-sequencing through the Illumina Hiseq platform using paired-end strategy and new techniques such as the optical mapping made it possible to improve genomes' quality and thereby explore new genes. The function prediction based on protein domains, especially when it has standard data from distant organisms, can generate function determination errors. So, more attention is needed to the final results obtained through this type of analysis. The identification of SNPs showed that possibly some changes in the use of codons by bacteria might exist due to environmental pressures. These mutations may influence the selection of targets to develop a more specific immunoprophylaxis tool for *C. pseudotuberculosis* infection.

5 Figures

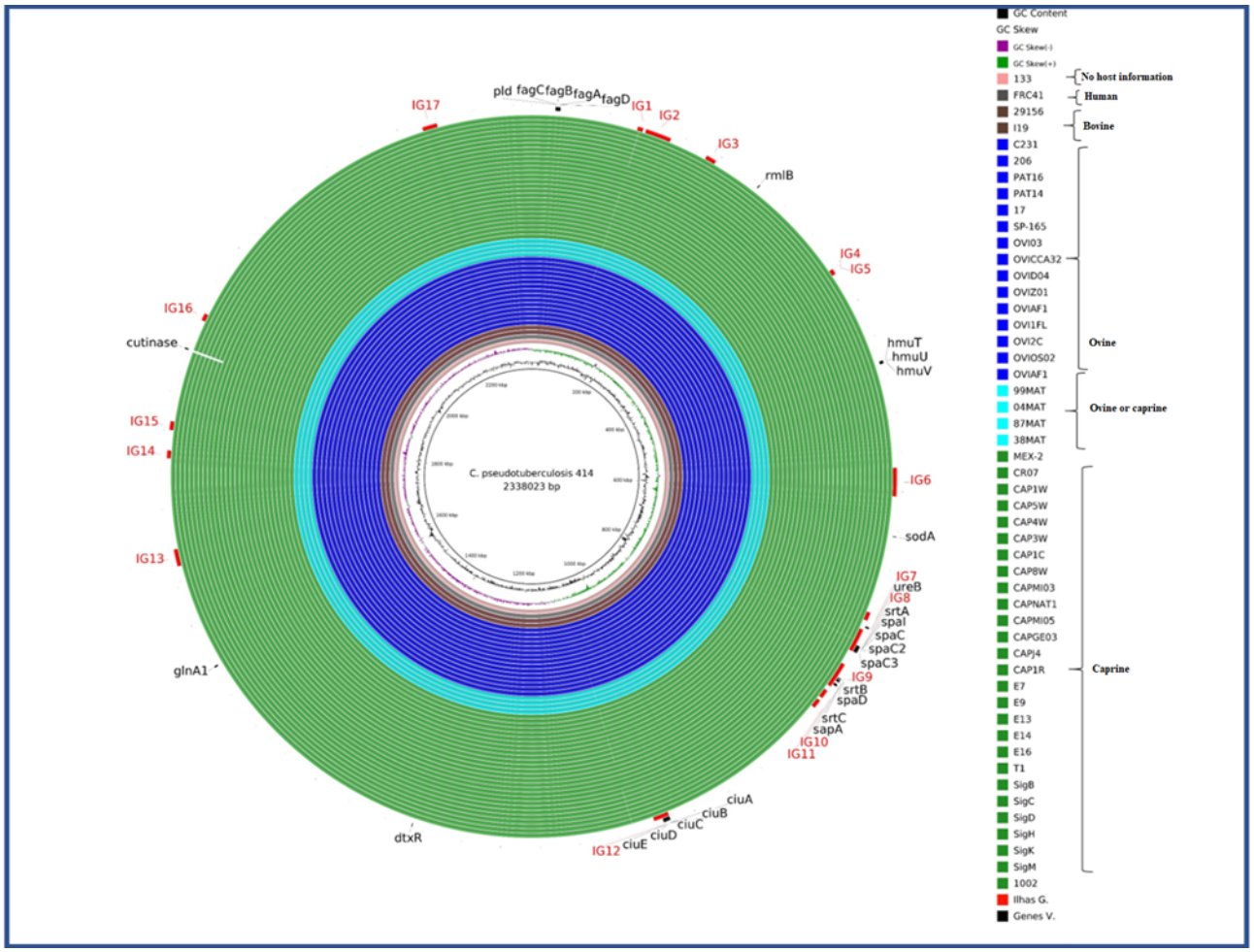


Figure 1 - Representation of the results obtained through the BRIG software with all the *C. pseudotuberculosis* biovar *ovis* 50 genomes.

Figure 2 - Visualization by the artemis software of the cutinase gene in the genome of the *C. pseudotuberculosis* 414 (below) and *C. pseudotuberculosis* 1002 (above) strains.

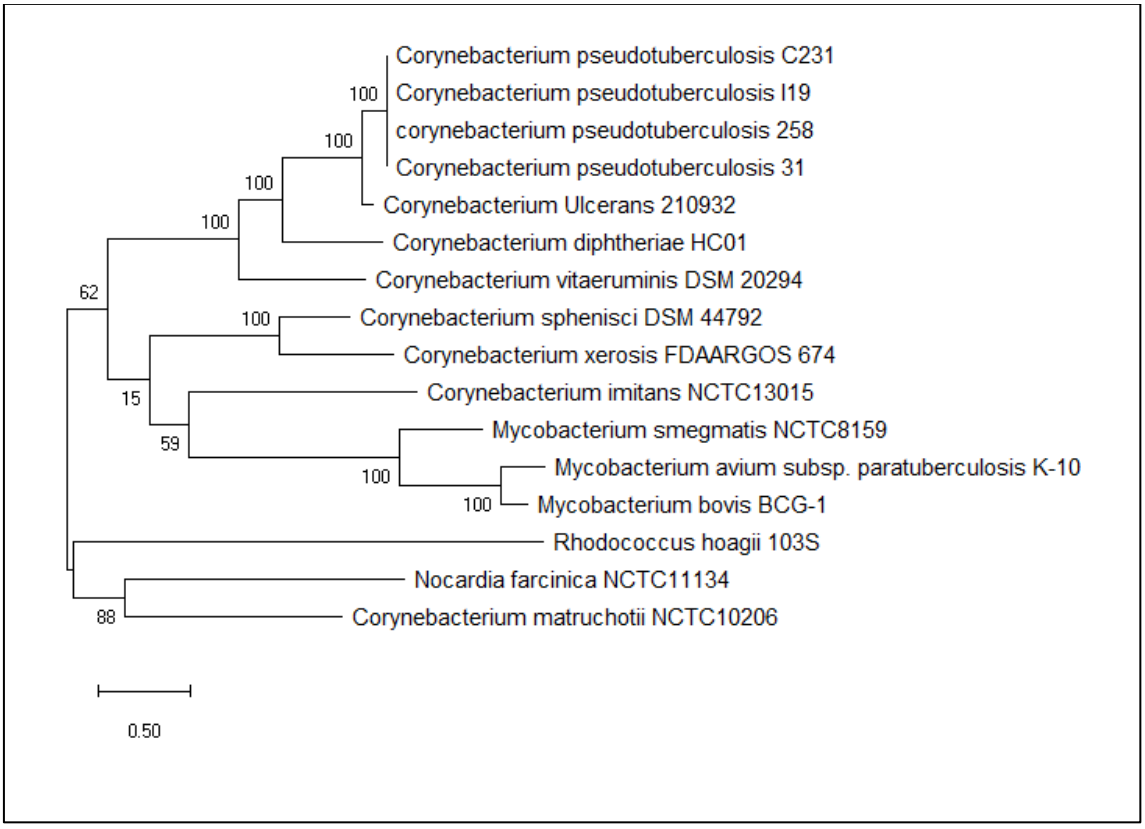


Figure 3 – Phylogenetic reconstruction of the cutinase gene using the maximum likelihood method in strains of the CMNR group.

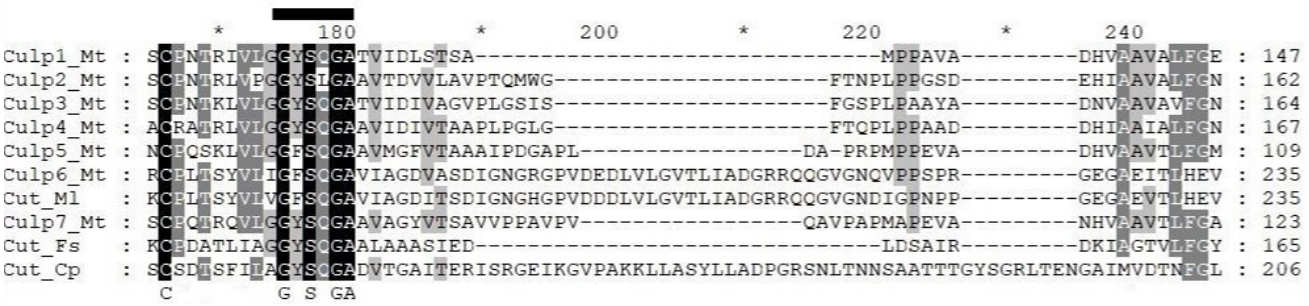


Figure 4 - Multiple alignment of cutinases.

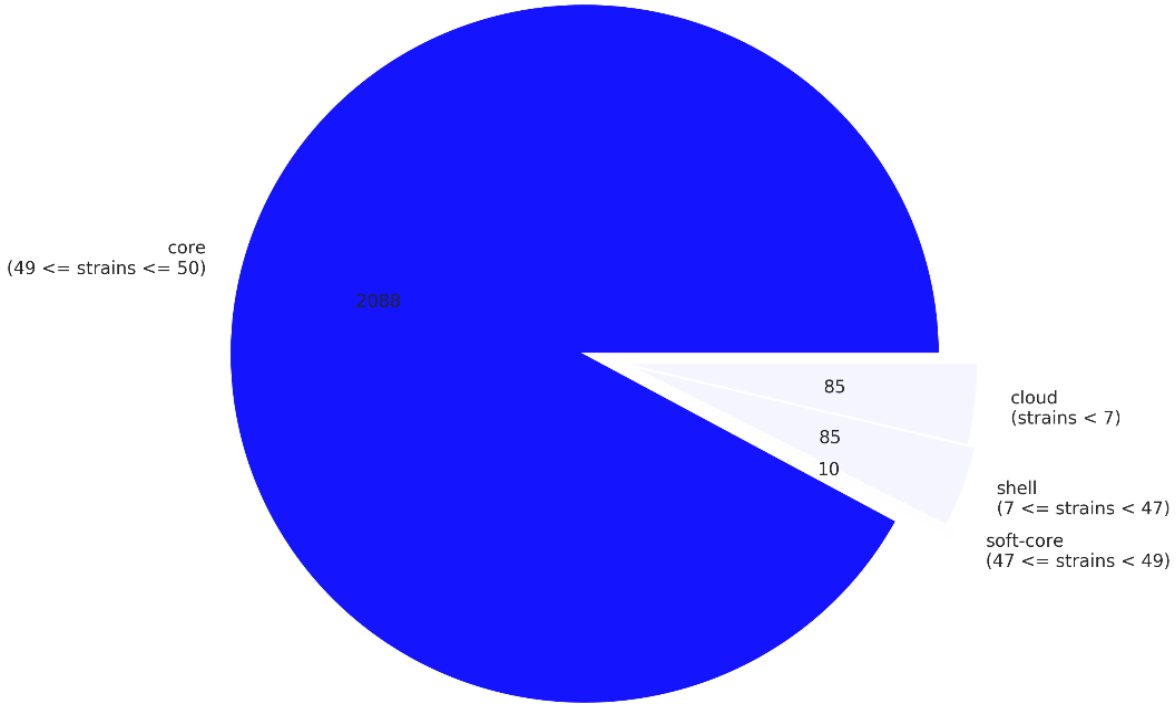


Figure 5 - Distribution of genes belonging to the pan-genome of the 50 strains of *C. pseudotuberculosis* biovar *ovis*. Core genes are found in > 99% of genomes, soft-core genes are found in 95-99% of genomes, shell genes are found in 15-95%, while cloud genes are present in less than 15 % of genomes.

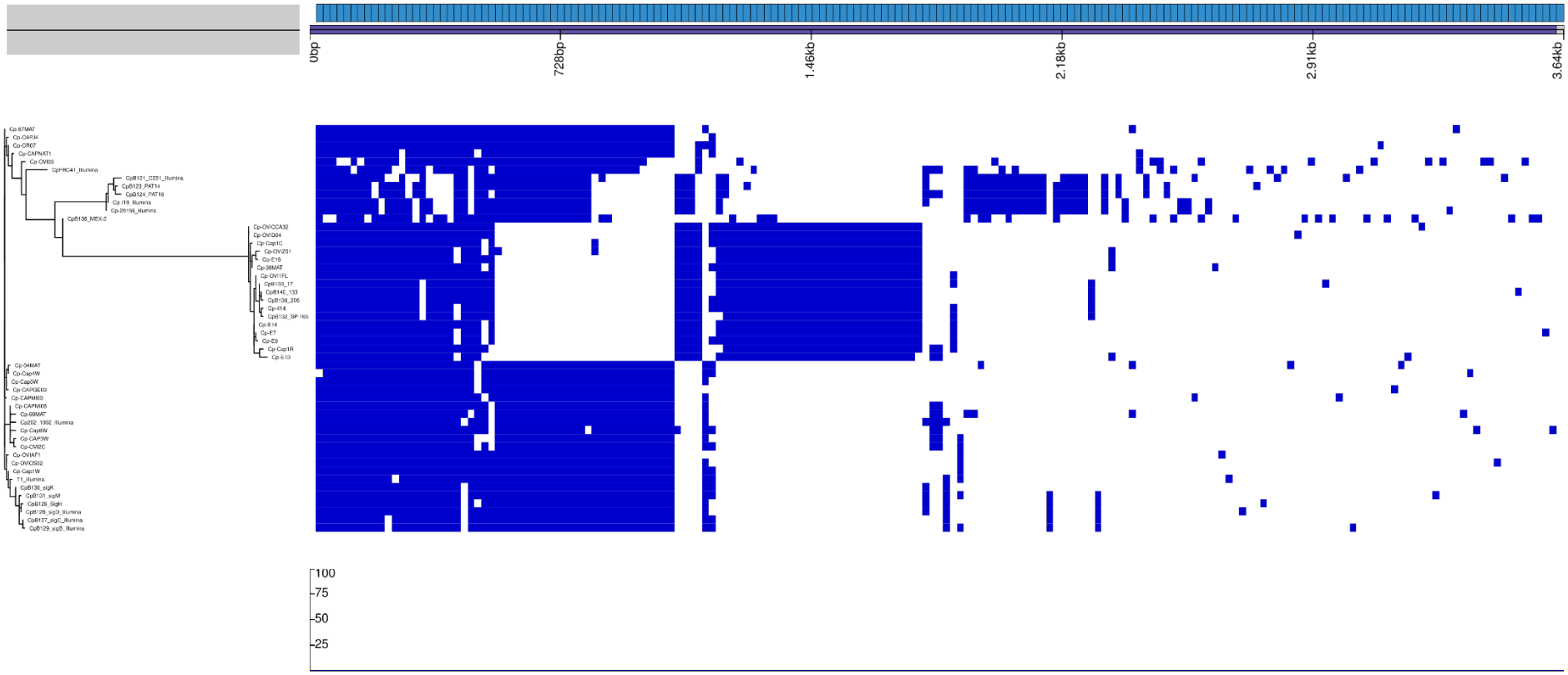


Figure 6 - Clusters between strains without core genome genes. The blocks of the image represents are genes belonging to the accessory genome or shared genome.

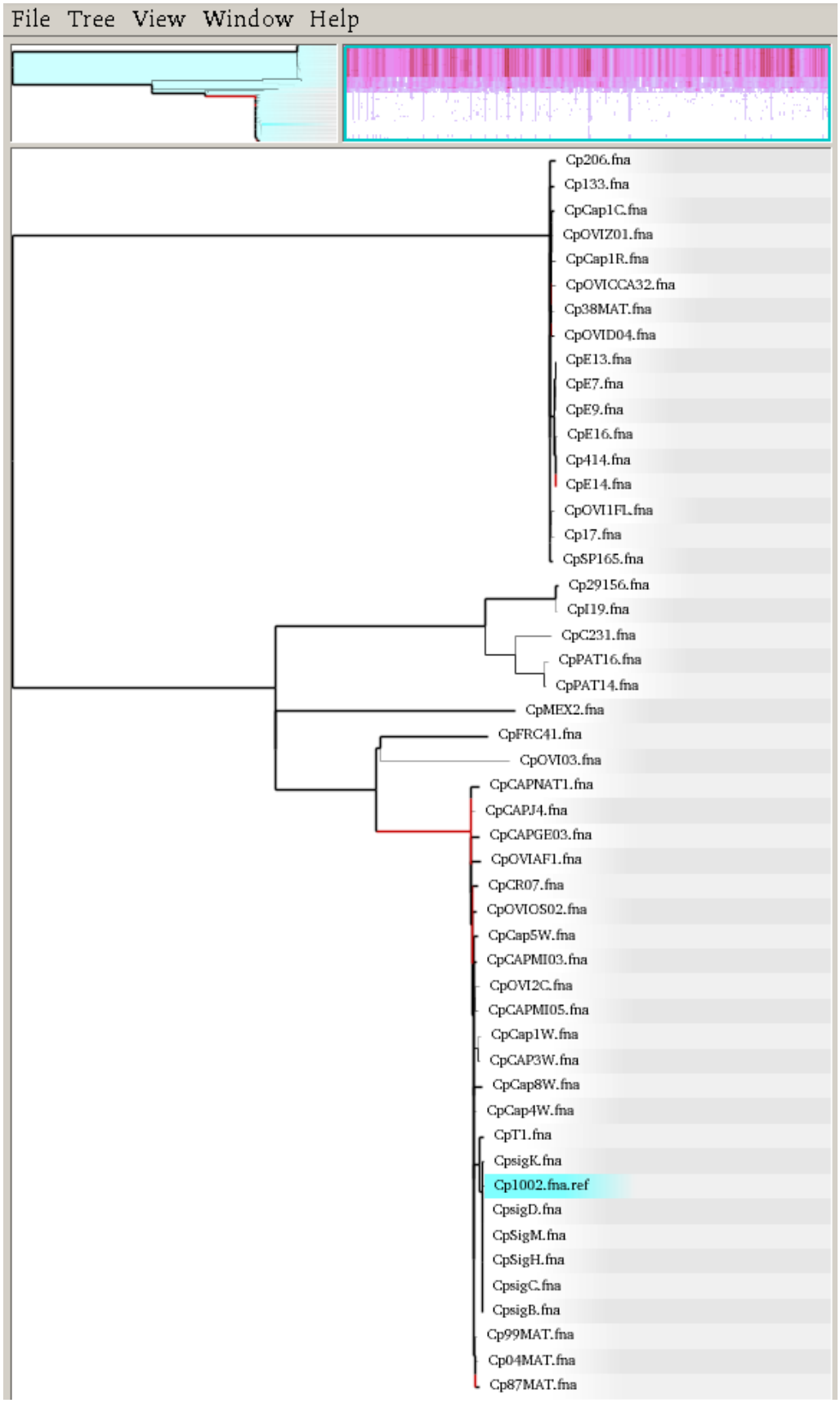


Figure 7 - Clustering and graphical representation based on the SNPs found in *C. pseudotuberculosis* genomes using the Cp1002 strain genome as reference.

6 Tables

Tabel 1 - Data on the 50 strains of *C. pseudotuberculosis* biovar *ovis* used in this study.

Strains	Host	Country	Access Number	Genome (Mb)
C. pseudotuberculosis C231*	Ovine	Australia	GCA_000144675.2	2,33725
C. pseudotuberculosis 1002*	Goat	Brazil	GCA_000144935.3	2,33832
C. pseudotuberculosis I19*	Bovine	Israel	GCA_000152065.3	2,33821
C. pseudotuberculosis FRC41*	Human	France	GCA_000143705.2	2,33822
C. pseudotuberculosis 29156	Bovine	Israel	GCA_001026945.2	2,33775
C. pseudotuberculosis T1	Goat	Brazil	GCA_001682255.2	2,33800
C. pseudotuberculosis CAP3W	Goat	Brazil	GCA_002953275.1	2,33818
C. pseudotuberculosis CAPJ4	Goat	Brazil	GCA_002953315.1	2,33808
C. pseudotuberculosis OVI2C	Ovine	Brazil	GCA_002953235.1	2,33802
C. pseudotuberculosis OVI03	Ovine	Brazil	GCA_002953955.1	2,33811
C. pseudotuberculosis Cap1W	Goat	Brazil	GCA_003955885.1	2,33817
C. pseudotuberculosis OVIAF1	Ovine	Brazil	GCA_003955905.1	2,33804
C. pseudotuberculosis 87MAT	Goat or Ovine	Brazil	GCA_004193955.1	2,33796
C. pseudotuberculosis CAPMI03	Goat	Brazil	GCA_004193675.1	2,33812
C. pseudotuberculosis CAPMI05	Goat	Brazil	GCA_004193695.1	2,33798
C. pseudotuberculosis CAPNAT1	Goat	Brazil	GCA_004193655.1	2,33792
C. pseudotuberculosis CR07	Goat	Brazil	GCA_004193635.1	2,33794
C. pseudotuberculosis 99MAT	Goat or Ovine	Brazil	GCA_004295625.1	2,33797
C. pseudotuberculosis E7	Goat	Brazil	GCA_004353245.1	2,33778
C. pseudotuberculosis Cap1R	Goat	Brazil	GCA_004323775.1	2,33772
C. pseudotuberculosis Cap8W	Goat	Brazil	GCA_004323755.1	2,33803
C. pseudotuberculosis OVI1FL	Ovine	Brazil	GCA_004941425.1	2,33807
C. pseudotuberculosis Cap4W	Goat	Brazil	GCA_005341465.1	2,33805
*C. pseudotuberculosis*Cap5W	Goat	Brazil	GCA_005341445.1	2,33795
C. pseudotuberculosis 04MAT	Goat or Ovine	Brazil	GCA_004332375.1	2,33801
C. pseudotuberculosis 38MAT	Goat or Ovine	Brazil	GCA_004332355.1	2,33771
C. pseudotuberculosis OVICCA32	Goat	Brazil	GCA_006974065.1	2,33767
C. pseudotuberculosis OVID04	Ovine	Brazil	GCA_006971245.1	2,33810
C. pseudotuberculosis OVIOS02	Ovine	Brazil	GCA_006971625.1	2,33793
C. pseudotuberculosis OVIZ01	Ovine	Brazil	GCA_006971425.1	2,33781
C. pseudotuberculosis CAPGE03	Goat	Brazil	GCA_006971065.1	2,33806
C. pseudotuberculosis E13	Goat	Brazil	GCA_008461625.1	2,33784
*C. pseudotuberculosis*E14	Goat	Brazil	GCA_008462025.1	2,33766
C. pseudotuberculosis E16	Goat	Brazil	GCA_008461845.1	2,33785
C. pseudotuberculosis Cap1C	Goat	Brazil	GCA_004771435.1	2,33814
C. pseudotuberculosis 17	Ovine	Brazil	GCA_009759745.1	2,33774
C. pseudotuberculosis MEX2	Goat	Mexico	GCA_009759765.1	2,33809
C. pseudotuberculosis PAT16	Ovine	Argentina	GCA_009759705.1	2,33815
C. pseudotuberculosis SP165	Ovine	Brazil	GCA_009759725.1	2,33779
C. pseudotuberculosis SigB	Goat	Brazil	GCA_009762635.1	2,33759

C. pseudotuberculosis SigH	Goat	Brazil	GCA_009762655.1	2,33762
C. pseudotuberculosis SigK	Goat	Brazil	GCA_009762615.1	2,33833
C. pseudotuberculosis 133	---	Brazil	GCF_009789115.1	2,33825
C. pseudotuberculosis 414	Goat	Brazil	GCF_009789135.1	2,338,02
C. pseudotuberculosis 206	Ovine	Brazil	GCF_009791415.1	2,33802
C. pseudotuberculosis SigM	Goat	Brazil	GCF_009791515.1	2,33922
C. pseudotuberculosis PAT14	Ovine	Argentina	GCA_009905275.1	2,33825
C. pseudotuberculosis SigD	Goat	Brazil	GCA_009905175.1	2,33760
C. pseudotuberculosis SigC	Goat	Brazil	GCA_009930775.1	2,33764

* Strains used in the first pan-genome (SOARES, SC *et al.*, 2013)

Table 2 - Strains used in the first *C. pseudotuberculosis* pan-genome study.

Nº	Strains	Genome size	Genes	*Singletons*	Old GenBank	New GenBank
1	1002	2,335,113	2,203	0	CP001809.2	CP001809.3
2	C231	2,328,208	2,204	3	CP001829.1	CP001829.2
3	42/02-A	2,337,606	2,164	5	CP003062.1	-
4	PAT10	2,335,323	2,200	1	CP002924.1	-
5	3/99-5	2,337,938	2,239	39	CP003152.1	-
6	267	2,337,628	2,249	8	CP003407.1	-
7	P54B96	2,337,657	2,205	2	CP003385.1	-
8	I19	2,337,730	2,213	0	CP002251.1	CP002251.3
9	FRC41	2,337,913	2,171	12	CP002097.1	CP002097.2
10	CIP52.97	2,320,595	2,194	30	CP003061.1	CP003061.3
11	316	2,310,415	2,234	25	CP003077.1	CP003077.2
12	258	2,314,404	2,195	29	CP003540.1	CP003540.3
13	1/06-A	2,279,118	2,127	20	CP003082.1	-
14	Cp162	2,293,464	2,150	13	CP003652.1	CP003652.3
15	31	2,297,010	2,170	50	CP003421.1	CP003421.4

6. CONCLUSÕES

O sequenciamento por plataformas, tais como Ion Torrent PGMTM HI-Q 400 pb Fragment e Illumina Hiseq 2500 450 pb *paired*-end, em conjunto, forneceu leituras de alta qualidade e elevada cobertura genômica. Isso facilitou substancialmente o processo de geração e montagem dos *contigs*, permitindo a obtenção de versões completas e mais refinadas de cada genoma. Desta forma, a combinação de diferentes tecnologias de sequenciamento aplicada neste trabalho possibilitou a geração de genomas completos de *C. pseudotuberculosis* que apresentam maior acurácia em comparação aos genomas anteriormente disponíveis para essa espécie.

A estratégia do mapa óptico mostrou-se altamente eficaz no processo de ordenação dos contigs. Com o genoma completo, foi possível identificar erros em montagens anteriores depositadas, bem como detectar inversões gênicas específicas de cada linhagem. Observou-se que as maiores inversões cromossômicas são consequência da presença de clusters de genes de *rRNAs*, cuja alta similaridade dentro do mesmo organismo favorece eventos de reordenação e modifica a sintenia entre linhagens da mesma espécie.

Por fim, conseguimos encontrar vários genes que não haviam sido representados nas primeiras versões depositadas. Destaque para a linhagem *C. pseudotuberculosis* 162 (biovar *equi*), cujo conteúdo gênico aumentou em cerca de 70 mil pares de bases. Considerando outras linhagens também do biovar *equi*, validou-se por PCR a presença do cluster *nar*, principal marcador diferencial deste *biovar*, em genomas que deveriam conter esse cluster, mas que foram anteriormente depositados sem ele.

O gene predito como *cutinase like family* em C. pseudotuberculosis apresenta atividade enzimática voltada para a degradação de ácidos graxos de cadeia curta ou média, não possuindo, portanto, função de degradação de cutina, como sugere a anotação automática incorreta. O mesmo gene está ausente em *C. pseudotuberculosis* 1002 e em *C. pseudotuberculosis* PAT10.

A identificação de SNPs dentro do biovar *ovis* contribuiu para destacar mutações pontuais nucleotídicas em genes específicos de determinadas linhagens. Algumas dessas variantes são não sinônimas, como é o caso de linhagens isoladas da mesma região que compartilham tais variantes nucleotídicas.

7. PERSPECTIVAS

a) Realizar uma nova análise de pan-genoma com todas as 115 linhagens completas de *C. pseudotuberculosis*, considerando as adições recentes de genomas completos. Além disso, o maior número de sequenciamentos de alta qualidade e novas versões genômicas disponíveis proporciona uma base mais robusta para análise;
b) Propor novos alvos vacinais e terapêuticos por meio da vacinologia reversa e ancoragem molecular, utilizando todas as linhagens de *C. pseudotuberculosis* atualmente depositadas, uma vez que as novas montagens genômicas revelaram genes e proteínas anteriormente não identificados.

8. REFERÊNCIAS

ALMEIDA, Sintia; LOUREIRO, Dan; *et al.* Complete Genome Sequence of the Attenuated Corynebacterium pseudotuberculosis Strain T1. *Genome Announcements*, v. 4, n. 5, p. e00947-16, 2016. Disponível em: <http://jb.asm.org/cgi/doi/10.1128/JB.06479-11%5Cnhttp://genomea.asm.org/lookup/doi/10.1128/genomeA.00947-16>.

ALMEIDA, Sintia *et al.* Exploration of Nitrate Reductase Metabolic Pathway in *Corynebacterium pseudotuberculosis*. *International Journal of Genomics*, v. 2017, p. 1–12, 2017. Disponível em: <https://www.hindawi.com/journals/ijg/2017/9481756/>.

ALMEIDA, Sintia; TIWARI, Sandeep; *et al.* The genome anatomy of Corynebacterium pseudotuberculosis VD57 a highly virulent strain causing Caseous lymphadenitis. *Standards in Genomic Sciences*, v. 11, n. 1, p. 29, 8 dez. 2016. Disponível em: <http://dx.doi.org/10.1186/s40793-016-0149-7>.

ALVES, F.S.F.; PINHEIRO, R.R.; PIRES, P.C. Linfadenite caseosa: patogenia, diagnóstico, controle. *Sobral:EMBRAPA-CNPC*, v. Doc. 27, p. 16, 1997.

AMERICA, North; AMHERST, Sir Jeffrey; JENNER, Edward. Edward Jenner and the History of Vaccines. p. 21–25, 2005.

ARAÚJO, Carlos Leonardo *et al.* Prediction of new vaccine targets in the core genome of Corynebacterium pseudotuberculosis through omics approaches and reverse vaccinology. *Gene*, v. 702, n. December 2018, p. 36–45, 2019. Disponível em: <https://doi.org/10.1016/j.gene.2019.03.049>.

ASTON, Christopher; MISHRA, Bud; SCHWARTZ, David C. Optical mapping and its potential for large-scale sequencing projects. *Trends in biotechnology*, v. 17, n. 7, p. 297–302, jul. 1999. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/10370237>.

AYLING, Martin; CLARK, Matthew D; LEGGETT, Richard M. New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics*, v. 00, n. February, p. 1–11, 28 fev. 2019. Disponível em: <https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz020/5363831>.

BAKER, Monya. De novo genome assembly: what every biologist should know. *Nature Methods*, v. 9, n. 4, p. 333–337, 27 abr. 2012. Disponível em: <http://www.nature.com/articles/nmeth.1935>.

BANKEVICH, Anton *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, v. 19, n. 5, p. 455–477, maio 2012. Disponível em: <http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0021>. Acesso em: 19 abr. 2016.

BARAÚNA, Rafael A. *et al.* Assessing the Genotypic Differences between Strains of Corynebacterium pseudotuberculosis biovar equi through Comparative Genomics. *PLOS ONE*, v. 12, n. 1, p. e0170676, 26 jan. 2017. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0170676>.

BARNABÉ, Nathanael Natércio da Costa *et al.* Characterization of caseous lymphadenitis in caprine animals slaughtered in a semi-arid region of Brazil. *Semina: Ciências Agrárias*, v. 40,

n. 5, p. 1867, 2019.

BASHIR, Ali *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nature biotechnology*, v. 30, n. 7, p. 701–7, jul. 2012. Disponível em: <http://www.nature.com/nbt/journal/v30/n7/full/nbt.2288.html?WT.ec_id=NBT-201207>.

BELBAHRI, Lassaad *et al.* Evolution of the cutinase gene family: Evidence for lateral gene transfer of a candidate Phytophthora virulence factor. *Gene*, v. 408, n. 1–2, p. 1–8, 2008.

BELLANGER, Xavier *et al.* Conjugative and mobilizable genomic islands in bacteria: Evolution and diversity. *FEMS Microbiology Reviews*, v. 38, n. 4, p. 720–760, 2014.

BRETTIN, Thomas *et al.* RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, v. 5, 2015.

BROOKES, Anthony J. The essence of SNPs. *Gene*, v. 234, n. 2, p. 177–186, jul. 1999. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S037811199900219X>.

BRYSON, Kevin *et al.* Protein structure prediction servers at University College London. *Nucleic acids research*, v. 33, n. Web Server issue, p. W36-8, 1 jul. 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15980489>.

CHAUDHARI, Narendrakumar M.; GUPTA, Vinod Kumar; DUTTA, Chitra. BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports*, v. 6, n. 1, p. 24373, 13 abr. 2016. Disponível em: <http://dx.doi.org/10.1038/srep24373>.

CHEN, Sheng *et al.* Identification and characterization of bacterial cutinase. *Journal of Biological Chemistry*, v. 283, n. 38, p. 25854–25862, 2008.

CHMIELA, Magdalena; KUPCINSKAS, Juozas. Review: Pathogenesis of Helicobacter pylori infection. *Helicobacter*, v. 24, n. S1, p. 1–5, 2019.

COELHO, Keila da Silva. Isolamento, clonagem e caracterização molecular do gene hsp60 de Corynebacterium pseudotuberculosis e sua utilização na construção de uma vacina de subunidade protéica e de DNA. *Dissertação (Mestrado em Genética) – Universidade Federal de Minas Gerais, BH, MG*, p. 1–165, 2007.

CONTZEN, M. *et al.* Corynebacterium ulcerans from Diseased Wild Boars. *Zoonoses and Public Health*, v. 58, n. 7, p. 479–488, 2011.

CRAIG, N. L. Tn7: a target site-specific transposon. *Molecular Microbiology*, v. 5, n. 11, p. 2569–2573, 1991.

CRISTINA, Ana *et al.* Whole-genome mapping reveals a large chromosomal inversion on Iberian Brucella suis biovar 2 strains. *Veterinary Microbiology*, v. 192, p. 220–225, 2016. Disponível em: <http://dx.doi.org/10.1016/j.vetmic.2016.07.024>.

DARLING, Aaron E.; MAU, Bob; PERNA, Nicole T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE*, v. 5, n. 6, p. e11147, 25 jun. 2010. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20593022>.

DARMON, E.; LEACH, D. R. F. Bacterial Genome Instability. *Microbiology and Molecular Biology Reviews*, v. 78, n. 1, p. 1–39, 2014.

DE BARSY, Marie *et al.* Regulatory (pan-)genome of an obligate intracellular pathogen in the PVC superphylum. *The ISME Journal*, v. 10, n. 9, p. 2129–2144, 8 set. 2016. Disponível em: <http://www.nature.com/articles/ismej201623>.

DE SÁ, Pablo H. C. G. *et al.* GapBlaster—A Graphical Gap Filler for Prokaryote Genomes. *PLOS ONE*, v. 11, n. 5, p. e0155327, 12 maio 2016. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0155327>.

DIGUISTINI, Scott *et al.* De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology*, v. 10, n. 9, p. R94, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19747388>.

DONG, Yang *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nature biotechnology*, v. 31, n. 2, p. 135–41, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23263233>.

DORELLA, F. A. *et al.* Construction and partial characterization of a Corynebacterium pseudotuberculosis bacterial artificial chromosome library through genomic survey sequencing. *Genetics and Molecular Research*, v. 5, n. 4, p. 653–663, 2006.

DORELLA, Fernanda A. *et al.* Antigens of Corynebacterium pseudotuberculosis and prospects for vaccine development. *Expert Review of Vaccines*, v. 8, n. 2, p. 205–213, 2009.

DORELLA, Fernanda Alves *et al.* Corynebacterium pseudotuberculosis : microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Veterinary Research*, mar. 2006. , v. 37, n. 2, p. 201–218Disponível em: <http://www.edpsciences.org/10.1051/vetres:2005056>. Acesso em: 14 maio 2014.

DRAGANOVA, Elizabeth B. *et al.* Heme Binding by Corynebacterium diphtheriae HmuT: Function and Heme Environment. *Biochemistry*, v. 54, n. 43, p. 6598–6609, 3 nov. 2015. Disponível em: <file:///C:/Users/Carla Carolina/Desktop/Artigos para acrescentar na qualificação/The impact of birth weight on cardiovascular disease risk in the.pdf>.

EDWARDS, David J; HOLT, Kathryn E. Comparative bacterial genome analysis. *Bioinformatics*, v. 25, n. 11, p. 1422–1423, 2009. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp163>.

EID, John *et al.* Real-time DNA sequencing from single polymerase molecules. *Science*, v. 323, n. 5910, p. 133–138, 2009.

EL SHAFEY, H. M.; GHANEM, S. Regulation of expression of sodA and msrA genes of corynebacterium glutamicum in response to oxidative and radiative stress. *Genetics and Molecular Research*, v. 14, n. 1, p. 2104–2117, 2015.

EMMS, David M.; KELLY, Steven. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, v. 16, n. 1, p. 1–14, 2015. Disponível em: <http://dx.doi.org/10.1186/s13059-015-0721-2>.

EWING, Brent *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research*, v. 8, n. 3, p. 175–85, 1 mar. 1998. Disponível em: <http://genome.cshlp.org/content/8/3/175.short%5Cnpapers3://publication/uuid/C78B978E-A9C6-4B0E-8666-71E2DA2334F3>.

FARIAS, J. L.de S. *et al.* Análise socioeconômica de produtores familiares de caprinos e ovinos no semiárido cearense, Brasil. *Archivos de Zootecnia*, v. 63, n. 241, p. 13–24, 2014.

FLEISCHMANN, Robert D *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, v. 269, n. 5223, p. 496–512, 28 jul. 1995. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/7542800>.

FONTAINE, M. C.; BAIRD, G. J. Caseous lymphadenitis. *Small Ruminant Research*, v. 76, n. 1–2, p. 42–48, 14 abr. 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/2336781>.

FONTAINE, Michael C *et al.* Vaccination confers significant protection of sheep against infection with a virulent United Kingdom strain of Corynebacterium pseudotuberculosis. *Vaccine*, v. 24, n. 33–34, p. 5986–96, 14 ago. 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16806606>. Acesso em: 15 ago. 2013.

FOUTS, Derrick E. *et al.* PanOCT: Automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Research*, v. 40, n. 22, p. 1–11, 2012.

GALARDINI, Marco *et al.* CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source code for biology and medicine*, v. 6, n. 1, p. 11, 2011. Disponível em: <http://www.scfbm.org/content/6/1/11>.

GHURYE, Jay; POP, Mihai. Modern technologies and algorithms for scaffolding assembled genomes. *PLOS Computational Biology*, v. 15, n. 6, p. e1006994, 5 jun. 2019. Disponível em: <http://dx.plos.org/10.1371/journal.pcbi.1006994>.

GUEDES, Maria T. *et al.* Infecção por Corynebacterium pseudotuberculosis em equinos: aspectos microbiológicos, clínicos e preventivos. *Pesquisa Veterinária Brasileira*, v. 35, n. 8, p. 701–708, ago. 2015. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-736X2015000800701&lng=pt&nrm=iso&tlng=en>.

GUIMARÃES, Luis Carlos *et al.* Inside the Pan-genome - Methods and Software Overview. *Current genomics*, v. 16, n. 4, p. 245–52, ago. 2015. Disponível em: <http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1389-2029&volume=16&issue=4&spage=245>.

GUIZELINI, Dieval *et al.* GFinisher: A new strategy to refine and finish bacterial genome assemblies. *Scientific Reports*, v. 6, n. October, p. 1–8, 2016. Disponível em: <http://dx.doi.org/10.1038/srep34963>.

GUREVICH, Alexey *et al.* QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, v. 29, n. 8, p. 1072–1075, 15 abr. 2013. Disponível em: <http://bioinf.spbau.ru/quast>. Acesso em: 17 ago. 2018.

HAAS, Dionei J. *et al.* Molecular epidemiology of Corynebacterium pseudotuberculosis

isolated from horses in California. *Infection, Genetics and Evolution*, v. 49, p. 186–194, 2017. Disponível em: <http://dx.doi.org/10.1016/j.meegid.2016.12.011>.

HASSAN, Syed Shah *et al.* Whole-genome sequence of Corynebacterium pseudotuberculosis strain Cp162, isolated from camel. *Journal of Bacteriology*, v. 194, n. 20, p. 5718–5719, 15 out. 2012. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23012291>.

HEMMERICH, Johannes *et al.* Combinatorial impact of Sec signal peptides from Bacillus subtilis and bioprocess conditions on heterologous cutinase secretion by Corynebacterium glutamicum. *Biotechnology and Bioengineering*, v. 116, n. 3, p. 644–655, 31 mar. 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.26873>.

HERNANDEZ, David *et al.* De novo finished 2.8 Mbp Staphylococcus aureus genome assembly from 100 bp short and long range paired-end reads. *Bioinformatics*, v. 30, n. 1, p. 40–49, 2014.

HOU, Xiao Gang *et al.* Genetic identification of members of the genus Corynebacterium at genus and species levels with 16s rDNA-targeted probes. *Microbiology and Immunology*, v. 41, n. 6, p. 453–460, jun. 1997. Disponível em: <http://doi.wiley.com/10.1111/j.1348-0421.1997.tb01878.x>.

HOU, Ya Ming. *Transfer RNAs and pathogenicity islands*. *Trends in Biochemical Sciences*. [S.l: s.n.]. , 1999

IBRAIM, Izabela Coimbra *et al.* Transcriptome profile of Corynebacterium pseudotuberculosis in response to iron limitation. *BMC Genomics*, v. 20, n. 1, p. 663, 20 dez. 2019. Disponível em: <https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6018-1>.

JING, J *et al.* Optical mapping of Plasmodium falciparum chromosome 2. *Genome research*, v. 9, n. 2, p. 175–81, fev. 1999. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=310721&tool=pmcentrez&rendertype=abstract>.

KAPUSTOVÁ, Veronika *et al.* The dark matter of large cereal genomes: Long tandem repeats. *International Journal of Molecular Sciences*, v. 20, n. 10, 2 maio 2019.

KOREN, Sergey *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, v. 30, n. 7, p. 693–700, 1 jul. 2012. Disponível em: <http://www.nature.com/doifinder/10.1038/nbt.2280>.

KOTEWICZ, Michael 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. Disponível em: <https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.2008/019026-0>.

KOTEWICZ, Michael L. *et al.* Optical maps distinguish individual strains of Escherichia coli O157 : H7. *Microbiology*, v. 153, n. 6, p. 1720–1733, 1 jun. 2007. Disponível em: <http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.2006/004507-0>.

KRAWIEC, S; RILEY, M. Organization of the bacterial chromosome. *Microbiological reviews*, v. 54, n. 4, p. 502–39, dez. 1990. Disponível em:

<http://www.ncbi.nlm.nih.gov/pubmed/2087223>.

LANGMEAD. Bowtie2. *Nature methods*, v. 9, n. 4, p. 357–359, 2013.

LEAL, Karen Silva *et al.* Recombinant M. bovis BCG expressing the PLD protein promotes survival in mice challenged with a C. pseudotuberculosis virulent strain. *Vaccine*, v. 36, n. 25, p. 3578–3583, 2018. Disponível em: <https://doi.org/10.1016/j.vaccine.2018.05.049>.

LEE, Hayan *et al.* Third-generation sequencing and the future of genomics. *bioRxiv*, n. Table 1, p. 048603, 2016. Disponível em: <http://biorxiv.org/content/early/2016/04/13/048603.abstract>.

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. January, p. 43149, 2017. Disponível em: <http://www.nature.com/articles/srep43149>.

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. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324>.

LI, Wei *et al.* RmlB and rmlC genes are essential for growth of mycobacteria. *Biochemical and Biophysical Research Communications*, v. 342, n. 1, p. 170–178, 2006.

LIN, Jieyi. Whole-Genome Shotgun Optical Mapping of Deinococcus radiodurans. *Science*, v. 285, n. 5433, p. 1558–1562, 3 set. 1999. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.285.5433.1558>. Acesso em: 26 set. 2019.

LING, Jielu *et al.* Aerobactin Synthesis Genes iucA and iucC Contribute to the Pathogenicity of Avian Pathogenic Escherichia coli O2 Strain E058. *PLoS ONE*, v. 8, n. 2, 2013.

MAGALHÃES, Klinger Aragão *et al. Pesquisa Pecuária Municipal 2017 : efetivo dos rebanhos caprinos e ovinos*. *Boletim do Centro de Inteligência e Mercado de Caprinos e Ovinos*. [S.l: s.n.], 2018. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/185392/1/CNPC-2018-BCIMn52018.pdf>.

MAHAIRAS, G G *et al.* Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis. *Journal of bacteriology*, v. 178, n. 5, p. 1274–82, mar. 1996. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/8631702>.

MARGULIES, Marcel *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, v. 437, n. 7057, p. 376–80, 31 jul. 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16056220>.

MAXAM, A M; GILBERT, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, v. 74, n. 2, p. 560–4, 1977. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/265521%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC392330>.

MEDINI, Duccio *et al.* The microbial pan-genome. *Current Opinion in Genetics & Development*, v. 15, n. 6, p. 589–594, dez. 2005. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0959437X05001759>.

MENZIES, P. I.; HWANG, Y. T.; PRESCOTT, J. F. Comparison of an interferon-γ to a phospholipase D enzyme-linked immunosorbent assay for diagnosis of Corynebacterium pseudotuberculosis infection in experimentally infected goats. *Veterinary Microbiology*, v. 100, n. 1–2, p. 129–137, 2004.

MERHEJ, Vicky *et al.* Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biology Direct*, v. 4, p. 1–25, 2009.

METZKER, Michael L. Sequencing technologies — the next generation. *Nature Reviews Genetics*, v. 11, n. 1, p. 31–46, 8 jan. 2010. Disponível em: <http://www.nature.com/articles/nrg2626>.

MEYER, Roberto *et al.* Avaliação da resposta imune humoral em caprinos inoculados com uma vacina viva atenuada liofilizada contra Corynebacterium pseudotubercusis. *REVISTA DE CIÊNCIAS MÉDICAS E BIOLÓGICAS*, v. 1, n. 1, p. 42–48, 2002.

MILLER, Jason R.; KOREN, Sergey; SUTTON, Granger. Assembly algorithms for next-generation sequencing data. *Genomics*, v. 95, n. 6, p. 315–327, 2010. Disponível em: <http://dx.doi.org/10.1016/j.ygeno.2010.03.001>.

MOORE, R. *et al.* Corynebacterium and Arcanobacterium. *Pathogenesis of Bacterial Infections in Animals: Fourth Edition*, n. April 2016, p. 133–147, 12 maio 2010. Disponível em: <http://doi.wiley.com/10.1002/9780470958209.ch8>. Acesso em: 14 abr. 2016.

MOURA-COSTA, L F *et al.* Evaluation of the humoral and cellular immune response to different antigens of Corynebacterium pseudotuberculosis in Canindé goats and their potential protection against caseous lymphadenitis. *Veterinary immunology and immunopathology*, v. 126, n. 1–2, p. 131–41, 15 nov. 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18752855>. Acesso em: 17 set. 2013.

NAGARAJAN, Niranjan; POP, Mihai. Sequence assembly demystified. *Nature Reviews Genetics*, v. 14, n. 3, p. 157–167, 29 jan. 2013. Disponível em: <http://dx.doi.org/10.1038/nrg3367>.

OLSEN, Remi-andre Andre *et al.* De novo assembly of Dekkera bruxellensis: a multi technology approach using short and long-read sequencing and optical mapping. *GigaScience*, v. 4, n. 1, p. 56, 26 dez. 2015. Disponível em: <http://dx.doi.org/10.1186/s13742-015-0094-1>.

ONMUS-LEONE, Fatma *et al.* Enhanced De Novo Assembly of High Throughput Pyrosequencing Data Using Whole Genome Mapping. *PLoS ONE*, v. 8, n. 4, p. 2–10, 2013.

OTT, Lisa. Adhesion properties of toxigenic corynebacteria. *AIMS Microbiology*, v. 4, n. 1, p. 85–103, 2018. Disponível em: <http://www.aimspress.com/article/10.3934/microbiol.2018.1.85>.

OXFORD NANOPORE TECHNOLOGIES. *Oxford Nanopore: How it works*. Disponível em: <https://nanoporetech.com/how-it-works>.

PACHECO, Luis G.C. C *et al.* Multiplex PCR assay for identification of Corynebacterium pseudotuberculosis from pure cultures and for rapid detection of this pathogen in clinical samples. *Journal of Medical Microbiology*, v. 56, n. 4, p. 480–486, abr. 2007.

PAGE, Andrew J. *et al.* Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, v. 31, n. 22, p. 3691–3693, 2015.

PANTOJA, Yan *et al.* PanWeb: A web interface for pan-genomic analysis. *PLoS ONE*, v. 12, n. 5, p. 1–9, 2017.

PARISE, Doglas *et al.* First genome sequencing and comparative analyses of Corynebacterium pseudotuberculosis strains from Mexico. *Standards in Genomic Sciences*, v. 13, n. 1, p. 1–10, 2018.

PAULIEN HOGEWEG. The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology*, v. 7, n. 3, p. e1002021, 31 mar. 2011. Disponível em: <https://dx.plos.org/10.1371/journal.pcbi.1002021>.

PEREIRA, Felipe L. *et al.* Evaluating the efficacy of the new Ion PGM Hi-Q Sequencing Kit applied to bacterial genomes. *Genomics*, v. 107, n. 5, p. 189–198, mar. 2016. Disponível em: <http://dx.doi.org/10.1016/j.ygeno.2016.03.004>.

PETERSEN, Randi Føns *et al.* Molecular Characterization of Salmonella Typhimurium Highly Successful Outbreak Strains. *Foodborne Pathogens and Disease*, v. 8, n. 6, p. 655–661, jun. 2011. Disponível em: <http://www.liebertpub.com/doi/10.1089/fpd.2010.0683>. Acesso em: 18 maio 2016.

POP, Mihai. Genome assembly reborn: Recent computational challenges. *Briefings in Bioinformatics*, v. 10, n. 4, p. 354–366, 1 jul. 2009. Disponível em: <http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbp026>.

PRAY, Leslie. *Discovery of DNA Structure and Function: Watson and Crick*.

PROSDÓCIMI, Francisco; MOREIRA, Leandro Marcio. Genômica comparativa. *Ciências genômicas: fundamentos e aplicações*. [S.l: s.n.], 2015. p. 81–99.

QUAIL, Michael *et al.* A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, v. 13, n. 1, p. 1, 2012. Disponível em: <http://dx.doi.org/10.1186/1471-2164-13-341>.

RAMOS, Rommel Tj *et al.* Analysis of quality raw data of second generation sequencers with Quality Assessment Software. *BMC Research Notes*, v. 4, n. 1, p. 130, 2011. Disponível em: <http://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-4-130>.

RAPPUOLI, R. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine*, v. 19, n. 17–19, p. 2688–91, 21 mar. 2001. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11257410>.

RAPPUOLI, Rino. *Bridging the knowledge gaps in vaccine design*. *Nature Biotechnology*. [S.l: s.n.]. Disponível em: <http://www.nature.com/naturebiotechnology>. Acesso em: 22 ago. 2019. , 2007

RAPPUOLI, Rino *et al.* Meningococcal B vaccine (4CMenB): the journey from research to real world experience. *Expert Review of Vaccines*, v. 17, n. 12, p. 1111–1121, 2018. Disponível em: <https://doi.org/10.1080/14760584.2018.1547637>.

RAPPUOLI, Rino *et al.* Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. *The Journal of Experimental Medicine*, v. 213, n. 4, p. 469–481, 4 abr. 2016. Disponível em: <www.jem.org/cgi/doi/10.1084/jem.20151960469>. Acesso em: 19 ago. 2019.

ROULI, L. *et al.* The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, v. 7, p. 72–85, 2015. Disponível em: <http://dx.doi.org/10.1016/j.nmni.2015.06.005>.

RUIZ, Jerônimo C. *et al.* Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two Corynebacterium pseudotuberculosis Strains. *PLoS ONE*, v. 6, n. 4, p. e18551, 18 abr. 2011. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0018551>. Acesso em: 6 maio 2014.

SAHA, Subrata; RAJASEKARAN, Sanguthevar. Efficient and scalable scaffolding using optical restriction maps. *BMC Genomics*, v. 15, n. 5, p. S5, 2014. Disponível em: <http://www.biomedcentral.com/1471-2164/15/S5/S5>.

SAMAD, A H *et al.* Mapping the genome one molecule at a time--optical mapping. *Nature*, v. 378, n. 6556, p. 516–7, 30 nov. 1995. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/7477412>.

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, 1977. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5463>.

SCHWARTZ, D. C *et al.* Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science (New York, N.Y.)*, v. 262, n. 5130, p. 110–4, 1 out. 1993. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.8211116>. Acesso em: 18 maio 2016.

SEEMANN, Torsten. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, v. 30, n. 14, p. 2068–2069, 15 jul. 2014. Disponível em: <http://bioinformatics.oxfordjournals.org/content/30/14/2068.full>.

SETTE, Alessandro; RAPPUOLI, Rino. Reverse Vaccinology: Developing Vaccines in the Era of Genomics. *Immunity*, v. 33, n. 4, p. 530–541, out. 2010. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1074761310003602>.

SEYFFERT, N. *et al.* High seroprevalence of caseous lymphadenitis in Brazilian goat herds revealed by Corynebacterium pseudotuberculosis secreted proteins-based ELISA. *Research in Veterinary Science*, v. 88, n. 1, p. 50–55, 2010. Disponível em: <http://dx.doi.org/10.1016/j.rvsc.2009.07.002>.

SHUKLA, Sanjay K. *et al.* Comparative whole-genome mapping to determine Staphylococcus aureus genome size, virulence motifs, and clonality. *Journal of Clinical Microbiology*, v. 50, n. 11, p. 3526–3533, 2012.

SIMS, David *et al.* Sequencing depth and coverage: key considerations in genomic analyses.

Nature Reviews Genetics, v. 15, n. 2, p. 121–32, 2014. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24434847>.

SIVASHANKARI, Selvarajan; SHANMUGHAVEL, Piramanayagam. Comparative genomics - a perspective. *Bioinformation*, v. 1, n. 9, p. 376–8, 2007. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17597925%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1891719>.

SOARES, SC *et al.* The Pan-Genome of the Animal Pathogen Corynebacterium pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains. *PLoS ONE*, v. 8, n. 1, p. e53818, 14 jan. 2013. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0053818>. Acesso em: 2 jul. 2014.

SOARES, Siomar C. *et al.* Genome sequence of Corynebacterium pseudotuberculosis biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. *Journal of Biotechnology*, v. 167, n. 2, p. 135–141, 2013. Disponível em: <http://dx.doi.org/10.1016/j.jbiotec.2012.11.003>.

SOARES, Siomar de Castro. *Pan-genomic analyses of Corynebacterium pseudotuberculosis and characterization of the biovars ovis and equi through comparative genomics.* 2013. 106 f. Universidade Federal de Minas Gerais, 2013.

SONGER, J G. Bacterial phospholipases and their role in virulence. *Trends in microbiology*, v. 5, n. 4, p. 156–61, abr. 1997. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/9141190>. Acesso em: 13 ago. 2013.

SOUCY, Shannon M.; HUANG, Jinling; GOGARTEN, Johann Peter. Horizontal gene transfer: Building the web of life. *Nature Reviews Genetics*, v. 16, n. 8, p. 472–482, 2015. Disponível em: <http://dx.doi.org/10.1038/nrg3962>.

SOUSA, Thiago De Jesus *et al.* Re-sequencing and optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes. *Scientific Reports*, v. 9, n. 1, p. 16387, 8 dez. 2019. Disponível em: <http://www.nature.com/articles/s41598-019-52695-4>.

SPIER, S. J. *et al.* Short communications: Survival of Corynebacterium pseudotuberculosis biovar equi in soil. *Veterinary Record*, v. 170, n. 7, p. 0–1, 18 fev. 2012. Disponível em: <http://veterinaryrecord.bmj.com/lookup/doi/10.1136/vr.100543>.

TACHEDJIAN, Mary *et al.* Caseous lymphadenitis vaccine development: site-specific inactivation of the Corynebacterium pseudotuberculosis phospholipase D gene. *Vaccine*, v. 13, n. 18, p. 1785–1792, 1995.

TATUSOVA, Tatiana *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, v. 44, n. 14, p. 6614–6624, 2016.

TETTELIN *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. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.0506758102>.

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

THOMSEN, René; CHRISTENSEN, Mikael H. MolDock: A New Technique for High-Accuracy Molecular Docking. *Journal of Medicinal Chemistry*, v. 49, n. 11, p. 3315–3321, jun. 2006. Disponível em: <https://pubs.acs.org/doi/10.1021/jm051197e>.

TREANGEN, Todd J *et al.* The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, v. 15, n. 11, p. 524, 19 nov. 2014. Disponível em: <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0524-x>.

TRIPATHI, Deeksha *et al.* Low expression level of glnA1 accounts for absence of cell wall associated poly-L-glutamate/glutamine in Mycobacterium smegmatis. *Biochemical and Biophysical Research Communications*, v. 458, n. 2, p. 240–245, 2015. Disponível em: <http://dx.doi.org/10.1016/j.bbrc.2015.01.079>.

TROST, Eva *et al.* The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC Genomics*, v. 11, n. 1, p. 728, jan. 2010. Disponível em: <http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-728>.

VAN DIJK, Erwin L. *et al.* Ten years of next-generation sequencing technology. *Trends in Genetics*, v. 30, n. 9, p. 418–426, set. 2014. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0168952514001127>.

VIANA, Marcus Vinicius Canário *et al.* Comparative genomic analysis between Corynebacterium pseudotuberculosis strains isolated from buffalo. *PLoS ONE*, v. 12, n. 4, p. e0176347, 26 abr. 2017. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0176347>.

VILELA RODRIGUES, Thaís Cristina *et al.* Reverse vaccinology and subtractive genomics reveal new therapeutic targets against Mycoplasma pneumoniae : a causative agent of pneumonia . *Royal Society Open Science*, v. 6, n. 7, p. 190907, 2019.

VOELKERDING, Karl V.; DAMES, Shale A.; DURTSCHI, Jacob D. Next-generation sequencing:from basic research to diagnostics. *Clinical Chemistry*, v. 55, n. 4, p. 641–658, 2009.

WALKER, J. *et al.* Identification of a novel antigen from Corynebacterium pseudotuberculosis that protects sheep against caseous lymphadenitis. *Infection and Immunity*, v. 62, n. 6, p. 2562–2567, 1994.

WANG, Guo Dong *et al.* Dog10K: The International Consortium of Canine Genome Sequencing. *National Science Review*, v. 6, n. 4, p. 611–613, 1 jul. 2019.

WEISSENSTEINER, Matthias H. *et al.* Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Research*, v. 27, n. 5, p. 697–708, maio 2017. Disponível em: <http://genome.cshlp.org/lookup/doi/10.1101/gr.215095.116>.

WEST, Nicholas P. *et al.* Cutinase-like proteins of Mycobacterium tuberculosis : characterization of their variable enzymatic functions and active site identification . *The FASEB Journal*, v. 23, n. 6, p. 1694–1704, 2009a.

WEST, Nicholas P *et al.* Cutinase-like proteins of Mycobacterium tuberculosis: characterization of their variable enzymatic functions and active site identification. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, v. 23, n. 6, p. 1694–704, jun. 2009b. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19225166>.

WINDSOR, P. A.; BUSH, R. D. Caseous lymphadenitis: Present and near forgotten from persistent vaccination? *Small Ruminant Research*, v. 142, p. 6–10, 2016. Disponível em: <http://dx.doi.org/10.1016/j.smallrumres.2016.03.023>.

WU, Chia-wei *et al.* Optical mapping of the Mycobacterium avium subspecies paratuberculosis genome. *BMC genomics*, v. 10, p. 25, 2009.

XAVIER, Basil Britto *et al.* Employing whole genome mapping for optimal de novo assembly of bacterial genomes. *BMC Research Notes*, v. 7, n. 1, p. 484, 2014. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4118782&tool=pmcentrez&rendertype=abstract>.

ZHAO, Yongbing *et al.* PGAP: Pan-genomes analysis pipeline. *Bioinformatics*, v. 28, n. 3, p. 416–418, 2012.

ZHOU, Shiguo *et al.* Validation of rice genome sequence by optical mapping. *BMC Genomics*, v. 8, n. 1, p. 278, 7 nov. 2007. Disponível em: <http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-8-278>.

9. APÊNDICES

Nesta seção estão descritas as demais atividades e trabalhos com o intuito de complementar minha formação acadêmica.

- Apêndice A – Tabela com todas as linhagens de *C. pseudotuberculosis* já depositadas no banco de dados do NCBI.
- Apêndice B – Pipeline para fechamento de *gaps* em genomas bacterianos.
- Apêndice C – Artigo: *Identification of membrane- associated proteins with pathogenic potential expressed by Corynebacterium pseudotuberculosis grow in animal serum*.
- Apêndice D – Artigo: *SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology*.
- Apêndice E – Artigo: *Complete Genome Sequence of the Attenuated Corynebacterium pseudotuberculosis T1*.
- Apêndice F - Artigo: *Cell wall glycolipids from Corynebacterium pseudotuberculosis strains with different virulences differ in terms of composition and immune recognition.*
- Apêndice G - Atividade presentes no currículo lattes - 2016 a 2020.

Apêndice A – Tabela com todas as linhagens de *C. pseudotuberculosis* já depositadas no NCBI.

Tabela 2 - Informações de todas as linhagens de *C. pseudotuberculosis* já depositadas no banco do *GenBank* no NCBI.

(continua)

Linhagem	Biovar	Hospedeiro	Pais	Plataforma de sequenciamento	Número de acesso
C. pseudotuberculosis C231	*ovis*	Ovino	Austrália	Illumina HiSeq *paired-end*	GCA_000144675.2
C. pseudotuberculosis 1002	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_000144935.3
C. pseudotuberculosis I19	*ovis*	Bovino	Israel	Illumina HiSeq *paired-end*	GCA_000152065.3
C. pseudotuberculosis 42/02-A	*ovis*	Ovino	Austrália	454 GS-FLX e Solexa 50-bp *paired-end*	GCA_000227175.1
C. pseudotuberculosis 1/06-A	*equi*	Cavalo	EUA	454 GS-FLX e Solexa 50-bp *paired-end*	GCA_000233735.1
C. pseudotuberculosis CIP 52.97	*equi*	Cavalo	Quênia	Illumina HiSeq *paired-end*	GCA_000227605.3
C. pseudotuberculosis PAT10	*ovis*	Ovino	Argentina	SOLiD *platform mate-paired* 25 bp	GCA_000221625.1
C. pseudotuberculosis 316	*equi*	Cavalo	EUA	Illumina HiSeq *paired-end*	GCA_000248375.2
C. pseudotuberculosis P54B96	*ovis*	Gnu	África do Sul	---	GCA_000255935.1
C. pseudotuberculosis 31	*equi*	Búfalo	Egito	Ion Torrent PGM 400 bp *fragment*	GCA_000259155.4
C. pseudotuberculosis 258	*equi*	Cavalo	Bélgica	Illumina HiSeq *paired-end*	GCA_000263755.3
C. pseudotuberculosis 3/99-5	*ovis*	Ovino	Escócia	454 GS-FLX e Solexa 50-bp *paired-end*	GCA_000241855.1
C. pseudotuberculosis 267	*ovis*	Lhama	EUA	SOLiD v3 Plus (Applied Biosystems)	GCA_000258385.1
C. pseudotuberculosis Cp162	*equi*	Camelo	Reino Unido	Illumina HiSeq *paired-end*	GCA_000265545.3
C. pseudotuberculosis CCUG 27541	*equi*	Cavalo	Noruega	---WTG---	GCA_000729725.1
C. pseudotuberculosis E9	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_006974085.1
C. pseudotuberculosis FRC41	*ovis*	Humano	França	Illumina HiSeq *paired-end*	GCA_000143705.2
C. pseudotuberculosis 48252	*ovis*	Humano	Noruega	PacBio	GCA_000730365.1
C. pseudotuberculosis CS_10	*ovis*	Caprino	Noruega	PacBio	GCA_000730405.1
C. pseudotuberculosis Ft_2193/67	*ovis*	Caprino	Noruega	PacBio	GCA_000730445.1
C. pseudotuberculosis 12C	*ovis*	Ovino	Brasil	Ion Torrent PGM	GCA_001017615.1
C. pseudotuberculosis 29156	*ovis*	Bovino	Israel	Illumina HiSeq *paired-end*	GCA_001026945.2
C. pseudotuberculosis PA01	*ovis*	Ovino	Brasil	Ion Torrent PGM	GCA_001456175.1
C. pseudotuberculosis 226	*ovis*	Caprino	EUA	Ion Torrent PGM	GCA_000972805.1
C. pseudotuberculosis 1002B	*ovis*	Caprino	Brasil	Ion Torrent PGM 200 bp *fragment*	GCA_001433475.1
C. pseudotuberculosis VD57	*ovis*	Caprino	Brasil	Ion Torrent PGM	GCA_000814865.1

Tabela 2 – Informações de todas as linhagens de *C. pseudotuberculosis* já depositadas no banco do *GenBank* no NCBI.

(continuação)

C. pseudotuberculosis 262	*equi*	Bovino	Bélgica	Ion Torrent PGM	GCA_001047215.2
C. pseudotuberculosis E19	*equi*	Cavalo	Chile	Ion Torrent PGM	GCA_001186445.1
C. pseudotuberculosis PO269-5	*ovis*	Caprino	Portugal	Ion Torrent PGM	GCA_001298505.1
C. pseudotuberculosis N1	*ovis*	Ovino	Equatorial Guiné	Ion Torrent PGM	GCA_001447295.1
C. pseudotuberculosis E56	*ovis*	Ovino	Egito	Ion Torrent PGM	GCA_001481755.1
C. pseudotuberculosis MEX25	*ovis*	Ovino	México	Ion Torrent PGM	GCA_001481675.1
C. pseudotuberculosis PO222/4-1	*ovis*	Caprino	Portugal	Ion Torrent PGM	GCA_001481715.1
C. pseudotuberculosis MB11	*equi*	Cavalo	EUA	Ion Torrent PGM *platform* 200bp	GCA_001579825.2
C. pseudotuberculosis MB14	*equi*	Cavalo	EUA	Ion Torrent PGM	GCA_001579865.1
C. pseudotuberculosis MB30	*equi*	Cavalo	EUA	Ion Torrent PGM	GCA_001579885.2
C. pseudotuberculosis MB66	*equi*	Cavalo	EUA	Ion Torrent PGM	GCA_001579925.1
C. pseudotuberculosis MEX29	*ovis*	Ovino	México	Ion Torrent PGM	GCA_001865765.1
C. pseudotuberculosis I37	*equi*	Bovino	Israel	Ion Torrent PGM	GCA_001889145.1
C. pseudotuberculosis E55	*ovis*	Ovino	Egito	Ion Torrent PGM	GCA_001653295.1
C. pseudotuberculosis MEX9	*ovis*	Caprino	México	Ion Torrent PGM	GCA_001653315.1
C. pseudotuberculosis XH02	*ovis*	Caprino	China	---WTG---	GCA_001719645.1
C. pseudotuberculosis PA02	*ovis*	Caprino	Brasil	Ion Torrent PGM	GCA_001663495.1
C. pseudotuberculosis 32	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867005.1
C. pseudotuberculosis 33	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867025.1
C. pseudotuberculosis 34	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001866985.1
C. pseudotuberculosis 35	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867045.1
C. pseudotuberculosis 36	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867065.1
C. pseudotuberculosis 38	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867085.1
C. pseudotuberculosis 39	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867105.1
C. pseudotuberculosis 43	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867125.1
C. pseudotuberculosis 46	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867145.1
C. pseudotuberculosis 48	*equi*	Búfalo	Egito	Ion Torrent PGM	GCA_001867165.1
C. pseudotuberculosis T1	*ovis*	Caprino	Brasil	Ion Torrent PGM 200 bp *fragment*	GCA_001682255.2
C. pseudotuberculosis Cp13	*ovis*	Caprino	Brasil	Ion Torrent PGM	GCA_001682275.1
C. pseudotuberculosis MB20	*equi*	Cavalo	EUA	Ion Torrent PGM	GCA_000764045.2
C. pseudotuberculosis MEX30	*equi*	Cavalo	México	Ion Torrent PGM	GCA_001922265.1

Tabela 2 – Informações de todas as linhagens de *C. pseudotuberculosis* já depositadas no banco do *GenBank* no NCBI. (continuação)

C. pseudotuberculosis MEX31	*equi*	Cavalo	México	Ion Torrent PGM	GCA_001922285.1
C. pseudotuberculosis MB122	*equi*	Cavalo	EUA	---WTG---	GCA_001969065.1
C. pseudotuberculosis MB336	*equi*	Cavalo	EUA	---WTG---	GCA_001969155.1
C. pseudotuberculosis MB44	*equi*	Cavalo	EUA	---WTG---	GCA_001969175.1
C. pseudotuberculosis PA05	*ovis*	Ovino	Brasil	---WTG---	GCA_001969465.1
C. pseudotuberculosis PA06	*ovis*	Ovino	Brasil	---WTG---	GCA_001969485.1
C. pseudotuberculosis MEX1	*ovis*	Caprino	México	Ion Torrent PGM	GCA_001975165.1
C. pseudotuberculosis MIC6	*ovis*	Ovino	Brasil	Ion Torrent PGM	GCA_002009415.1
C. pseudotuberculosis Phop	*ovis*	Caprino	Brasil	Ion Torrent PGM	GCA_002009385.1
C. pseudotuberculosis SigmaE	*ovis*	Caprino	Brasil	Ion Torrent PGM	GCA_002072715.1
C. pseudotuberculosis ATCC 19410	*ovis*	Ovino	EUA	Ion Torrent PGM 400 pb *fragment*	GCA_002155265.1
C. pseudotuberculosis MB302	*equi*	Cavalo	EUA	Illumina HiSeq *paired-end*	GCA_001969165.3
C. pseudotuberculosis PA04	*ovis*	Ovino	Brasil	Ion Torrent PGM	GCA_001989595.1
C. pseudotuberculosis MB278	*equi*	Cavalo	EUA	Ion Torrent PGM	GCA_001969145.2
C. pseudotuberculosis MB154	*equi*	Cavalo	EUA	Ion Torrent PGM	GCA_001969095.2
C. pseudotuberculosis PA07	*ovis*	Ovino	Brasil	Ion Torrent PGM	GCA_001942005.2
C. pseudotuberculosis PA08; A19	*ovis*	Ovino	Brasil	Ion Torrent PGM	GCA_002794435.1
C. pseudotuberculosis MB295	*equi*	Cavalo	EUA	Ion Torrent PGM	GCA_001969085.2
C. pseudotuberculosis CAP3W	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_002953275.1
C. pseudotuberculosis CAPJ4	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_002953315.1
C. pseudotuberculosis OVI2C	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_002953235.1
C. pseudotuberculosis OVI03	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_002953955.1
C. pseudotuberculosis KM01	*ovis*	Caprino	China	Illumina HiSeq	GCA_002869805.1
C. pseudotuberculosis DSM 20689	*ovis*		EUA	PacBio RS_II	GCA_003634885.1
C. pseudotuberculosis Cap1W	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_003955885.1
C. pseudotuberculosis OVIAF1	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_003955905.1
C. pseudotuberculosis 87MAT	*ovis*	Caprino ou Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_004193955.1
C. pseudotuberculosis CAPMI03	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_004193675.1
C. pseudotuberculosis CAPMI05	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_004193695.1
C. pseudotuberculosis CAPNAT1	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_004193655.1
C. pseudotuberculosis CR07	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_004193635.1

Tabela 2 – Informações de todas as linhagens de *C. pseudotuberculosis* já depositadas no banco do *GenBank* no NCBI. (continuação)

C. pseudotuberculosis 99MAT	*ovis*	Caprino ou Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_004295625.1
C. pseudotuberculosis E7	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_004353245.1
C. pseudotuberculosis Cap1R	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_004323775.1
C. pseudotuberculosis Cap8W	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_004323755.1
C. pseudotuberculosis OVI1FL	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_004941425.1
C. pseudotuberculosis Cap4W	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_005341465.1
C. pseudotuberculosis Cap5W	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_005341445.1
C. pseudotuberculosis 04MAT	*ovis*	Caprino ou Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_004332375.1
C. pseudotuberculosis 38MAT	*ovis*	Caprino ou Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_004332355.1
C. pseudotuberculosis OVICCA32	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_006974065.1
C. pseudotuberculosis OVID04	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_006971245.1
C. pseudotuberculosis OVIOS02	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_006971625.1
C. pseudotuberculosis OVIZ01	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_006971425.1
C. pseudotuberculosis CAPGE03	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_006971065.1
C. pseudotuberculosis E13	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_008461625.1
C. pseudotuberculosis E14	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_008462025.1
C. pseudotuberculosis E16	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_008461845.1
C. pseudotuberculosis Cap1C	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_004771435.1
C. pseudotuberculosis NCTC4656	*equi*	Cavalo	Reino Unido	---	GCA_901482545.1
C. pseudotuberculosis NCTC4681	*ovis*	Ovino	Austrália	---	GCA_901482565.1
C. pseudotuberculosis 17	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_009759745.1
C. pseudotuberculosis MEX2	*ovis*	Caprino	México	Illumina HiSeq *paired-end*	GCA_009759765.1
C. pseudotuberculosis PAT16	*ovis*	Ovino	Argentina	Illumina Hiseq *paired-end*	GCA_009759705.1
C. pseudotuberculosis SP165	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCA_009759725.1
C. pseudotuberculosis MB16	*equi*	Cavalo	EUA	Illumina HiSeq *paired-end*	GCA_009762695.1
C. pseudotuberculosis SigB	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_009762635.1
C. pseudotuberculosis Sigh	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_009762655.1
C. pseudotuberculosis SigK	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_009762615.1
C. pseudotuberculosis 133	*ovis*	---	Brasil	Illumina HiSeq *paired-end*	GCF_009789115.1

(conclusão)

C. pseudotuberculosis 414	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCF_009789135.1
C. pseudotuberculosis 206	*ovis*	Ovino	Brasil	Illumina HiSeq *paired-end*	GCF_009791415.1
C. pseudotuberculosis SigM	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCF_009791515.1
C. pseudotuberculosis PAT14	*ovis*	Ovino	Argentina	Illumina Hiseq *paired-end*	GCA_009905275.1
C. pseudotuberculosis SigD	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_009905175.1
C. pseudotuberculosis SigC	*ovis*	Caprino	Brasil	Illumina HiSeq *paired-end*	GCA_009930775.1

Apêndice B – Pipeline para fechamento de *gaps* em genomas bacterianos.

Durante a tese, foi construído e padronizado um pipeline para fechamento de *gaps* em genomas bacterianos. Isso ocorreu devido a crescente necessidade do laboratório em automatizar o processo de montagem de genomas. Para isso, criei um *script* em *shell* que auxilia neste processo. O *script* atua criando pastas, arquivos e gerenciando programas e outros *script*s para fechamento de *gaps*. Assim, com *contigs* gerados por montagens alternativas com mais de um software, associado a um genoma de referência completo, acurado e com alta similaridade nucleotídica. É possível submeter um arquivo com vários *contigs* e sair com um genoma completo, pronto para a etapa de anotação estrutural e funcional, seguida de depósito.

```
#!/bin/she
# Program: manage_contigs.sh
# Description: A simple script for help on close gap genome process.
# Written by: Thiago de Jesus Sousa. Laboratory of Cellular and Molecular
Genetics, UFMG, Brazil.
# Version: 1.3
# Date: 06/01/2019.
# Command line: sh manage_contigs.sh or ./manage_contigs.sh

# Requirements:
# I: Install programs and scripts ContiguatorF v.2.7.3, movednaa.py v.2,
GenomeFinisher v.1.4, GapBlaster v1.1.2, RNAmmer 1.2, hmmer-2.2g and Blast+
2.9 installed in usr/local/bin path.
# II: In addition, standard requirements such as: Linux Terminal and BioPython
v.1.73.
# Pipeline usage: You only need to change the paths of files and folders on
the variable box.
# key1 - Workspace full path.
# key2 - Name of the folder that will be created to organize the files.
# key3 - The file path of contig or scaffold (Inside workspace path).
# key4 - The reference genome file path, format ".fna" (Inside workspace
path).
# key5 - The reference genome file path, format ".gbk" (Inside workspace
path).
# key6 - The first word of the reference genome header, format ".fna". OBS:
Numbers with "." don't work.

#Variable box:
key1=/home/thiagojs/Genomes
key2=Tentativa_1
```

```
key3=scaffold_edena.fasta
key4=ATCC_19410.fasta
key5=ATCC_19410.gbk
key6=ATCC19410

#1-Managing folders and files generated.
echo"------------------------------------------------------------"
echo"Creating Folders"
cd $key1
mkdir $key2 && cp $key1/$key3 $key2
cd $key2 &&mkdirContiguatorF GapBlaster GenomeFinisherRNAmmerFinal_files
cd $key1/$key2/ContiguatorF&&mkdirmovednaa&&cdmovednaa&&mkdir ContiguatorF_2

#2-Sorting with the ContiguatorF.
echo"------------------------------------------------------------"
echo"Running the ContiguatorF"
cd $key1/$key2/ContiguatorF
CONTIGuatorF.py \
-c $key1/$key3 \
-r $key1/$key4 \
-g $key1/$key5

evince $key1/$key2/ContiguatorF/Map_$key6/*.pdf &
cp $key1/$key2/ContiguatorF/UnMapped*Contigs*/Excluded.fsa \
$key1/$key2/GapBlaster

#3-Fixing the start of the genome from the dnaA gene, script movednaa.py v.2.
echo"------------------------------------------------------------"
echo"Running the movednaa"
cd $key1/$key2/ContiguatorF/movednaa
python2.7 /usr/local/bin/movednaa.py \
$key1/$key2/ContiguatorF/Map_$key6/PseudoContig.fsa \
$key1/$key4

#4-Sorting with the ContiguatorF, after movednaa.py.
echo"------------------------------------------------------------"
echo"Running ContiguatorF after movednaa.py"
cd $key1/$key2/ContiguatorF/movednaa/ContiguatorF_2
CONTIGuatorF.py \
-c $key1/$key2/ContiguatorF/movednaa/f2.fasta \
-r $key1/$key4 \
-g $key1/$key5
```

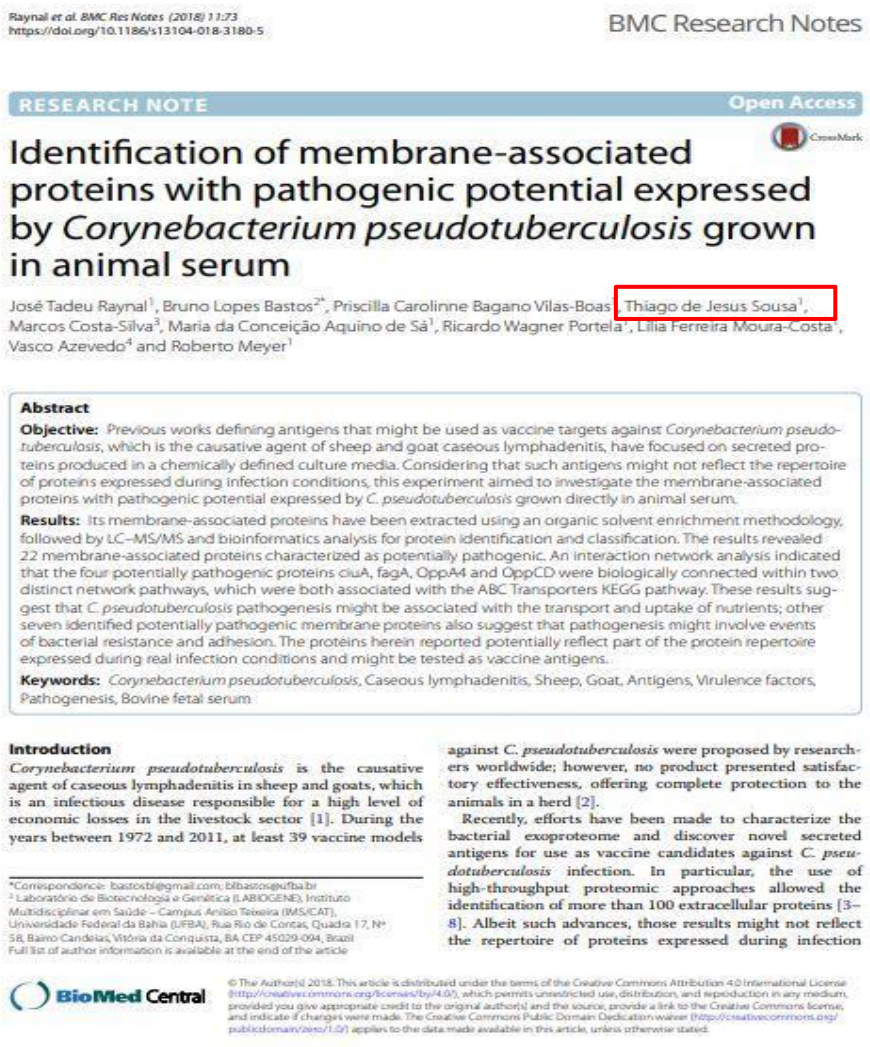
```
evince $key1/$key2/ContiguatorF/movednaa/ContiguatorF_2/Map_$key6/*.pdf &
cp $key1/$key2/ContiguatorF/movednaa/ContiguatorF_2/Map_$key6/*.fsa \
$key1/$key2/GapBlaster


#5-Gap closing with GapBlaster.
echo"----------------------------------------------------------------"
echo"Running GapBlaster"
cd $key1/$key2/GapBlaster
java -jar /usr/local/bin/GapBlaster_v1.1.2.jar

#6-Gap closing with GenomeFinisher v.1.4, after GapBlaster.
echo"----------------------------------------------------------------"
echo"Running GenomeFinisher"
cd $key1/$key2/GenomeFinisher
java -jar /usr/local/bin/GenomeFinisher.jar

cp $key1/$key2/GenomeFinisher/out/*.fasta \
$key1/$key2/Final_files
cd $key1/$key2/Final_files
mv*.fastaGenomeFinisher_final.fasta

#8-Sorting with the ContiguatorF with the final file.
echo"----------------------------------------------------------------"
echo"Running ContiguatorF"
cd $key1/$key2/Final_files
CONTIGuatorF.py \
-c $key1/$key2/Final_files/GenomeFinisher_final.fasta \
-r $key1/$key4 \
-g $key1/$key5

#9- Merge of all generated synteny representation.
cd $key1/$key2/Final_files
pdfunite \
$key1/$key2/ContiguatorF/Map_$key6/*.pdf \
$key1/$key2/ContiguatorF/movednaa/ContiguatorF_2/Map_$key6/*.pdf \
$key1/$key2/Final_files/Map_$key6/*.pdf \
Todas_sintenias.pdf

evince $key1/$key2/Final_files/*.pdf &

#10- Run rnammer
```

```
echo"------------------------------------------------------------"
echo"Running rnammmer"


cd $key1/$key2/RNAmmer
rnammer -S bac -m lsu,ssu,tsu -gff $key6.gff -h $key6.hmmreport -f $key6.fasta
< $key1/$key2/Final_files/GenomeFinisher_final.fasta

echo"------------------------------------------------------------"
echo"If the pipeline was useful, give me a star on GitHub ;)-
https://github.com/Thiagojsousa/Genome_Scripts"
echo"Thanks a lot and Be Happy!"
```

Disponível no GitHub: https://github.com/Thiagojsousa/Genome_Scripts

Apêndice C – Artigo*: Identification of membrane- associated proteins with pathogenic potential expressed by Corynebacterium pseudotuberculosis grow in animal serum.*

Raynal e colaboradores (RAYNAL *et al.*, 2018) construíram com um método para a identificação de proteínas associadas à membrana com potencial patogênico isolado do soro de animais infectados com *C. pseudotuberculosis*. Contribui na extração, identificação e análises de bioinformática dessas moléculas durante o período de iniciação científica em 2012 a 2014 no Laboratório de Imunologia e Biologia Molecular (LABIMUNO) no Instituto de Ciência da Saúde na Universidade Federal da Bahia. Com isso, surgiu o interesse inicial em trabalhar com *C. pseudotuberculosis,* antes da entrada no mestrado e doutorado da UFMG.



Raynal et al. BMC Res Notes (2018) 11:73
https://doi.org/10.1186/s13104-018-3180-5

BMC Research Notes

RESEARCH NOTE — Open Access

Identification of membrane-associated proteins with pathogenic potential expressed by *Corynebacterium pseudotuberculosis* grown in animal serum

José Tadeu Raynal[1], Bruno Lopes Bastos[2*], Priscilla Carolinne Bagano Vilas-Boas[1], Thiago de Jesus Sousa[1], Marcos Costa-Silva[3], Maria da Conceição Aquino de Sá[1], Ricardo Wagner Portela[1], Lília Ferreira Moura-Costa[1], Vasco Azevedo[4] and Roberto Meyer[1]

Abstract

Objective: Previous works defining antigens that might be used as vaccine targets against *Corynebacterium pseudotuberculosis*, which is the causative agent of sheep and goat caseous lymphadenitis, have focused on secreted proteins produced in a chemically defined culture media. Considering that such antigens might not reflect the repertoire of proteins expressed during infection conditions, this experiment aimed to investigate the membrane-associated proteins with pathogenic potential expressed by *C. pseudotuberculosis* grown directly in animal serum.

Results: Its membrane-associated proteins have been extracted using an organic solvent enrichment methodology, followed by LC–MS/MS and bioinformatics analysis for protein identification and classification. The results revealed 22 membrane-associated proteins characterized as potentially pathogenic. An interaction network analysis indicated that the four potentially pathogenic proteins ciuA, fagA, OppA4 and OppCD were biologically connected within two distinct network pathways, which were both associated with the ABC Transporters KEGG pathway. These results suggest that *C. pseudotuberculosis* pathogenesis might be associated with the transport and uptake of nutrients; other seven identified potentially pathogenic membrane proteins also suggest that pathogenesis might involve events of bacterial resistance and adhesion. The proteins herein reported potentially reflect part of the protein repertoire expressed during real infection conditions and might be tested as vaccine antigens.

Keywords: *Corynebacterium pseudotuberculosis*, Caseous lymphadenitis, Sheep, Goat, Antigens, Virulence factors, Pathogenesis, Bovine fetal serum

Introduction

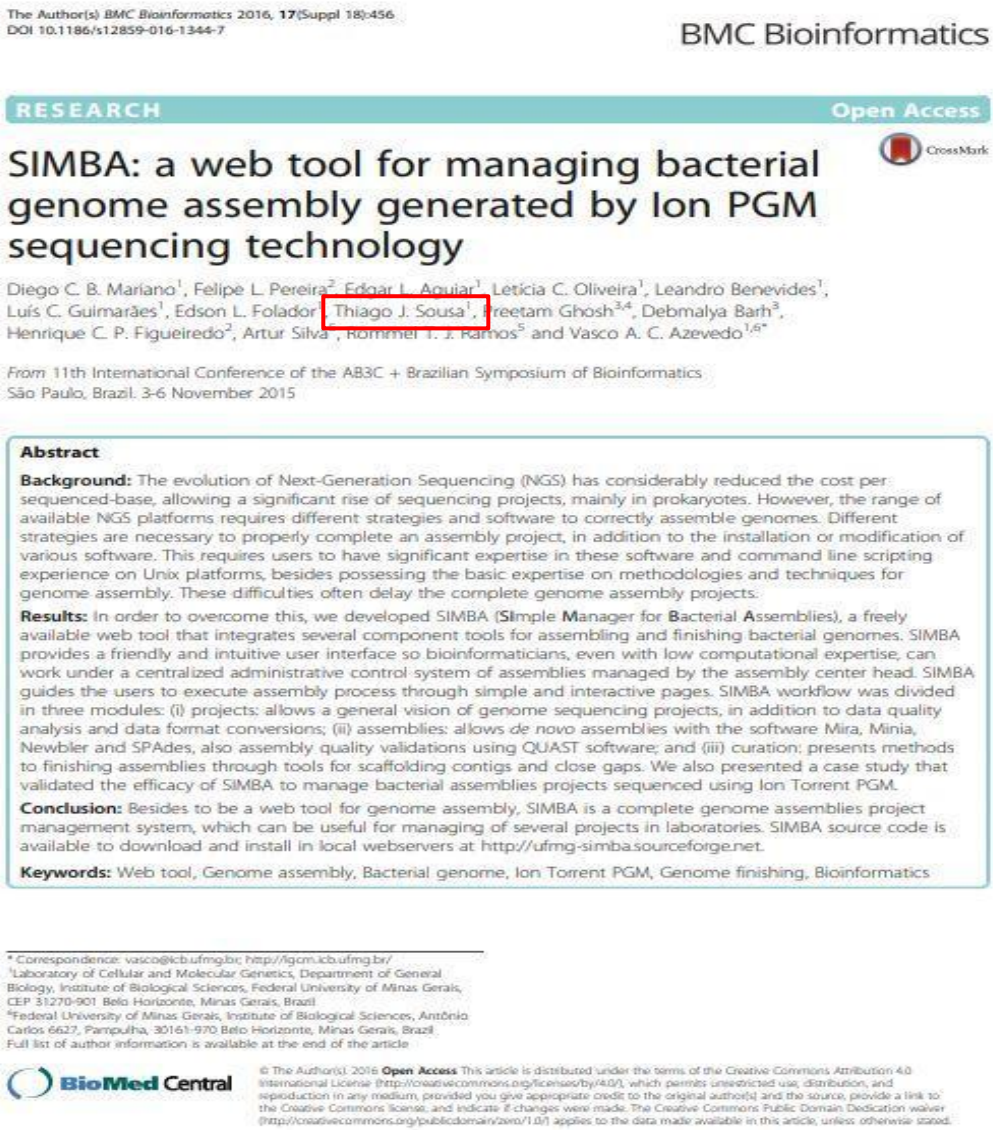
Corynebacterium pseudotuberculosis is the causative agent of caseous lymphadenitis in sheep and goats, which is an infectious disease responsible for a high level of economic losses in the livestock sector [1]. During the years between 1972 and 2011, at least 39 vaccine models against *C. pseudotuberculosis* were proposed by researchers worldwide; however, no product presented satisfactory effectiveness, offering complete protection to the animals in a herd [2].

Recently, efforts have been made to characterize the bacterial exoproteome and discover novel secreted antigens for use as vaccine candidates against *C. pseudotuberculosis* infection. In particular, the use of high-throughput proteomic approaches allowed the identification of more than 100 extracellular proteins [3–8]. Albeit such advances, those results might not reflect the repertoire of proteins expressed during infection

*Correspondence: bastosbl@gmail.com; blbastos@ufba.br
[2] Laboratório de Biotecnologia e Genética (LABIOGENE), Instituto Multidisciplinar em Saúde – Campus Anísio Teixeira (IMS/CAT), Universidade Federal da Bahia (UFBA), Rua Rio de Contas, Quadra 17, Nº 58, Bairro Candeias, Vitória da Conquista, BA CEP 45029-094, Brazil
Full list of author information is available at the end of the article

Apêndice D – Artigo: SIMBA: *a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology*.

Mariano e colaboradores (MARIANO, DIEGO C.B. *et al.*, 2016) no LGCM, construíram o programa SIMBA, que é ferramenta web para gerenciamento e montagem de genomas bacterianos. Nesse projeto contribui com as ideias de elaboração e testes de implementação da ferramenta. Nesse período, estava em meu primeiro período do mestrado, sendo um importante aprendizado dentro da bioinformática.



The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 18):456
DOI 10.1186/s12859-016-1344-7

BMC Bioinformatics

RESEARCH Open Access

SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology

Diego C. B. Mariano[1], Felipe L. Pereira[2], Edgar L. Aguiar[1], Letícia C. Oliveira[1], Leandro Benevides[1], Luís C. Guimarães[1], Edson L. Folador[1], Thiago J. Sousa[1], Preetam Ghosh[3,4], Debmalya Barh[3], Henrique C. P. Figueiredo[2], Artur Silva[2], Rommel T. J. Ramos[5] and Vasco A. C. Azevedo[1,6*]

From 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics
São Paulo, Brazil. 3-6 November 2015

Abstract

Background: The evolution of Next-Generation Sequencing (NGS) has considerably reduced the cost per sequenced-base, allowing a significant rise of sequencing projects, mainly in prokaryotes. However, the range of available NGS platforms requires different strategies and software to correctly assemble genomes. Different strategies are necessary to properly complete an assembly project, in addition to the installation or modification of various software. This requires users to have significant expertise in these software and command line scripting experience on Unix platforms, besides possessing the basic expertise on methodologies and techniques for genome assembly. These difficulties often delay the complete genome assembly projects.

Results: In order to overcome this, we developed SIMBA (**SI**mple **M**anager for **B**acterial **A**ssemblies), a freely available web tool that integrates several component tools for assembling and finishing bacterial genomes. SIMBA provides a friendly and intuitive user interface so bioinformaticians, even with low computational expertise, can work under a centralized administrative control system of assemblies managed by the assembly center head. SIMBA guides the users to execute assembly process through simple and interactive pages. SIMBA workflow was divided in three modules: (i) projects: allows a general vision of genome sequencing projects, in addition to data quality analysis and data format conversions; (ii) assemblies: allows *de novo* assemblies with the software Mira, Minia, Newbler and SPAdes, also assembly quality validations using QUAST software; and (iii) curation: presents methods to finishing assemblies through tools for scaffolding contigs and close gaps. We also presented a case study that validated the efficacy of SIMBA to manage bacterial assemblies projects sequenced using Ion Torrent PGM.

Conclusion: Besides to be a web tool for genome assembly, SIMBA is a complete genome assemblies project management system, which can be useful for managing of several projects in laboratories. SIMBA source code is available to download and install in local webservers at http://ufmg-simba.sourceforge.net.

Keywords: Web tool, Genome assembly, Bacterial genome, Ion Torrent PGM, Genome finishing, Bioinformatics

* Correspondence: vasco@icb.ufmg.br; http://lgcm.icb.ufmg.br/
[1]Laboratory of Cellular and Molecular Genetics, Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, CEP 31270-901 Belo Horizonte, Minas Gerais, Brazil
[6]Federal University of Minas Gerais, Institute of Biological Sciences, Antônio Carlos 6627, Pampulha, 30161-970 Belo Horizonte, Minas Gerais, Brazil
Full list of author information is available at the end of the article

Apêndice E – Artigo: *Complete Genome Sequence of the Attenuated Corynebacterium pseudotuberculosis T1.*

Através desse trabalho conseguimos depositar o genoma de *C. pseudotuberculosis* T1 (ALMEIDA *et al.*, 2016). Nesse projeto realizei a curadoria e depósito do genoma sobre a supervisão da Drª Sintia Almeida. Isso possibilitou minha compreensão do processo, desde o isolamento bacteriano até a etapa de depósito de um genoma no NCBI.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Complete Genome Sequence of the Attenuated *Corynebacterium pseudotuberculosis* Strain T1

Sintia Almeida,[a] Dan Loureiro,[b] Ricardo W. Portela,[b] Diego C. B. Mariano,[a] Thiago J. Sousa,[a] Felipe L. Pereira,[c] Fernanda A. Dorella,[c] Alex F. Carvalho,[c] Lilia F. Moura-Costa,[b] Carlos A. G. Leal,[c] Henrique C. Figueiredo,[c] Roberto Meyer,[b] Vasco Azevedo[a]

Laboratory of Cellular and Molecular Genetics, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[a]; Laboratory of Immunology and Molecular Biology, Federal University of Bahia, Salvador, Bahia, Brazil[b]; Aquacen, National Reference Laboratory for Aquatic Animal Diseases of Ministry of Fisheries and Aquaculture, Veterinary School, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[c]

We present here the genome sequence of the attenuated *Corynebacterium pseudotuberculosis* strain T1. The sequencing was performed with an Ion Torrent Personal Genome Machine platform. The genome is a circular chromosome of 2,337,201 bp, with a G+C content of 52.85% and a total of 2,125 coding sequences (CDSs), 12 rRNAs, 49 tRNAs, and 24 pseudogenes.

Received 12 July 2016 **Accepted** 19 July 2016 **Published** 8 September 2016

Citation Almeida S, Loureiro D, Portela RW, Mariano DCB, Sousa TJ, Pereira FL, Dorella FA, Carvalho AF, Moura-Costa LF, Leal CAG, Figueiredo HC, Meyer R, Azevedo V. 2016. Complete genome sequence of the attenuated *Corynebacterium pseudotuberculosis* strain T1. Genome Announc 4(5):e00947-16. doi:10.1128/genomeA.00947-16.

Address correspondence to Ricardo W. Portela, rwportela@ufba.br.

Corynebacterium pseudotuberculosis is the etiologic agent of caseous lymphadenitis, a chronic disease that affects small ruminants worldwide. *C. pseudotuberculosis* is a Gram-positive bacterium that belongs to the *Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rhodococcus* (CMNR) group (1). Here, we present the complete genome sequence of *C. pseudotuberculosis* strain T1. This strain was isolated from a goat lymph node in Bahia State, Brazil, and belongs to *C. pseudotuberculosis* bv. ovis; after several passages in culture, it was considered to present low virulence and was used as an attenuated immunogen in goats, producing 33.3% protection against caseous lymphadenitis clinical signs (2). It was described that the secreted/excreted antigens of the T1 strain showed 89% sensitivity and 99% specificity in the detection of specific anti-*C. pseudotuberculosis* IgG antibodies in sheep (1). Later, it was found that these T1 strain antigens were able to stimulate the production of gamma interferon by peripheral blood mononuclear cells of goats and sheep infected with the bacteria (3). The T1 strain could be used as an antigenic model for the detection of specific anti-*C. pseudotuberculosis* IgM antibodies in sheep (4).

The genome was sequenced using the Ion Torrent Personal Genome Machine (PGM) system, a 200-bp-fragment library kit, and a coverage of 110-fold. The quality of the reads was analyzed using the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), and *de novo* assembly was performed using Newbler 2.9 (Roche, USA). The assembly process produced seven contigs with an N_{50} value of 367,637 bp. The contigs were oriented and positioned based on an optical map. The optical mapping system measures the lengths of DNA fragments after digestion with restriction enzymes. This high-resolution technique can generate ordered maps of whole genomes and can also be used in the discrimination of closely related bacterial strains (5). Last, the Argus MapSolver software (OpGen, Inc., Gaithersburg, MD) was employed to import the DNA sequence and convert to *in silico* map data. For adjacent contigs with overlapping edges, SIMBA (http://ufmg-simba.sourceforge.net) was used. Repetitive regions were mapped with the CLC Genomics Workbench 7.0 software from Qiagen, USA (CLC bio), using as a reference the genome of *C. pseudotuberculosis* strain 1002. The gap-filling process was done with SIMBA (http://ufmg-simba.sourceforge.net), CLC Genomics Workbench 7.0, and in-house scripts. Automatic annotation was performed by transferring information from a curated database using in-house scripts. Genes encoding tRNAs, rRNAs, and some coding sequences (CDSs) that were absent following the transfer by in-house scripts were predicted using RAST (http://rast.nmpdr.org). All CDSs were manually curated using the Artemis software (6) and the UniProt database (http://www.uniprot.org).

The complete genome of *C. pseudotuberculosis* T1 includes one circular chromosome with a length of 2,337,201 bp, a G+C content of 52.85%, and a total of 2,125 CDSs, 12 rRNAs (5S, 16S, and 23S), 49 tRNAs, and 24 pseudogenes.

Accession number(s). The complete genome sequence has been deposited in GenBank under the accession no. CP015100.

ACKNOWLEDGMENTS

This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Apoio à Pesquisa e Extensão (FAPEX), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and Fundação de Apoio à Pesquisa do Estado da Bahia (FAPESB). The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

We declare that we do not have any conflict of interest in the publication of this work.

Fundação de Apoio à Pesquisa e Extensão (FAPEX) provided funding to Roberto Meyer. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) provided funding to Vasco Azevedo. Fundação de Apoio à Pesquisa do Estado da Bahia (FAPESB) provided funding to Dan Loureiro and Ricardo W. Portela.

Apêndice F – Arigo: ***Cell wall glycolipids from Corynebacterium pseudotuberculosis strains with 2 different virulences differ in terms of composition and immune recognition*.**

Neste trabalho, contribui na extração e moléculas durante o período de iniciação científica em 2012 a 2014 no Laboratório de Imunologia e Biologia Molecular (LABIMUNO) no Instituto de Ciência da Saúde na Universidade Federal da Bahia. Com isso, surgiu o interesse inicial em trabalhar com *C. pseudotuberculosis,* antes da entrada no mestrado e doutorado da UFMG.

Cell wall glycolipids from *Corynebacterium pseudotuberculosis* strains with different virulences differ in terms of composition and immune recognition

Miriam Flores Rebouças[1]; Dan Loureiro[1]; Thiago Doria Barral[1]; Nubia Seyffert[2]; José Tadeu Raynal[1]; Thiago Jesus Sousa[3]; Henrique Cesar Pereira Figueiredo[4], Vasco Azevedo[3]; Roberto Meyer[1]; Ricardo Wagner Portela[1*]

[1]Laboratory of Immunology and Molecular Biology, Health Sciences Institute, Federal University of Bahia, Salvador, Bahia, Brazil, 40110-100

[2]Post-graduation Program in Microbiology, Institute of Biology, Federal University of Bahia, Salvador, Bahia, Brazil, 40170-115

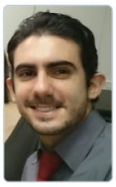
[3]Laboratory of Cellular and Molecular Genetics, Biological Sciences Institute, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, 31270-901

[4]National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

ORCID: MFR -; DL -; TDB - 0000-0002-1252-3605; NS - 0000-0002-2193-6508; JTR - 0000-0002-2771-0235; TJS - 0000-0001-9809-8883; HCPF - 0000-0002-1022-6842; VA - 0000-0002-4775-2280; RM - 0000-0002-4727-4805; RWP - 0000-0001-9095-776X

***Corresponding author:** rwportela@ufba.br

Apêndice G - Atividade presentes no currículo lattes - 2016 a 2020.



Thiago de Jesus Sousa

Endereço para acessar este CV: **http://lattes.cnpq.br/3620345994509054**

Última atualização do currículo em 09/01/2020

Resumo informado pelo autor

Possui graduação em Biotecnologia pela Universidade Federal da Bahia (2013). Mestrado em Bioinformática pelo Programa Interunidades de Pós-Graduação em Bioinformática pela Universidade Federal de Minhas Gerais (2016). Atualmente, doutorando em Bioinformática no Laboratório de Genética Celular e Molecular da UFMG, no qual desenvolve pesquisas nas áreas de Microbiologia, Genética e Bioinformática, com aplicação em montagem de genomas, mapa óptico e genômica comparativa (http://lgcm.icb.ufmg.br/site/).

(Texto informado pelo autor)

Links para Outras Bases:

SciELO - Artigos em texto completo

Producão

Produção bibliográfica

Artigos completos publicados em periódicos

1. doi **SOUSA, THIAGO DE JESUS**; PARISE, DOGLAS; PROFETA, RODRIGO; PARISE, MARIANA TEIXEIRA DORNELLES; GOMIDE, ANNE CYBELLE PINTO; KATO, RODRIGO BENTOS; PEREIRA, FELIPE LUIZ; FIGUEIREDO, HENRIQUE CESAR PEREIRA; RAMOS, ROMMEL; BRENIG, BERTRAM; COSTA DA SILVA, ARTUR LUIZ DA; GHOSH, PREETAM; BARH, DEBMALYA; GÓES-NETO, ARISTÓTELES; AZEVEDO, VASCO
Re-sequencing and optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes. Scientific Reports. JCR, v.9, p.16387 - , 2019.

2. doi RAYNAL, JOSÉ TADEU; BASTOS, BRUNO LOPES; VILAS-BOAS, PRISCILLA CAROLINNE BAGANO; **SOUSA, THIAGO DE JESUS**; COSTA-SILVA, MARCOS; DE SÁ, MARIA DA CONCEIÇÃO AQUINO; PORTELA, RICARDO WAGNER; MOURA-COSTA, LÍLIA FERREIRA; AZEVEDO, VASCO; MEYER, ROBERTO
Identification of membrane-associated proteins with pathogenic potential expressed by Corynebacterium pseudotuberculosis grown in animal serum. BMC RESEARCH NOTES. , v.11, p.11:73 - , 2018.

3. doi GOMIDE, ANNE CYBELLE PINTO; DE SÁ, PABLO GOMES; CAVALCANTE, ANA LIDIA QUEIROZ; **DE JESUS SOUSA, THIAGO**; GOMES, LUCAS GABRIEL RODRIGUES; RAMOS, ROMMEL THIAGO JUCA; AZEVEDO, VASCO; SILVA, ARTUR; FOLADOR, ADRIANA RIBEIRO CARNEIRO
Heat shock stress: Profile of differential expression in Corynebacterium pseudotuberculosis biovar Equi. GENE. JCR, v.645, p.124 - 130, 2017.

4. doi LOUREIRO, DAN; PORTELA, RICARDO W.; **SOUSA, THIAGO J.**; ROCHA, FLÁVIA; PEREIRA, FELIPE L.; DORELLA, FERNANDA A.; CARVALHO, ALEX F.; MENEZES, NILDO; MACEDO, EDUARDO S.; MOURA-COSTA, LILIA F.; MEYER, ROBERTO; LEAL, CARLOS A. G.; FIGUEIREDO, HENRIQUE C.; AZEVEDO, VASCO
Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Viscerotropic Strain N1. Genome Announcements. , v.4, p.e01673-15 - , 2016.

5. doi ALMEIDA, SINTIA; LOUREIRO, DAN; PORTELA, RICARDO W.; MARIANO, DIEGO C. B.; **SOUSA, THIAGO J.**; PEREIRA, FELIPE L.; DORELLA, FERNANDA A.; CARVALHO, ALEX F.; MOURA-COSTA, LILIA F.; LEAL, CARLOS A. G.; FIGUEIREDO, HENRIQUE C.; MEYER, ROBERTO; AZEVEDO, VASCO
Complete Genome Sequence of the Attenuated Strain T1. Genome Announcements. , v.4, p.e00947-16 - , 2016.

6. doi SOUZA, BIANCA C.; SENA, LUDMILLA S.; LOUREIRO, DAN; RAYNAL, JOSÉ T.; **SOUSA, THIAGO J.**; BASTOS, BRUNO L.; MEYER, ROBERTO; PORTELA, RICARDO W.
Determinação de valores de referência séricos para os eletrólitos magnésio, cloretos, cálcio e fósforo em ovinos das raças Dorper e Santa Inês. Pesquisa Veterinária Brasileira (Online). JCR, v.36, p.167 - 173, 2016.

7. doi MARIANO, DIEGO C. B.; PEREIRA, FELIPE L.; AGUIAR, EDGAR L.; OLIVEIRA, LETÍCIA C.; BENEVIDES, LEANDRO; GUIMARÃES, LUÍS C.; FOLADOR, EDSON L.; **SOUSA, THIAGO J.**; GHOSH, PREETAM; BARH, DEBMALYA; FIGUEIREDO, HENRIQUE C. P.; SILVA, ARTUR; RAMOS, ROMMEL T. J.; AZEVEDO, VASCO A. C.
SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology. BMC Bioinformatics. JCR, v.17, p.65 - 72, 2016.

8. doi MARIANO, DIEGO CÉSAR BATISTA; **SOUSA, THIAGO DE JESUS**; PEREIRA, FELIPE LUIZ; ABURJAILE, FLÁVIA; BARH, DEBMALYA; ROCHA, FLÁVIA; PINTO, ANNE CYBELLE; HASSAN, SYED

//wwws.cnpq.br/cvlattesweb/pkg_impcv.trata

SHAH; SARAIVA, TESSÁLIA DINIZ LUERCE; DORELLA, FERNANDA ALVES; DE CARVALHO, ALEX FIORINI; LEAL, CARLOS AUGUSTO GOMES; FIGUEIREDO, HENRIQUE CÉSAR PEREIRA; SILVA, ARTUR; RAMOS, ROMMEL THIAGO JUCÁ; AZEVEDO, VASCO ARISTON CARVALHO
Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002. BMC Genomics. JCR, v.17, p.1 - 7, 2016.

Apresentação de trabalho e palestra

1. SANTOS, L. N. Q.; ALVES, L. G.; OLIVEIRA, L. C.; SOARES, S. S.; **SOUSA, T. J.**
ANÁLISE DAS LINHAGENS DE Acinetobacter lwoffii VIA GENÔMICA COMPARATIVA, 2019. (Comunicação,Apresentação de Trabalho)

2. **SOUSA, T. J.**; GOMIDE, ANNE CYBELLE PINTO; OLIVEIRA, L. C.; SEYFFERT, N.; BRENIG, B.; COSTA, M. M.; SOARES, S. S.; AZEVEDO, V. A. C.
Comparative genomic analysis of Corynebacterium pseudotuberculosis: A quest for biofilm biosynthesis genes., 2019. (Congresso,Apresentação de Trabalho)

3. MIRANDA, F.M.; **SOUSA, T. J.**; Kato, R.B.; CAVALCANTE, A.L.Q.,; AZEVEDO, V. A. C.; SILVA, A.L.C.; RAMOS, R. T. J.
DEVELOPMENT OF A PIPELINE TO ASSEMBLE SHORT READS SEQUENCED BY ILLUMINA HISEQ, 2019. (Congresso,Apresentação de Trabalho)

4. SANTOS, L. N. Q.; ALVES, L. G.; OLIVEIRA, L. C.; SOARES, S. S.; **SOUSA, T. J.**
Genômica Comparativa: Estudo da plasticidade das linhagens de Acinetobacter Lwoffii, 2019. (Comunicação,Apresentação de Trabalho)

5. MADEIRA, J. C. C.; TREVISAN, R. O.; SANTOS, M. M.; DESIDERIO, C. S.; SILVA, M. O.; BOVI, W. G.; **SOUSA, T. J.**; OLIVEIRA, L. C.; ALVES, L. G.; JAISWAL, A. K.; TIWARI, S.; SOARES, S. S.; OLIVEIRA, C. J. F.; SILVA, M. V.
IMMUNOINFORMATICS-AIDED DESIGN OF POTENTIAL VACCINE CANDIDATES AND DRUGS' TARGETS AGAINST Trypanosoma CRUZI, 2019. (Congresso,Apresentação de Trabalho)

6. **SOUSA, T. J.**; PARISE, DOGLAS; Kato, R.B.; GOMIDE, ANNE CYBELLE PINTO; Profeta, R.; PEREIRA, FELIPE L.; FIGUEIREDO, H. C. P.; SILVA, A.; RAMOS, R. T. J.; AZEVEDO, VASCO
Detection of inversion in the genome of Corynebacterium pseudotuberculosis strains by optical mapping, 2018. (Congresso,Apresentação de Trabalho)

7. GANTOIS, R. C.; Hurtado, R.; Profeta, R.; **SOUSA, T. J.**; VIANA, MARCUS VINICIUS CANÁRIO; GOMIDE, ANNE CYBELLE PINTO; SILVA, A.; BARAUNA, R. A.; AZEVEDO, V. A. C.
Genome assembly completeness and its effect on phylogenetic estimation, 2017. (Congresso,Apresentação de Trabalho)

8. **SOUSA, THIAGO J.**
Aplicação do mapa ótico na detecção e correção de erros de montagens em genomas de Corynebacterium pseudotuberculosis, 2016. (Seminário,Apresentação de Trabalho)

9. PARISE, DOGLAS; **SOUSA, T. J.**; PARISE, M. T. D.; BUCIO, A. V. M.; PEREIRA, F. L.; DORELLA, FERNANDA A.; APARICIO, E. D.; FIGUEIREDO, H. C. P.; COSTA, D. A.; AZEVEDO, V. A. C.
Assembly, annotation and comparison of Corynebacterium pseudotuberculosis lineages, 2016. (Congresso,Apresentação de Trabalho)

10. GANTOIS, R. C.; **SOUSA, THIAGO J.**; PARISE, DOGLAS; COSTA, D. A.; PINTO, ANNE CYBELLE; FIGUEIREDO, H. C. P.; AZEVEDO, V. A. C.
Comparative Genomics between two different biovars of Corynebacterium pseudotuberculosis isolated in the same host, 2016. (Congresso,Apresentação de Trabalho)

11. VIANA, MARCUS VINICIUS CANÁRIO; PARISE, DOGLAS; **SOUSA, THIAGO J.**; BENEVIDES, LEANDRO JESUS; MARIANO, D. C. B.; ROCHA, F. S.; BAGANO, PRISCILLA; GUIMARÃES, LUIS CARLOS; PEREIRA, F. L.; DORELLA, FERNANDA A.; RAMOS, R. T. J.; SELIM, S. A. K.; SALAHELDEAN, M.; SILVA, A.; WATTAM, A. R.; AZEVEDO, V. A. C.
Complete genome sequence of Corynebacterium pseudotuberculosis 33, 2016. (Congresso,Apresentação de Trabalho)

12. **SOUSA, T. J.**; PARISE, DOGLAS; MARIANO, D. C. B.; COSTA, D. A.; PEREIRA, F. L.; FIGUEIREDO, H. C. P.; SILVA, A.; RAMOS, R. T. J.; AZEVEDO, V. A. C.
Detection and correction mis-assemblies in genome of Corynebacterium pseudotuberculosis, 2016. (Congresso,Apresentação de Trabalho)

13. **SOUSA, T. J.**
Mapa óptico:Inovação na geração de dados genômicos com alta precisão, 2016. (Conferência ou palestra,Apresentação de Trabalho)