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

Dissertação

Genômica comparativa entre os biovares Equi e Ovis de *Corynebacterium pseudotuberculosis* isolados no México

MESTRANDO: **Doglas Parise**
ORIENTADOR: **Prof. Dr. Vasco Ariston de Carvalho Azevedo**
CO-ORIENTADORAS: **Dra. Anne Cybelle Pinto Gomide**
Dra. Daniela Arruda Costa

BELO HORIZONTE
Novembro – 2016

Doglas Parise

Genômica comparativa entre os biovares Equi e Ovis de *Corynebacterium pseudotuberculosis* isolados no México

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo programa Interunidades de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo

Co-orientadora: Dra. Anne Cybelle Pinto Gomide

Dra. Daniela Arruda Costa

BELO HORIZONTE

Novembro – 2016

Parise, Doglas.
043 Genômica comparativa entre os biovares Equi e Ovis de Corynebacterium pseudotuberculosis isolados no México [manuscrito] / Doglas Parise. – 2016.

94 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo. Coorientadoras: Dra. Anne Cybelle Pinto Gomide, Dra. Daniela Arruda Costa.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Corynebacterium pseudotuberculosis - Teses. 2. Genômica - Teses. 3. Vacinas. 4. Bioinformática - Teses. I. Azevedo, Vasco Ariston de Carvalho. II. Gomide, Anne Cybelle Pinto. III. Costa, Daniela Arruda. IV. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. V. Título.

CDU: 573:004

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE DISSERTAÇÃO

28/2016
entrada
1º/2015
CPF:
836.962.020-53

Doglas Parise

Às quatorze horas do dia **29 de novembro de 2016**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Genômica comparativa entre os biovares Equi e Ovis de Corynebacterium pseudotuberculosis isolados no México"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Vasco Ariston de Carvalho Azevedo**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dr. Vasco Ariston de Carvalho Azevedo	UFMG	233.141.225-49	APROVADO
Dra. Anne Cybelle Pinto	UFMG	00123274680	APROVADO
Dr. Gabriel da Rocha Fernandes	FIOCRUZ	05270621622	Aprovado
Dr. Siomar de Castro Soares	UFTM	05695182611	APROVADO
Dr. Thiago Luiz de Paula Castro	UFMG	074 003 946-97	Aprovado
Dr. Eric Roberto G. Rocha Aguiar	UFMG	02507164 5-08	Aprovado
Dr. José Miguel Ortega	UFMG	05950126807	Aprovado

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

Belo Horizonte, 29 de novembro de 2016.

Dr. Vasco Ariston de Carvalho Azevedo - Orientador _____

Dra. Anne Cybelle Pinto _____

Dr. Gabriel da Rocha Fernandes _____

Dr. Siomar de Castro Soares _____

Dr. Thiago Luiz de Paula Castro _____

Dr. Eric Roberto Guimarães Rocha Aguiar _____

Dr. José Miguel Ortega _____

“Carregue dentro si apenas o bem. O amor, a bondade e a paz são sempre boas companhias.”

Autor Desconhecido

“Decidi não esperar as oportunidades e sim, buscá-las. Decidi ver cada dia como uma nova oportunidade de ser feliz.”

Walt Disney

Agradecimentos

Agradeço em primeiro lugar a Deus pelo dom da vida e por todas as oportunidades que tenho recebido. Ao Professor Vasco Azevedo por ter aceitado me orientar e me guiar neste trabalho e em minha vida acadêmica e também as Doutoras Daniela Costa e Anne Gomide por todo auxilio prestado e contribuições neste trabalho.

Ao Professor Efrén Díaz e ao Adrian Valentín Muñoz Bucio da *Universidad Nacional Autónoma de México* por ter disponibilizado as linhagens estudadas neste trabalho. A toda equipe do AQUACEN e em especial ao Felipe Pereira por sua disponibilidade em me auxiliar sempre que precisei.

A CAPES pela concessão da bolsa e a Universidade Federal de Minas Gerais.

A Sheila, Natália, Stefane e Fernanda por toda a atenção, carinho e auxilio em tudo que precisei.

A toda minha família e em especial aos meus pais Moacir e Imgrid Parise pela educação e apoio em todos os momentos e também a meu irmão Diogo Parise. Aos meus sogros Elvis Dornelles e Daniela Teixeira por todo o apoio em todos estes anos, e também pelo Elvis ter-me “apresentado” à bioinformática.

A minha esposa Mariana Teixeira Dornelles Parise por todo amor, companheirismo, apoio, paciência e auxílio incondicionais em todos os momentos e principalmente durante o desenvolvimento deste trabalho. E também por ter contribuído para o desenvolvimento deste principalmente discutindo, lendo e contribuindo com sua opinião. Obrigado por estar ao meu lado em mais esta conquista.

A todo time do LGCM que sempre esteve disposto a me auxiliar e esclarecer todas as minhas dúvidas o meu muito obrigado. Em especial gostaria de agradecer a Thiago Sousa, Marcus, Raquel, Alberto, Yasmim, Izabela, Flávia, Nilson, Sandeep, Jamal e Arun.

A Siomar, Letícia e Rommel que muitas vezes me auxiliaram e esclareceram minhas dúvidas.

E por fim, mas não menos importante a Diego, Paulo e Edgar por ter me ajudado quando entrei no laboratório e por sua amizade.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E TERMOS EM INGLÊS

ab initio	Do início.
ATP	*Adenosine triphosphate*. Trifosfato de adenosina.
beads	Microesferas.
BHI	*Brain Heart Infusion*. Infusão Cérebro Coração.
BLAST	*Basic Local Alignment Search Tool*.
BLASTn	*Nucleotide-nucleotide* BLAST.
BLASTp	*Protein-protein* BLAST.
BRIG	*BLAST Ring Image Generator*.
CDS	*Coding Sequence.* Sequência codificante.
chip	Dispositivo microeletrônico.
CLA	*Caseous lymphadenitis*. Linfadenite caseosa.
clusters	Grupos.
CMNR	*Corynebacterium*, *Mycobacterium*, *Nocardia* e *Rhodococcus*.
CRISPR-Cas	*Clustered regularly interspaced short palindromic repeats and associated genes*. Repetições palindrômicas curtas agrupadas e regularmente interespaçadas e genes associados.
crRNA	CRISPR RNA
GC	Guanina e Citosina.
Gb	Giga bases.
contigs	Fragmento de DNA formado pelo consenso de *reads*.
default	Padrão.
DNA	*Deoxyribonucleic acid*. Ácido desoxirribonucléico.
fag(ABCD)	*Fe acquisition genes*. Genes de aquisição de ferro.
frameshifts	Deslocamento do quadro de leitura em uma CDS.
FIGfams	*Protein Families*. Famílias proteicas.

gaps	Sequências genômicas não conhecidas em uma fita de DNA.
Formato gbk	Formato GenBank.
GEI	*Genomic island*. Ilha genômica.
GOLD	*Genomes On-line Database*.
indels	Inserção e deleção de nucleotídeos.
k-mer	Sequência de caracteres de tamanho k que se repete mais de uma vez em uma sequência.
mate-pair	Técnica que gera longos *pair-ends*.
Mb	Megabases.
MI	*Metabolic island.* Ilha metabólica.
MntR	*Manganese transport regulator*. Regulador de transporte de magnésio.
MT	Metilação.
MTase	DNA metiltransferase.
NCBI	*National Center for Biotechnology Information.*
ncRNA	RNA não codificante.
NGS	*Next Generation Sequencing.* Sequenciamento de nova geração.
NIH	*National Institutes of Health.*
NJ	*Neighbor-joining*. Aproximação dos vizinhos.
nt	Nucleotídeos.
OLC	*Overlap Layout Consensus.*
ORF	*Open Reading Frame*. Fase de leitura aberta.
PAI	*Pathogenicity island*. Ilha de patogenicidade.
pair-end	Técnica que realiza o sequenciamento das duas extremidades de curtos fragmentos de DNA, gerando *reads* nas duas direções.
PATRIC	*Pathosystems Resource Integration Center*.
pb	Pares de base.
PCR	*Polimerase Chain Reaction*. Reação em cadeia da polimerase.
PGfams	PATRIC *cross-genus families*.

PGM	*Personal Genome Machine*.
Phred	Medida de qualidade das bases nucleotídicas identificadas por sequenciamento automático.
PLD	Fosfolipase D.
PLfams	PATRIC *genus-specific families*.
QUAST	*Quality assessment tool*.
RAST	*Rapid Annotation using Subsystem Technology.*
RE	Restrição.
reads	Leituras geradas pelo sequenciador dos fragmentos de DNA.
REase	*Restriction endonuclease*. Endonuclease de restrição.
RI	*Resistance island*. Ilha de resistência.
RM *systems*	*Restriction-modification systems*. Sistemas de restrição de modificação.
RNA	*Ribonucleic Acid*. Ácido ribonucléico.
RNR	*Ribonucleotide reductase*. Ribonucleotídeo redutase.
rRNA	RNA ribossomal.
scaffolds	*Contigs* ordenados.
score	Pontuação.
SI	*Symbiotic island*. Ilha simbiótica.
single-end	Fragmentos simples.
single-reads	Leituras únicas geradas pelo sequenciador.
SMRT	*Single Molecule, Real Time*.
SNP	*Single Nucleotide Polymorphism*.
SOLiD	*Sequencing by Oligonuclotide Ligation and Detection*.
spacer	Espaçador.
start codon	Códon de iniciação.
stop codon	Códon de parada.
substrings	Subcadeia de uma cadeia de caracteres.
template	Molde utilizado para realizar o sequenciamento.

threads	É uma forma de dividir um processo que está sendo executado pelo computador em subprocessos.
throughput	Quantidade total de dados gerada pelo sequenciador.
tRNA	RNA transportador.
WGS	*Whole Genome Sequencing.* Sequenciamento de genoma completo.

RESUMO

O desenvolvimento das tecnologias de sequenciamento de nova geração (NGS) tem permitido campos de estudos como o da genômica comparativa crescerem e levado a uma melhor compreensão de diferentes organismos. Dentre os quais podemos destacar bactérias patogênicas como a *Corynebacterium pseudotuberculosis*, a qual pode ser dividida em dois biovares. As pertencentes ao biovar Ovis afetam principalmente pequenos ruminantes e as do biovar Equi afetam animais de maior porte como equinos, bubalinos e bovinos. As doenças causadas por este patógeno têm levado a grandes perdas financeiras ao agronegócio mundial. Neste trabalho foram estudadas seis linhagens de *C. pseudotuberculosis* isoladas de caprinos, ovinos e equinos todas provenientes do México, sendo o primeiro estudo *in silico* deste patógeno isolado neste país. Para tal os DNAs genômicos foram sequenciados utilizando a tecnologia de sequenciamento Ion PGM™, seguido da montagem, anotação e depósito no NCBI. Nos estudos de genômica comparativa foram utilizados os seguintes softwares: *Protein Family Sorter* do PATRIC (pangenoma), Gegenees e FigTree (geração da árvore filogenômica), GIPSy e BRIG (predição e visualização das ilhas de patogenicidade (PAIs), respectivamente) e o *Specialty Genes Search tool* do PATRIC (predição de alvos para drogas). Na análise de pangenoma dos seis organismos o valor do pangenoma foi de 2295, do genoma central foi de 1903, do genoma acessório foi de 343 e quantidade de genes únicos foi de 49. Referente a análise filogenômica chegou-se a um agrupamento por biovar e também uma maior similaridade genômica pôde ser observada por hospedeiro. Em conformidade com os resultados da filogenômica, nas análises de PAIs encontrou-se semelhanças destas por biovar e uma PAI presente em todos os organismos. Além disso, foi encontrado um possível alvo para drogas em cinco dos seis genomas estudados. Este estudo visou contribuir com o conhecimento biológico de *C. pseudotuberculosis* e na identificação de prováveis fatores de virulência que podem ser utilizados em estudos que visam o desenvolvimento de vacinas e drogas para erradicação das diferentes doenças causadas por este organismo.

ABSTRACT

Next generation sequencing (NGS) technologies development have allowed study fields such as comparative genomics to grow and led to a better comprehension of different organisms. Among those, we can highlight pathogenic bacteria such as *Corynebacterium pseudotuberculosis*, which can be divided in two biovars. The ones belonging to biovar Ovis affect mainly small ruminants and ones belonging to biovar Equi affects larger animals such as equines, buffaloes and cattle. These diseases have been causing huge financial losses in the global agribusiness. In this work, six strains of *C. pseudotuberculosis* isolated from goats, sheep and horses from Mexico were studied; this is the first *in silico* research concerning strains of this organism isolated from this country. The genomic DNAs were sequenced by using Ion PGM™ sequencing technology, followed by the assembly, annotation and deposit of the resulting genomes at NCBI. In the comparative genomics studies, the following software were used: Protein Family Sorter from PATRIC (pan-genome), Gegenees and FigTree (generation of phylogenomic tree), GIPSy and BRIG (prediction and visualization of pathogenic islands (PAIs), respectively) and the Specialty Genes Search tool from PATRIC (prediction of drug targets). In the pan-genome analyses of the six organisms, the value of the pan-genome was 2295, the core genome was 1903, the accessory genome was 343 and the singletons was 49. Concerning the phylogenomic analyses, it was noticed that the strains clustered by biovar, and the strains from the same host present higher genomic similarity. In accordance with the phylogenomic results, it was found similarities by biovar in the PAI analysis, and one of the found PAIs was present in all organisms. Furthermore, it was found a putative drug target in five out of six studied genomes. This work aimed to contribute with the biological knowledge of *C. pseudotuberculosis* and in the identification of putative virulence factors that can be used in researches that focus in the development of vaccines and drugs for eradication of different diseases caused by this organism.

ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está dividida em sete seções. A primeira seção apresenta a revisão bibliográfica apresentando um breve histórico da genômica, incluindo tecnologias de sequenciamento, genômica comparativa e organismo de interesse. Na seção dois é apresentada a justificativa deste trabalho. A seção três é dividida em objetivo geral e objetivos específicos. Na seção quatro é apresentada a metodologia desenvolvida nesse trabalho. A seção cinco apresenta os resultados obtidos em cada etapa e a discussão dos mesmos. As seções seis e sete apresentam, respectivamente, as conclusões e perspectivas deste trabalho.

1. REVISÃO BIBLIOGRÁFICA

1.1. Breve histórico genômico

A genômica é a área de conhecimento que estuda o genoma (MCKUSICK & RUDDLE, 1987). Projetos de sequenciamento de genomas podem ser realizados com o intuito de identificar e caracterizar um determinado organismo, assim como analisar questões evolutivas (VERLI *et al.*, 2014). O primeiro genoma sequenciado foi o bacteriófago *phi* X174 no ano de 1977 por Sanger e colaboradores através do sequenciamento por dideoxinucleotídeos, que em 1980 rendeu o prêmio Nobel de Química aos autores (SANGER *et al.*, 1977). O desenvolvimento desta metodologia permitiu o sequenciamento do primeiro organismo bacteriano, o qual ocorreu em 1995 e desvendou a sequência genômica da bactéria *Haemophilus influenzae* (FLEISCHMANN *et al*., 1995). O contínuo avanço das tecnologias de sequenciamento possibilitou que código genético humano fosse revelado, o que levou 13 anos para ser concluído (LANDER *et al*., 2001; VENTER *et al*., 2001; MARIANO, 2015).

Após o sequenciamento do genoma humano, o contínuo aperfeiçoamento das tecnologias de sequenciamento levou ao desenvolvimento dos sequenciadores de nova geração (NGS), o qual revolucionou os projetos de sequenciamento de genomas, reduzindo o custo destes em comparação às metodologias anteriores, além de diminuir o tempo necessário para o processamento das amostras (COSTA, 2015; LOMAN & PALLEN, 2015). Com isso, a quantidade de genomas sequenciados e depositados cresceu exponencialmente. Atualmente, no GOLD existem 51.302 projetos de genomas bacterianos, como pode ser observado na Figura 1 (GOLD, 2016).

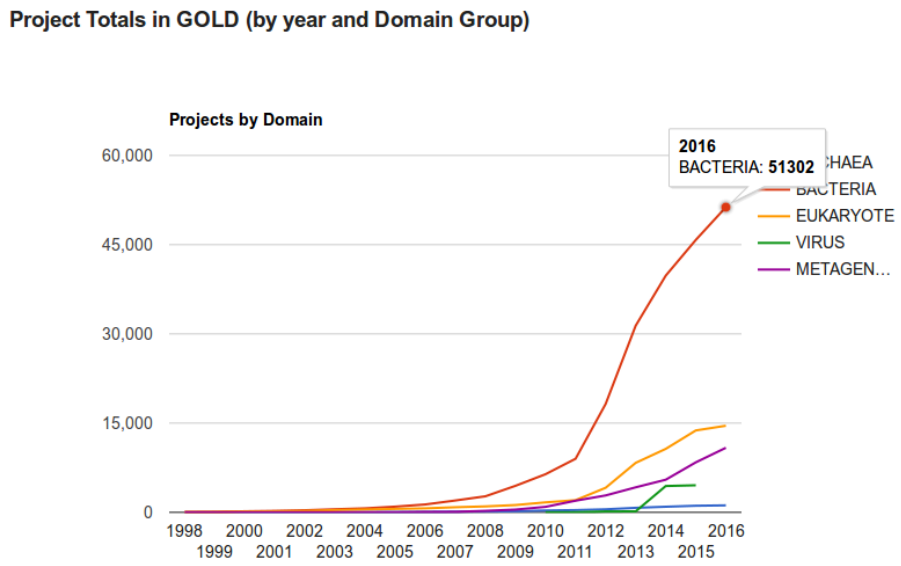


Figura 1 – Número de projetos de genomas bacterianos depositados no GOLD.

Fonte: GOLD (2016).

Esse crescimento no número de projetos genomas tem colaborado para uma melhor compreensão dos micro-organismos, como por exemplo, a sua evolução. Conhecimento que propicia a identificação de patógenos, bem como o acompanhamento de sua disseminação e surtos epidêmicos como os que ocorreram em 2011 na Alemanha, ocasionado pela bactéria *Escherichia coli* (PAULING *et al*., 2012) e em 2014 na África, produzido pelo vírus ebola (GIRE *et al.*, 2014). Essa grande quantidade de informação tem auxiliado no desenvolvimento de vacinas e contribuído com o diagnóstico de doenças (LUCIANI *et al.*, 2012; LOMAN *et al.,* 2012; LECUIT & ELOIT, 2015).

1.2. Sequenciamento de genomas

As tecnologias de sequenciamento de genomas conhecidas na literatura como primeira geração são os métodos de Sanger e o de Maxam e Gilbert, sendo que o mais utilizado é o de Sanger (LIU *et al*., 2012). Este último consiste no sequenciamento de genomas pelo método de terminação de cadeia (SANGER *et al.*, 1977; COSTA, 2015).

Buscando o aperfeiçoamento destas técnicas de sequenciamento, em 1987 foi lançado o primeiro sequenciador automático conhecido como AB370, que foi comercializado pela Applied Biosystems, utilizando eletroforese por

capilaridade (LIU *et al*., 2012). Este sendo muito utilizado nos anos seguintes propiciou o início da chamada corrida do sequenciamento genômico, sendo utilizado inclusive no projeto genoma humano (LANDER *et al.*, 2001; VENTER *et al.*, 2001).

Mesmo com o surgimento dos sequenciadores automáticos, como o AB370, ainda era necessário que os insertos fossem clonados em vetores de clonagem, o que era laborioso e exigia mão de obra qualificada. Estas desvantagens impulsionaram o surgimento das tecnologias NGS, nas quais esta etapa é substituída pela produção de uma biblioteca. Nesta, adaptadores são inseridos nas extremidades dos fragmentos de DNA e a amplificação destes ocorre através de PCR em emulsão ou em fase sólida (MARDIS, 2011).

A partir de 2004 intensificou-se o desenvolvimento das tecnologias NGS, a introdução destas tecnologias reduziu o custo do sequenciamento e aumentou a capacidade de produção de dados por corrida (KAUR & MALIK, 2013), permitindo assim que um genoma bacteriano fosse sequenciado em questão de horas ou poucos dias (LOMAN *et al.,* 2012).

O custo do sequenciamento do primeiro projeto genoma humano foi de aproximadamente 3 bilhões de dólares (VINCENT *et al.,* 2016). Devido ao alto custo deste projeto, o *National Institutes of Health* (NIH) passou a incentivar o desenvolvimento de tecnologias para reduzir o custo do sequenciamento (MARDIS, 2011). O gráfico representado na Figura 2 apresenta a variação do custo do sequenciamento do genoma humano, no período compreendido entre 2001 e 2015, sendo que atualmente o custo está em torno de mil dólares.

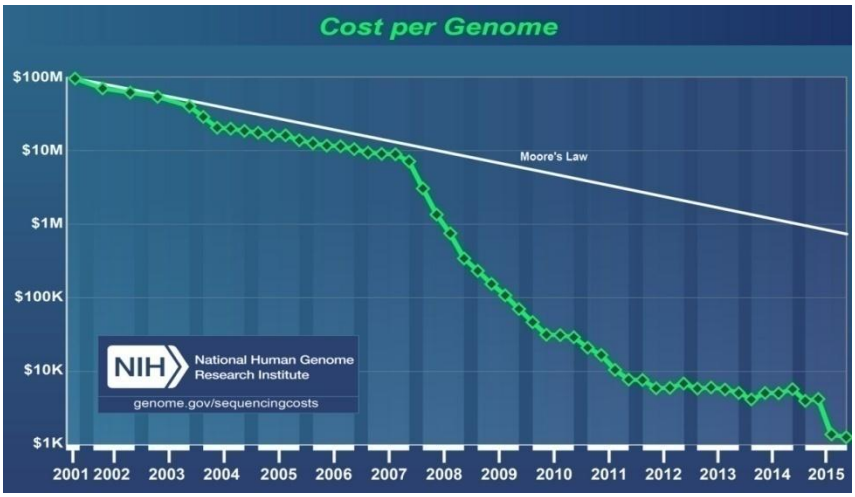


Figura 2 – Custo do sequenciamento de genomas no período de 2001 a 2015.

Fonte: National Human Genome Research Institute (2016).

As tecnologias NGS têm contribuído para a realização de diferentes estudos dentro da biologia, tais como: sequenciamento de genoma completo (WGS), ressequenciamento genômico, metagenômica, desenvolvimento de marcadores (geralmente genes alvos para o tratamento de doenças/infecções), transcriptômica, análises epigenéticas (regulações celulares) e estudos de interação proteína-DNA (KAUR & MALIK, 2013).

De modo geral, o processo de sequenciamento das tecnologias NGS pode ser compreendido em três etapas principais: (i) extração e purificação do DNA; (ii) construção de bibliotecas; e (iii) sequenciamento. Primeiramente o DNA é extraído e purificado, para posteriormente passar pela etapa de construção de bibliotecas, na qual este é fragmentado de forma aleatória (biblioteca de fragmentos) ou através de bibliotecas pareadas em um processo que pode ser mecânico ou enzimático (HUSEMANN, 2011; BERGLUND *et al*., 2011; MARIANO, 2015). As bibliotecas pareadas podem ser divididas em *paired-end* e *mate-pair*. Na *paired-end* são colocados adaptadores nas duas extremidades de um curto fragmento de DNA. Na *mate-pair,* apesar dos adaptadores também serem colocados nas duas extremidades, os fragmentos são maiores, e então o fragmento é circularizado e cortado ao meio. Em ambas, a distância entre os adaptadores é conhecida, o que facilita o alinhamento dos fragmentos na montagem *ab initio* de genomas quando comparado as bibliotecas de fragmentos (MARDIS, 2013). Após esta etapa,

adaptadores são ligados em ambas extremidades dos fragmentos para então passar pelo processo de amplificação por reação em cadeia da polimerase (PCR) (nos sequenciadores que necessitam desta etapa), e então estes são sequenciados (BERGLUND *et al*., 2011; MARIANO, 2015).

O primeiro sequenciador com tecnologia NGS a ser lançado foi o 454 da Life Sciences (Roche) em 2005, que utiliza o método de pirosequenciamento. Esse método é capaz de detectar o pirofosfato liberado no momento da incorporação do nucleotídeo (LIU *et al*., 2012). Um ano após o lançamento do 454, a Solexa lança o seu sequenciador, que foi adquirido pela Illumina no ano seguinte, o qual utiliza o sequenciamento por síntese. Essa metodologia utiliza nucleotídeos marcados por fluoróforos que são inseridos a uma *flowcell* que contém adaptadores fixos, e uma DNA polimerase. Os nucleotídeos são terminadores reversíveis, o que garante que apenas um nucleotídeo seja inserido por vez na fita crescente de DNA prevenindo assim problemas com homopolímeros. A próxima tecnologia de sequenciamento foi o SOLiD, o qual utiliza a detecção por ligação de oligonucleotídeos e um código de cores para identificar as bases desenvolvido pela Applied Biosystems em 2007 (atualmente pertence à Thermo Fisher Scientific) (VAN DIJK *et al.*, 2014).

Em 2010, surge o PacBio RS (Pacific Biosciences) que realiza o sequenciamento pelo processo conhecido como *Single Molecule, Real-Time* (SMRT) (PACBIO SYSTEMS, 2016). No ano de 2015 o Oxford Nanopore Technologies MinION foi lançado, este tem o tamanho de um *pendrive* o que possibilita ser facilmente carregado para fazer o sequenciamento do DNA genômico no campo ou em qualquer lugar no momento da coleta do mesmo (OXFORD NANOPORE TECHNOLOGIES, 2016). O grande diferencial dessas tecnologias é a capacidade de sequenciamento por molécula única e o tamanho das *reads* (VAN DIJK *et al.*, 2014; PACBIO SYSTEMS, 2016; OXFORD NANOPORE TECHNOLOGIES, 2016).

Com o desenvolvimento das tecnologias NGS surgiram os chamados sequenciadores de bancada. Como exemplos, podemos citar o 454 GS Junior (Roche), o Ion PGM™ (Thermo Fisher Scientific) e o MiSeq (Illumina). Estes diminuíram os custos, o tempo do sequenciamento e são mais simples na utilização quando comparado aos sequenciadores convencionais, porém com menor *throughput* (LOMAN *et al.,* 2012).

1.2.1. Ion Torrent

Em 2010 surge no mercado o Ion PGM™, atualmente pertencente à Thermo Fisher Scientific, o qual foi desenvolvido por Jonathan Rothberg, que já havia sido responsável pelo desenvolvimento do 454 (VAN DIJK *et al.*, 2014). O processo de sequenciamento inicia-se pela extração e fragmentação do DNA alvo, sendo posteriormente inseridos adaptadores para serem incorporados aos fragmentos de DNA em *beads* magnéticas e estes amplificados através de PCR. A tecnologia semicondutora é utilizada na etapa de leitura dos dados, esta detecta mudanças no pH a cada vez que a DNA polimerase incorpora um nucleotídeo à fita crescente de DNA (KAUR & MALIK, 2013; HEAD *et al*., 2014). Na versão original o Ion PGM™ gerava até 270 Mega bases (Mb) com comprimento de *read* de até 100 nucleotídeos (nt) (VAN DIJK *et al.*, 2014). Atualmente existem três *chips* de sequenciamento do Ion PGM™ como pode ser visto na Tabela 1.

Tabela 1 – Especificações dos chips Ion PGM™.

Chip	Tempo de sequenciamento (horas)		*Throughput*		Nº de *reads* (em milhões)
	200 pb*	400 pb	200 pb	400 pb	
Ion 314™ Chip v2	2,3	3,7	30 – 50 Mb	60 – 100 Mb	0,4 – 0,55
Ion 316™ Chip v2	3	4,9	300 – 600 Mb	600 Mb – 1 Gb	2 – 3
Ion 318™ Chip v2	4,4	7,3	600 Mb – 1 Gb	1,2 – 2 Gb	4 – 5,5

*pares de base (pb)

Fonte: adaptado de: THERMO FISHER SCIENTIFIC (2016).

Posteriormente surgiram outros sequenciadores desta linha que são o Ion Proton™, o Ion S5 e Ion S5 XL. O *chip* do Ion Proton™ pode gerar um *throughput* de até 10 Gb, com comprimento de *read* de 200 nt, o número de *reads* pode chegar a 80 milhões com tempo de sequenciamento de 2 a 4 horas. Já os sequenciadores Ion S5 e Ion S5 XL, novos sequenciadores da Thermo Fisher Scientific, podem chegar a 15 Gb de *throughput*, com *reads* de 200 e 400 nt de comprimento e tempo de corrida de 2,5 a 4 horas (THERMO FISHER SCIENTIFIC, 2016). As principais vantagens dos sequenciadores Ion

Torrent são: utiliza tecnologia semicondutora e apresenta baixo tempo de sequenciamento. Como desvantagens pode-se destacar a alta taxa de erros com homopolímeros, e o Ion PGM™ tem uma taxa de erros total de 1,71% e a taxa erros de inserção e deleção de nucleotídeos durante o sequenciamento (*indels)* de 1,5% e uma queda da acurácia nas extremidades das *reads* (VAN DIJK *et al.*, 2014; MARINIER *et al.*, 2015).

A Tabela 2 traz um comparativo resumido das características de cada tecnologia de sequenciamento.

Tabela 2 – Comparação entre tecnologias de sequenciamento.

Tecnologia/ Sequenciador	Química	*Throughput* (Gb)	Tamanho da *read* (pb)	Vantagens	Desvantagens
Sanger / ABI3730	Método de terminação de cadeia	0,0003	1 Kb	Acurácia e tamanho das *reads*	Custo e *throughput*
454	Pirosequen-ciamento	0,7	1 Kb	Tamanho das *reads* e tempo	Custo, *throughput* e alta taxa de homopolímeros
Illumina	Terminação reversível	1800	300	Acurácia e *throughput*	Tempo, tamanho das *reads* (*HiSeq* e *NextSeq*) e custo inicial (*HiSeq*)
SOLiD	Ligação	320	75	Acurácia e *throughput*	Tamanho das *reads* e tempo
Ion Torrent	Detecção de próton	15	400	Tamanho das *reads*, *throughput* e tempo	Alta taxa de homopolímeros e taxa de erros total de 1,71%
PacBio RS	Sequencia-mento em tempo real	0,5 (3 por dia)	~20 Kb	Tamanho das *reads* e tempo	Custo, taxa de erros e *throughput*
MinION	Nanopore	1	~10 Kb	Tamanho das *reads*, tempo e portável	Alta taxa de erros e *throughput*

Fonte: adaptado de: (LOMAN *et al.*, 2012; VINCENT *et al.*, 2016; AMBARDAR *et al*., 2016; CHIU & MILLER, 2016).

1.3. Montagem de genomas

Após a etapa de sequenciamento do genoma, chega-se na parte conhecida como montagem de genomas, a qual utiliza diferentes abordagens para manuseio dos dados. Uma analogia muito utilizada na literatura para

exemplificar a montagem é o quebra-cabeça, onde a pessoa que vai realizar o trabalho tem uma figura recortada em vários pedaços e então precisa encaixar as peças, uma por uma, para obter a figura inteira. No caso da montagem de genomas há milhares de pedaços (*reads*), sendo que muitos destes são repetidos, e é necessário encontrar a melhor maneira de encaixar os mesmos para obter o genoma finalizado (POP, 2009).

Metodologias de montagem de genomas de dados provenientes do método de Sanger consistem no alinhamento entre as próprias *reads*, buscando a construção dos *contigs* e repetindo esta etapa até o genoma estar completo. Porém, com as metodologias NGS novos desafios surgiram: o tamanho das *reads* diminuiu, aumentou-se o *throughput* e também a taxa de erros de sequenciamento. Além disso, o tamanho reduzido das *reads* dificultou a resolução de regiões repetitivas (LIU *et al*., 2013; VERLI *et al.*, 2014; CHIN *et al.*, 2014). Outro ponto observado por Chen e colaboradores (2013) é o viés de conteúdo de Guaninas e Citosinas (GC) na montagem dos genomas, sendo este explicado por regiões com conteúdo GC diferenciado e baixa cobertura.

A correta resolução destes vieses é importante, pois genomas incompletos ou com erros de montagem podem afetar os resultados dos estudos que se seguem à montagem como a anotação, genômica comparativa e estudos funcionais. Para tal fim, diferentes abordagens *in silico* e experimentais têm sido utilizadas, como aumentar a profundidade da cobertura do sequenciamento (número de vezes, em média, que cada base é representada), produção de *reads mate-pair* e *paired-end*, além da utilização de mapeamento óptico (SIMS *et al.*, 2014; PURANIK *et al*., 2015; MARIANO, 2015).

Uma ferramenta que combina muitas das abordagens *in silico* utilizadas para resolver os vieses do NGS é a plataforma SIMBA[1], a qual é gratuita e de código aberto. Ela é disponibilizada através de uma interface *web* o que a torna de fácil operação quando comparada com outros *pipelines* dependentes de uma interface de linha de comando. O SIMBA pode receber como dados de entrada *single-reads*, *paired-end*, *mate-pair* ou mesmo dados de mapa óptico para ordenação dos dados (MARIANO, 2015).

[1] Disponível em: < http://ufmg-simba.sourceforge.net>. Acesso em: 21 de outubro, 2016.

Para realizar a montagem dos genomas o SIMBA utiliza 4 diferentes montadores, o Mira (CHEVREUX *et al.* 1999; CHEVREUX *et al*. 2004), o Newbler[2], o Minia (CHIKHI & RIZK, 2013) e o SPAdes (BANKEVICH *et al*., 2012).

Os montadores Mira e Newbler são baseados na abordagem de grafos OLC, a qual possui três etapas: (i) *Overlap* - é a etapa na qual as *reads* são identificadas e sobrepostas, (ii) *Layout* – trata da construção do grafo que interliga as sobreposições encontradas na primeira etapa, e (iii) *Consensus* – busca encontrar um caminho único que possa ligar todos os nós do grafo. Porém, devido à complexidade dos organismos, aos erros de sequenciamento e ao fato de normalmente haver regiões repetitivas, encontrar um caminho único torna-se uma tarefa difícil (POP, 2009; MILLER *et al.*, 2010; SIMPSON & POP, 2015). Os montadores que utilizam essa abordagem trabalham melhor com *reads* maiores que 200 pb, como as provenientes de Sanger, 454 ou Ion Torrent, mesmo que, como no caso dos dois últimos, estas sejam de menor acurácia (NAGARAJAN & POP, 2013; MARIANO, 2015). Além disso, tanto o Newbler quanto o Mira são otimizados para solucionar problemas de homopolímeros (MILLER *et al*., 2010, MARIANO, 2015).

Já os montadores Minia e SPAdes são baseados em grafos De Bruijn. Grafos De Bruijn dividem as *reads* em *substrings* (*k-mers*) de tamanho k, onde o k pode ser determinado pelo usuário ou determinado por algum *software*, e estas *substrings* são representadas pelos nós do grafo. Em tais *substrings* busca-se alinhar suas últimas k-1 bases com outras *substrings*, sendo que as sobreposições encontradas são representadas por arestas. Desta forma é gerado um grafo que representa a sequência do genoma. Estes são indicados para sequenciamento de *reads* curtas de alta acurácia, como dados provenientes de Illumina e SOLiD (MILLER *et al.*, 2010; NAGARAJAN & POP, 2013; SIMPSON & POP, 2015).

Outra ferramenta que pode ser usada para montagem de genomas é o CLC Workbench da QIAGEN Company, o qual é uma solução comercial. O CLC Workbench é uma ferramenta desenvolvida para trabalhar com *reads* provenientes de NGS, realizar o mapeamento destas *reads* e obter a fita

[2] Disponível em: <http://www.454.com/products/analysis-software/>. Acesso em: 26 de abril, 2016.

consenso (MILLER *et al.*, 2010; WAHEED *et al.*, 2012), podendo ser usado na chamada montagem por referência (MARIANO, 2015). Além disso, é possível realizar o fechamento de *gaps* e verificar os *contigs* que foram gerados e não estão incluídos na montagem (MARIANO, 2015; CLCBIO, 2016).

1.4. Anotação de genomas

Com o sequenciamento realizado e o genoma montado é necessário encontrar e dar um significado biológico aos genes contidos nesse genoma, processos chamados de predição e anotação genômica. O estudo do código genético objetiva compreender os diferentes processos que ocorrem em determinado organismo. Dentre eles podemos destacar a descoberta de quais genes, vias metabólicas e mecanismos de infecção estão presentes no genoma deste organismo. Stein (2001) divide o processo de anotação em 3 etapas: (i) localizar e caracterizar regiões dos genomas já conhecidas em outros organismos (*i.e. Open Reading Frames* (ORFs), ncRNAs (RNAs não codificantes), polimorfismos de nucleotídeos únicos (SNPs)), (ii) catalogar as proteínas e (iii) realizar a anotação funcional, onde as proteínas são comparadas com outros genomas procurando funções conhecidas destas (STEIN, 2001; STOTHARD & WISHART, 2006). A anotação deve ser feita cuidadosamente, uma vez que anotações incompletas ou com erros (de anotação estrutural ou mesmo funcional) acabam sendo propagados através dos bancos de dados, nos quais estes dados são depositados (MARKOWITZ *et al.*, 2009).

Com o objetivo de realizar a anotação de genomas de procariotos, de fagos e de plasmídeos de maneira rápida, acurada e automatizada, em 2008 foi desenvolvido o *pipeline* do *Rapid Annotation using Subsystem Technology* (RAST) (OVERBEEK *et al*., 2014). O *pipeline* utiliza como base o SEED, que é uma plataforma para predição de função gênica e vias através de FIGfams e suas coleções funcionais (subsistemas). O RAST identifica e atribui função aos genes, prediz os RNAs e pode reconstruir a rede metabólica (OVERBEEK *et al*., 2014; BRETTIN *et al*., 2015).

Porém, a acurácia da anotação automática não depende somente da qualidade da ferramenta utilizada, mas também da qualidade dos dados de

sequenciamento e da montagem. Dessa forma, após o RAST gerar a anotação automática do genoma, uma etapa de curadoria manual pode aumentar a precisão desta anotação (MÉDIGUE & MOSZER, 2007).

1.5. Genômica comparativa

Estes estudos iniciaram-se a partir do primeiro genoma bacteriano sequenciado em 1995 (FLEISCHMANN *et al*., 1995; COSTA, 2015). A descoberta de quais genes estão presentes no genoma de determinado organismo possibilita análises mais complexas, como as de genômica comparativa. Esta caracteriza-se pelo alinhamento de um conjunto de genes ou do genoma do organismo estudado com um conjunto de genes ou genomas de outros organismos buscando compreender aspectos relacionados a divergência evolutiva dos organismos (COSTA, 2015).

1.5.1. Pangenômica

O pangenoma define o conjunto total dos diferentes genes que compõe o repertorio gênico de um conjunto de organismos bacterianos (organismos de interesse) (TETTELIN *et al*. 2005; VERNIKOS *et al*., 2015). O pangenoma é constituído pelo genoma central, que é composto por genes presentes em todos os organismos analisados; o genoma acessório, composto por genes que estão em pelo menos dois organismos, mas não em todos que estão sendo analisados; e os genes únicos, os quais estão presentes em um único organismo dentre todos os analisados (TETTELIN *et al*. 2005; SOARES *et al.*, 2013).

No pangenoma os genes são agrupados em conjuntos chamados *clusters*, de acordo com a similaridade encontrada no alinhamento destes. Estes *clusters* são formados por grupos de genes ortólogos, parálogos e únicos. Como ilustrado na Figura 3, ortólogos são genes que se originaram a partir de um evento de especiação de um gene ancestral comum, geralmente mantendo a função gênica; parálogos são genes que sofreram um evento de duplicação gênica; e, genes únicos são aqueles que são exclusivos de um

genoma (TETTELIN *et al*. 2005; SOARES *et al.*, 2013; RICHARDSON & WATSON, 2013).

O pangenoma pode ser considerado aberto ou fechado. Este é considerado aberto quando novos genes são adicionados ao pangenoma a cada novo genoma incluído nas análises, e é considerado fechado se o fato de adicionar novos genomas não adiciona um número substancial de novos genes no pangenoma sendo estudado (Tettelin *et al*., 2005).

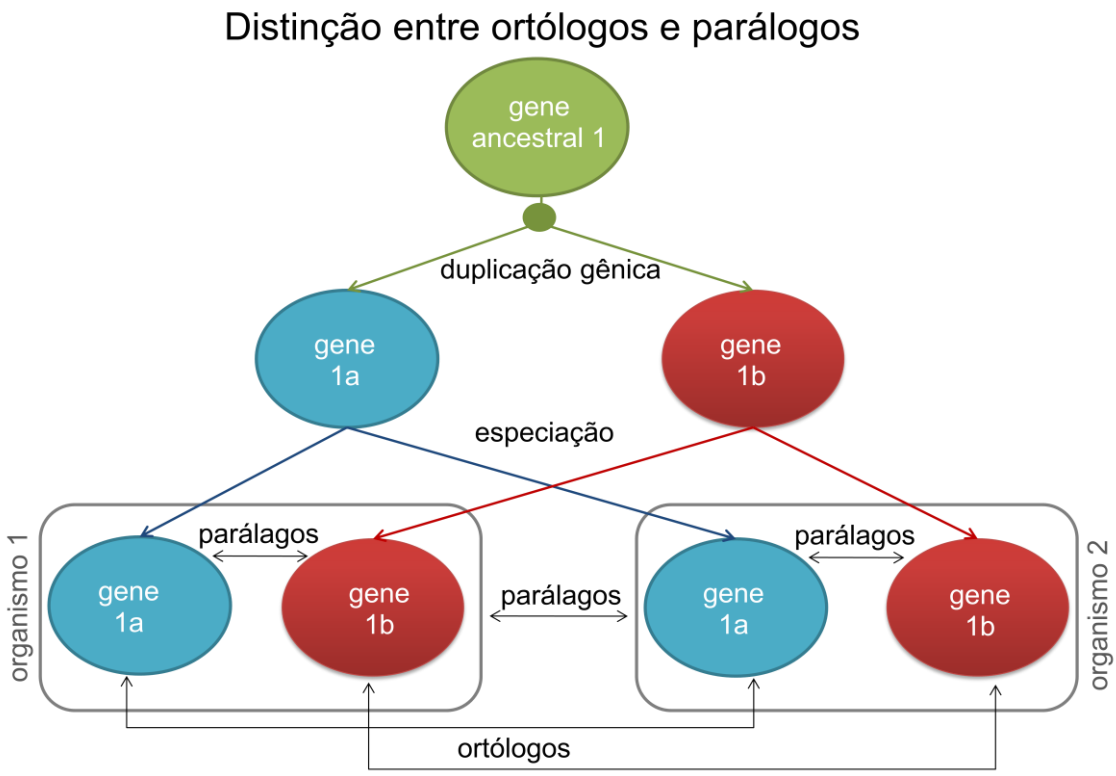


Figura 3 – Diagrama ilustrativo dos conceitos de ortologia e paralogia.

Relacionamentos evolutivos entre genes podem ser criados através de eventos de duplicação gênica e especiação.

Fonte: adaptado de (RICHARDSON & WATSON, 2013)

Para realização de análises de pangenoma é possível utilizar a ferramenta *Protein Family Sorter* do *Pathosystems Resource Integration Center* (PATRIC[3]) (SNYDER *et al*., 2007; WATTAM *et al*., 2013). O PATRIC é um centro de recursos de bioinformática focado em bactérias que armazena, integra e disponibiliza informações de diversas ômicas como a genômica,

[3] Disponível em: <https://www.patricbrc.org/portal/portal/patric/Home>. Acesso em: 14 de outubro, 2016.

transcriptômica, proteômica e seus metadados. Além disso, fornece ferramentas on-line para análise destes dados. O *Protein Family Sorter* realiza a clusterização das famílias proteicas relativas aos genes presentes nos genomas de interesse, através da análise dos genes que compõem cada genoma, analisando e comparando com as famílias proteicas previamente armazenadas em seu banco de dados. Existem três tipos de famílias proteicas no *Protein Family Sorter* que são: (i) PLfams (famílias proteicas específicas de gênero do PATRIC); (ii) PGfams (famílias proteicas através de gênero do PATRIC) e; (iii) FIGFams (famílias protéicas construídas através de anotação curada manualmente a partir da literatura pela equipe do PATRIC) (MEYER *et al*., 2009; WATTAM *et al*., 2013; DAVIS *et al*., 2016).

1.5.2. Métodos de visualização de genomas

Ferramentas que possibilitam a visualização de comparações genômicas podem contribuir em estudos de genômica comparativa, pois possibilitam determinar diferenças genotípicas entre organismos próximos. Essas ferramentas usam dois tipos de métodos para visualização os circulares ou os lineares (ALIKHAN *et al*., 2011).

Dentre as ferramentas que utilizam métodos de visualização circular pode-se destacar o BRIG (BLAST *Ring Image Generator*) (ALIKHAN *et al*., 2011), o qual compara os genomas de múltiplos procariotos e gera uma imagem circular apresentando a similaridade entre o genoma de referência selecionado e os demais genomas inseridos para realizar a comparação. Trata-se de uma ferramenta multiplataforma de fácil usabilidade através de sua interface gráfica. O mesmo utiliza o BLAST (ALTSCHUL *et al.*, 1990) para realizar as comparações genômicas e o CGView (STOTHARD & WISHART, 2005) para gerar a imagem (ALIKHAN *et al*., 2011).

1.5.3. Filogenômica

A filogenômica consiste na análise comparativa em escala genômica para o estudo dos relacionamentos evolutivos entre espécies buscando uma

melhor compreensão dos mecanismos de evolução molecular (EISEN, 1998; PHILIPPE *et al*., 2005; CHAN & RAGAN, 2013).

Geralmente, utilizar o genoma inteiro ou mesmo uma grande quantidade de genes torna a filogenia mais confiável do que utilizando um ou poucos genes. Assim pode-se obter informação detalhada das diferenças entre os genomas e não somente uma classificação geral destes (ÅGREN *et al*., 2012).

Para realizar a filogenômica de organismos é possível utilizar o *software* Gegenees (ÅGREN *et al*., 2012), que pode ser executado em múltiplas plataformas (Linux, Windows e MAC OS). Este fragmenta os genomas e busca realizar o alinhamento destes fragmentos utilizando o BLAST+ (CAMACHO *et al*., 2009) produzindo assim alinhamentos locais. O Gegenees representa os alinhamentos por um *score*, o qual é inserido em uma matriz, e também utiliza múltiplas *threads* para realizar os alinhamentos que posteriormente são tratados como um banco de dados, assim, ao adicionar um novo genoma não é necessário realizar o alinhamento inteiro novamente. Com os alinhamentos prontos é possível exportar a matriz de distância em um arquivo Nexus e utilizar a ferramenta Figtree (PYBUS *et al*., 2000; SUCHARD & RAMBAUT, 2009) para gerar o dendograma (árvore filogenética/filogenômica) (ÅGREN *et al*., 2012).

1.5.4. Plasticidade genômica

A capacidade de adaptação bacteriana pode influenciar no estilo de vida bacteriano, por exemplo, organismos patogênicos podem ter genomas menores que organismos não patogênicos devido a sua habilidade de utilizar metabólitos produzidos pelo hospedeiro (OCHMAN & DAVALOS, 2006; BARBOSA *et al*., 2014). Estudos de genômica comparativa têm auxiliado para uma melhor compreensão de características como a patogenicidade e também de evolução bacteriana. A evolução bacteriana está relacionada à plasticidade destes genomas, a qual pode ser caracterizada por mecanismos de transferência horizontal de genes, pontos de mutação e rearranjos genômicos (ex:. translocação e inversão) (DOBRINDT & HACKER, 2001).

Uma ferramenta que auxilia no estudo da plasticidade genômica é o GIPSy (SOARES *et al*., 2015), a qual identifica ilhas de patogenicidade (PAIs);

ilhas metabólicas (MIs), ilhas de resistência (RIs) e ilhas simbióticas (SIs). Estas são conhecidas como ilhas genômicas (GEIs), as quais são formadas pela incorporação de grandes blocos de sequências genômicas através da transferência horizontal de genes (SOARES *et al*., 2015). PAIs são ilhas que possuem genes de fatores de virulência (DOBRINDT *et al*., 2000), MIs são ilhas que possuem genes associados à biossíntese de metabólicos (TUMAPA *et al*., 2008), SIs que são ilhas que possuem genes que contribuem para um relacionamento simbiótico com o hospedeiro (BARCELLOS *et al*., 2007) e RIs são ilhas que possuem genes de fatores de resistência (KRIZOVA & NEMEC, 2010). O GIPSy identifica as GEIs através de um *pipeline* que analisa informações como o uso de códon, desvio de conteúdo GC, RNAs transportadores (tRNAs) flanqueadores, genes de transposase e fatores específicos de cada GEI (virulência, metabolismo, simbiose, ou resistência) (SOARES *et al*., 2015).

1.5.5. Alvos para drogas

Após realizar as análises comparativas, pode-se buscar por alvos para vacina e medicamentos, terapia, diagnóstico, fatores de virulência (BARH *et al*., 2011b; TROST *et al*., 2011). Abordagens utilizadas na descoberta de alvos para drogas costumam buscar por genes que são essenciais para a sobrevivência do patógeno e que não sejam homólogos aos seus hospedeiros (BARH *et al*., 2011a; BARH *et al*., 2011b). Essas abordagens computacionais possibilitam a descoberta de possíveis alvos de forma rápida e com maior custo benefício que abordagens não computacionais (BARH *et al*., 2011a).

Dentre os programas utilizados para este fim, podemos destacar o *Specialty Genes Search tool* (MAO *et al*., 2015) do PATRIC. Esta ferramenta objetiva fornecer alvos para drogas e genes envolvidos com fatores de virulência através da integração de bancos de dados de fatores de virulência, alvos para drogas, resistência antibiótica e de proteínas homólogas com humanos (MAO *et al*., 2015).

1.6. *Corynebacterium pseudotuberculosis*

Abordagens de genômica comparativa são muito utilizadas para o estudo de diversos organismos (SOARES *et al*., 2013; OLIVEIRA, 2014; COSTA, 2015). Neste trabalho o organismo de interesse é o patógeno bacteriano *Corynebacterium pseudotuberculosis* pertencente ao grupo CMNR (*Corynebacterium*, *Mycobacterium*, *Nocardia* e *Rhodococcus*) e à classe das Actinobactérias. O grupo CMNR é caracterizado pelo alto conteúdo GC (46 – 74%) e pela estrutura da parede celular, composta principalmente por peptidoglicanos, arabinogalactanos e ácidos micólicos. *C. pseudotuberculosis* é considerada um patógeno animal de importância veterinária que causa perdas financeiras ao agronegócio em diversos países (DORELLA *et al.*, 2006; TAUCH & SANDBOTE, 2014).

Como características morfológicas e bioquímicas a *C. pseudotuberculosis* é uma bactéria Gram-positiva, pleomórfica – podendo apresentar formas que variam de cocóides à bastões filamentosos – e é um organismo intracelular facultativo. Seu tamanho varia de 0,5 à 0,6 µm por 1 à 3 µm (DORELLA *et al.*, 2006; PINTO, 2011). Outras características apresentadas por este organismo é que possui fimbrias, é imóvel, não esporula, não possui cápsula e é anaeróbico facultativo (DORELLA *et al.*, 2006; GUEDES *et al.*, 2015).

Além disso, por se tratar de um organismo intracelular facultativo e ser capaz de sobreviver em diferentes condições ambientais pode sobreviver vários meses no solo (GUEDES *et al.*, 2015).

A transmissão ocorre através do contato do animal com superfícies contaminadas, ou de um animal infectado para outro com alguma lesão de pele e através de moscas, que também podem ser potenciais transmissores (DORELLA *et al.*, 2006; SOARES *et al.*, 2013; BARBA *et al.*, 2015).

As diferentes linhagens de *C. pseudotuberculosis* são divididas em dois biovares: o biovar Equi e o biovar Ovis. Estes se diferenciam pela capacidade de redução de nitrato das bactérias do biovar Equi, o que não ocorre nas bactérias do biovar Ovis. Outra diferença está no hospedeiro, sendo o primeiro normalmente encontrado em cavalos e bovinos e, o segundo em pequenos

ruminantes como ovinos e caprinos, porém os dois biovares podem afetar bovinos (BIBERSTEIN *et al*., 1971; SUTHERLAND *et al.*, 1996).

Entre os diferentes biovares, nota-se também uma diferença nos sintomas das doenças nos diferentes hospedeiros (camelídeos, bubalinos, equinos, ovinos, caprinos e bovinos), tendo em vista que *C. pseudotuberculosis* pode causar linfadenite caseosa (CLA), mastite, linfangite ulcerativa, peito de pombo e pode também afetar órgãos internos, além de linfadenite subaguda à crônica em humanos (PINTO, 2011; SOARES *et al.*, 2013; BARBA *et al*., 2015). Apesar de não ser frequente, a infecção em humanos ocorre devido ao contato com animais, locais infectados ou mesmo através do consumo de alimentos crus infectados. Alguns destes casos foram descritos por Peel *et al*. (1997) e por TROST *et al*. (2010). O diagnóstico pode ser feito de maneira acurada através da reação em cadeia da polimerase (PCR) multiplex (ALGAMMAL, 2016).

Mesmo com todos os estudos, diagnosticar este organismo ainda é um desafio para veterinários e outras autoridades em saúde (MUÑOZ-BUCIO *et al*., 2016), mesmo porque os sintomas variam entre biovares e hospedeiros. Assim, faz-se necessário a identificação de novos alvos de virulência e patogenicidade, e assim muitos estudos e testes experimentais ainda são necessários para encontrar um alvo universal (TIWARI *et al*., 2014; JEBER *et al*., 2016). A exotoxina fosfolipase D (PLD) é o fator determinante de virulência melhor caracterizado. Outro determinante de virulência melhor estudado são as proteínas codificadas pelos genes envolvidos na aquisição de ferro (*fagA, fagB, fagC e fagD*) (DORELLA *et al.*, 2006; TROST *et al.*, 2010).

Mesmo sendo o determinante de virulência da *C. pseudotuberculosis* melhor estudado, a exotoxina PLD possui propriedades que não são bem compreendidas (PINTO, 2011). Porém, estudos mostraram que a inativação do gene *pld* impede a disseminação no hospedeiro da *C. pseudotuberculosis*, contudo não impede a formação do abscesso. Essa capacidade de propagação da *C. pseudotuberculosis* se deve a capacidade da PLD hidrolisar lisofosfatidilcolina e esfingomielina no hospedeiro, o que permite a bactéria se mover através dos vasos linfáticos (TACHEDJIAN *et al.*, 1995; DORELLA *et al.*, 2006).

As proteínas codificadas pelo *cluster* gênico *fagABC* e *fagD*, o qual possivelmente auxilia na sobrevivência e virulência bacteriana em ambientes com baixa disponibilidade de ferro. A biossíntese de DNA e a respiração são exemplos de importantes funções celulares bacterianas nas quais o ferro age como um cofator para proteínas envolvidas nestas. Porém em excesso o ferro pode se tornar tóxico para a bactéria, assim a bactéria regula a nível transcricional os sistemas de captação de alta afinidade com o ferro (BILLINGTON *et al.*, 2002; TROST *et al.*, 2010).

O tratamento utilizando antibióticos é considerado prolongado e caro e os mesmos não penetram na cápsula formada como defesa e, portanto o tratamento com antibióticos não é recomendado (COLLETT *et. al.*, 1994; PINTO, 2011). O desenvolvimento de vacinas é algo que tem sido estudado, porém ainda não foi desenvolvido um método eficiente e, além disso, a espécie do hospedeiro pode também influenciar na efetividade das vacinas (SEYFFERT *et al.*, 2014).

Para se ter uma profilaxia efetiva no combate as doenças causadas pela *C. pseudotuberculosis* é necessário identificar novos alvos para drogas e vacinas que possam proteger e produzir uma resposta imune adequada no hospedeiro. Para tal, abordagens *in silico* como estudos comparativos são importantes estratégias para descoberta de novos alvos em potencial, os quais já podem ter sido testados em outros organismos, que possam ser posteriormente testados *in vivo* e *in vitro* (TIWARI *et al*., 2014).

Estudos de genômica estrutural e análise pangenômica (DE RESENDE, 2011; SOARES *et al*., 2013; DIAS, 2015), já demonstraram a importância dos estudos comparativos desta espécie. Além disso, estudos com novas linhagens podem contribuir, futuramente, nas análises de pangenoma, pois na análise realizada por Soares e colaboradores (2013) este é aberto, ou seja, novos genomas devem incluir mais genes ao pangenoma.

Assim, este trabalho visou à caracterização de seis novas linhagens de *C. pseudotuberculosis* provenientes do México, sendo a primeira vez que linhagens desta espécie provenientes deste país foram sequenciadas, permitindo, assim, uma melhor compreensão biológica do organismo, além de observar as diferenças e semelhanças genômicas dos mesmos.

2. JUSTIFICATIVA

Como pôde ser visto na secção anterior *C. pseudotuberculosis* é um importante patógeno médico veterinário que vêm causando perdas no agronegócio brasileiro e mundial. Perdas estas provenientes da redução da produção agroindustrial ou mesmo da morte dos animais afetados pelas doenças causadas por este organismo (DORELLA *et al.*, 2006). *C. pseudotuberculosis* também pode atingir humanos através do contato com animais ou mesmo por alimentos infectados (PEEL *et al.*, 1997; TROST *et al.*, 2010).

Apesar da importância deste patógeno, pouco se conhece sobre os mecanismos moleculares e os fatores de virulência associados a este organismo (TIWARI *et al*., 2014). Principalmente com relação ao biovar Equi (BARBA *et al*., 2015), no qual o número de casos vem aumentando significativamente nos últimos anos nos Estados Unidos (KILCOYNE *et al*., 2014). Mesmo com a crescente quantidade de estudos em relação a esse organismo, muitas vezes veterinários e outras autoridades em saúde pouco conhecem sobre o organismo e enfrentam dificuldades na identificação do mesmo (MUÑOZ-BUCIO *et al*., 2016). Adicionalmente, novos estudos são necessários para a compreensão de características desconhecidas deste organismo o que pode auxiliar os estudos com importantes questões em aberto, como por exemplo: como ocorre à transmissão, melhores práticas para tratamento e mesmo vacinas e medicamentos para prevenção (TIWARI *et al*., 2014; BARBA *et al*., 2015; MUÑOZ-BUCIO *et al*., 2016).

Buscando compreender melhor este organismo, este trabalho estuda seis linhagens de *C. pseudotuberculosis* provenientes do México (quatro pertencentes ao biovar Ovis e duas pertencentes ao biovar Equi). Estas são as primeiras linhagens de *C. pseudotuberculosis* isoladas neste país a serem sequenciadas. Além disso, é a primeira vez que *C. pseudotuberculosis* biovar Equi é isolada no México como pode ser visto em (MUÑOZ-BUCIO *et al*., 2016).

Desta forma, a caracterização destas linhagens se torna importante para a compreensão desta espécie, pois estas podem apresentar características relevantes e ainda não conhecidas em outras linhagens. Isto pode levar a

futuras descobertas de novos alvos para drogas ou novos fatores de virulência. Estes alvos podem contribuir com as pesquisas que abrangem este organismo e o tratamento desta zoonose, assim como diminuir as perdas econômicas no agronegócio mundial devido às doenças causadas pela *C. pseudotuberculosis*.

3. OBJETIVO PRINCIPAL

Realizar a análise de genômica comparativa de seis linhagens de *C. pseudotuberculosis* dos dois biovares isoladas no México.

3.1. Objetivos Específicos

Sequenciar, montar, anotar e depositar no *National Center for Biotechnology Information* (NCBI) as linhagens de *C. pseudotuberculosis* MEX1, MEX9, MEX25, MEX29, MEX30 e MEX31.

Identificar o pangenoma: genoma central, genoma acessório e os genes únicos dos novos isolados de *C. pseudotuberculosis* utilizando a ferramenta *Protein Family Sorter* do PATRIC e comparar os dois biovares de acordo com os resultados desta análise.

Executar a análise filogenômica entre as linhagens estudadas.

Investigar prováveis Ilhas de Patogenicidade (PAIs) dos genomas estudados.

Identificar possíveis genes alvos para drogas dentro do genoma central dos genomas estudados.

4. METODOLOGIA

Para uma melhor compreensão do desenvolvimento do trabalho a Figura 4 apresenta o desenho experimental do trabalho desenvolvido.

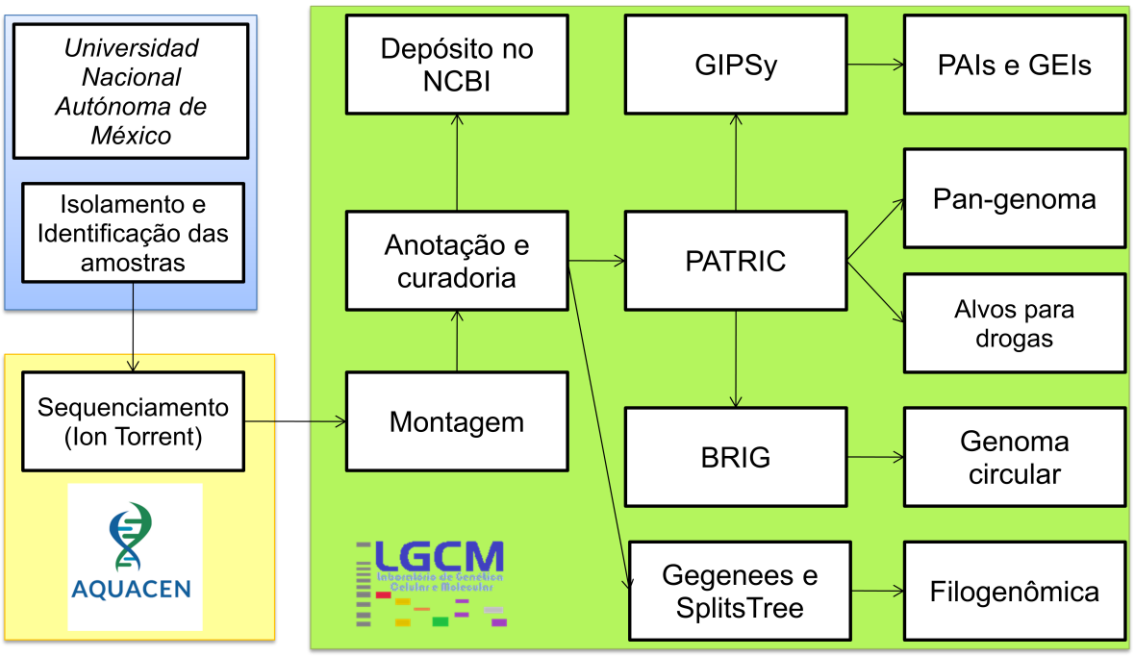


Figura 4 – Desenho experimental das etapas realizadas nesse trabalho.

Retângulo em azul apresenta a parte que foi realizada no México (isolamento e identificação das amostras). Retângulo em amarelo apresenta a parte que foi realizada no Laboratório AQUACEN (UFMG). Retângulo em verde apresenta a parte desenvolvida no LGCM.

4.1. Isolamento e identificação das seis linhagens de *C. pseudotuberculosis*

As seis linhagens de *C. pseudotuberculosis* (Tabela 3) utilizadas neste trabalho foram isoladas de cabras, ovelhas e cavalos no México, contemplando os dois biovares conhecidos para esta espécie, as quais foram identificadas através de PCR multiplex (PACHECO *et al.*, 2007; MUÑOZ-BUCIO *et al.*, 2016).

Tabela 3 – Linhagens de *C. pseudotuberculosis* provenientes do México que foram utilizadas neste trabalho e depositadas no NCBI.

Linhagem	BioProject e BioSample	Data de depósito	Hospedeiro	Material biológico	Origem (Cidade)	Ano de Isolamento	Biovar
Cp MEX1	PRJNA348354 SAMN05907792	10/2016	Cabra	retrofaríngea	Ixtenco	2014	Ovis
Cp MEX9	PRJNA312392 SAMN04497724	02/2016	Cabra	pré-escapular	Salamanca	2012	Ovis
Cp MEX25	PRJNA294672 SAMN04028217	09/2015	Ovelha	parótida	Guanajuato	2012	Ovis
Cp MEX29	PRJNA335634 SAMN05449619	08/2016	Ovelha	retrofaríngea	Río Frio	2013	Ovis
Cp MEX30	PRJNA343017 SAMN05771874	09/2016	Cavalo	músculo peitoral	Valparaiso	2013	Equi
Cp MEX31	PRJNA341961 SAMN05729795	09/2016	Cavalo	músculo peitoral	Valparaiso	2013	Equi

4.2. Extração do DNA e sequenciamento

Para realizar a extração do DNA, primeiramente as linhagens foram cultivadas em meio "*Brain Heart Infusion*" (BHI) suplementado com 0,5% de Tween 80, a 37°C por 72 horas, sob agitação e para culturas em meio sólido, foram adicionados 1,5% de ágar bacteriológico. E posteriormente, a extração de DNA foi realizada de acordo com Pacheco (2006).

Com o DNA extraído passou-se para a etapa de construção da biblioteca, a qual seguiu as recomendações do fabricante (IonXpress™ Plusg DNA Fragment Library Preparation). O primeiro passo foi a fragmentação do DNA genômico utilizando o kit Ion Shear™ Plus Reagents. O tamanho dos fragmentos de DNA gerados pode variar de tamanho por se tratar de um reagente tempo-dependente, por isso, utilizou-se o protocolo pré-estabelecido de 5 minutos a 37ºC.

Posteriormente, ligaram-se adaptadores (Ion Xpress™ Barcode Adapters) aos fragmentos de DNA gerados na etapa anterior, seguido de *nick-repair* para garantir a correta ligação entre os adaptadores e o inserto de DNA. Neste momento, para obter fragmentos do tamanho desejado selecionou-se a biblioteca mais adequada, etapa conhecida como *Size-selection*. Em sequência foi realizada a quantificação das bibliotecas em tempo real.

A próxima etapa foi a de amplificação da biblioteca, de acordo com as recomendações do fabricante (Ion PGM™ Template OT2 200 Kit), na qual realizou-se a amplificação dos fragmentos de DNA através de PCR em emulsão, utilizando-se partículas esféricas (Ion Sphere™ particles) e o sistema Ion One Touch™. No processo de enriquecimento, ocorreu a separação das partículas ligadas a fragmentos de DNA (partículas positivas) no Ion OneTouch™ ES. Na etapa seguinte o Control Ion Sphere™, o Sequencing Primer e o Ion PGM™ Sequencing Polymerase, este último para as linhagens de *C. pseudotuberculosis* MEX9, MEX25, MEX29, MEX30 e MEX31, ou o HiQ *Polymerase*, para a linhagem MEX1, foram adicionados às esferas positivas. Essa reação foi carregada no chip semicondutor (Ion 318 TM Chip v2) e este inserido no Ion PGM™.

4.3. Estratégia de Montagem

Para realizar a montagem destes genomas criou-se um projeto, no SIMBA, para cada um dos genomas sequenciados. Todos os genomas foram montados através da estratégia "*ab initio*" utilizando os *softwares* Mira versão 3.9 e Newbler versão 2.9. Antes de montar os genomas das linhagens MEX1, MEX30 e MEX 31, o SIMBA recebeu uma atualização e passou a contar com o montador SPAdes versão 3.6.0 e a ferramenta QUAST (*quality assessment tool*) (GUREVICH *et al.*, 2013), estes foram utilizados apenas nestes três genomas. Posteriormente foi selecionada a melhor montagem para cada genoma, considerando as características: número de *contigs*, tamanho do genoma, tamanho máximo e mínimo de *contig* e N50. O N50 é obtido ordenando-se os *contigs* em ordem de tamanho, do maior para o menor, e este corresponde ao *contig* no qual se atingiu pelo menos 50% do tamanho total do genoma. Essas características são apresentadas no SIMBA e obtidas através do *script python* CONTIGinfo[4].

Após realizar a montagem os *contigs* foram alinhados contra uma referência utilizando uma versão adaptada do programa CONTIGuator (GALARDINI *et al.*, 2011). A referência escolhida para cada genoma é apresentada na Tabela 4, e foi selecionada através de busca por similaridade utilizando o BLASTn no site do NCBI. O alinhamento no CONTIGuator gera três arquivos, um com os *contigs* que não foram alinhados à referência, chamado de *excluded.fsa*, um com *scaffolds* e um pdf contendo o gráfico de sintenia. Para que o gene *dnaA* possa estar na região OriC (região na qual a replicação se inicia em bactérias (SERNOVA & GELFAND, 2008)) utilizou-se o *script python in house moveDNA*[5]. Para analisar sobreposições entre *contigs* vizinhos (Figura 5A) foi utilizada a ferramenta BLAST (ALTSCHUL *et al.*, 1990) e posteriormente a região de sobreposição foi removida em um dos *contigs* (Figura 5B) e estes foram concatenados formando um único *contig* (Figura 5C). Todos os passos acima foram executados através da interface da ferramenta SIMBA.

[4] Disponível em: <https://github.com/dcbmariano/scripts>. Acesso em: 06 de maio, 2016.
[5] Disponível em: <https://github.com/dcbmariano/scripts>. Acesso em: 06 de maio, 2016.

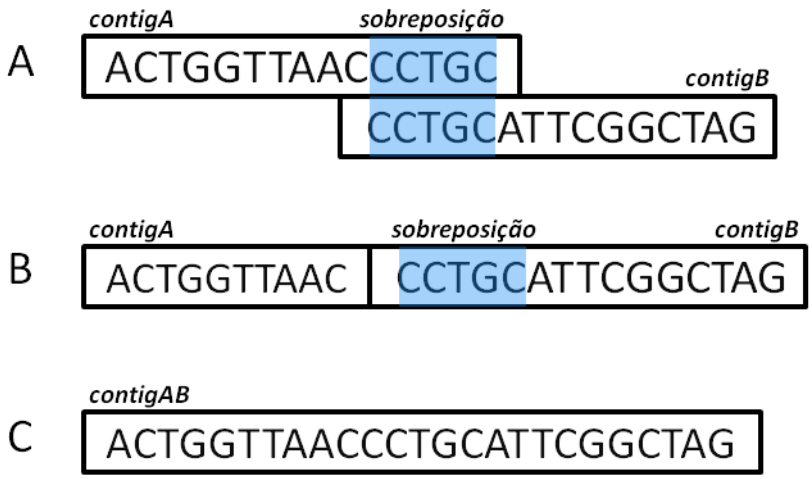


Figura 5 – Alinhamento e concatenação de dois *contigs* vizinhos e sobrepostos.

(A) Representação de dois *contigs* sobrepostos. (B) *contigA* depois de remover a região de sobreposição, com o *contigB* já alinhado. (C) *contigs* A e B concatenados e formando o *contigAB*.

Para realizar o fechamento dos *gaps* que não possuíam sobreposição utilizou-se a ferramenta CLC, na qual os dados brutos foram mapeados contra a mesma referência (Tabela 4), utilizada no CONTIGuator e foi extraído o consenso. Com base neste consenso e nos *contigs* contidos no arquivo *excluded.fsa* os genomas puderam ser fechados.

Tabela 4 – Referência utilizada para realizar o alinhamento e o fechamento de *gaps* para cada genoma.

	Referência			
Linhagem	Linhagem	Hospedeiro	País	Biovar
Cp MEX1	*Cp* MEX9	Caprino	México	Ovis
Cp MEX9	*Cp* 29156	Bovino	Israel	Ovis
Cp MEX25	*Cp* 29156	Bovino	Israel	Ovis
Cp MEX29	*Cp* 29156	Bovino	Israel	Ovis
Cp MEX30	*Cp* 316	Equino	EUA	Equi
Cp MEX31	*Cp* E19	Equino	Chile	Equi

4.4. Anotação

A anotação dos genomas foi realizada em três etapas: (i) anotação automática utilizando o *pipeline* RAST; (ii) transferência da anotação do genoma de referência; (iii) junção dos arquivos.

O primeiro passo foi executado utilizando-se o *pipeline* automático do RAST pela interface *web*, no qual os genomas no formato fasta foram submetidos. No segundo passo realizou-se a transferência de anotação para o novo genoma através do *script in house transfere.pl*. Para o terceiro passo utilizou-se o *software* Artemis (RUTHERFORD *et al.*, 2000; CARVER *et al.*, 2012) onde foi adicionado o arquivo gerado na montagem, o arquivo gerado pelo RAST e o arquivo gerado na transferência de anotação que foram importados dentro deste, sendo todos salvos em um arquivo único no formato gbk.

4.5. Curadoria

Para realizar a curadoria dos genomas utilizou-se o *software* Artemis para editar na sequência genômica os problemas encontrados nas CDSs. As principais correções realizadas foram: verificar e eliminar CDSs com sobreposição; CDSs com *start codon* (códon inicial) diferente de Metionina, Leucina ou Valina e/ou sem *stop* códon (códon de parada), ou mesmo em outra posição que não seja no término da CDS; e correção de *frameshifts* e da anotação destas CDSs.

Para iniciar o processo de curadoria o arquivo foi analisado em sua totalidade e as CDS sobrepostas foram apagadas com o objetivo de criar um consenso entre as anotações.

Com o objetivo de verificar as CDSs preditas e suas anotações, as mesmas foram verificadas utilizando o BLASTp do UNIPROT (UNIPROT CONSORTIUM *et al.*, 2011). Este possui anotações manuais revisadas (Swiss-Prot) e não revisadas (TrEMBL).

Para realizar a correção de *frameshifts* utilizou-se os *softwares* CLC e Artemis. O primeiro continha o mapeamento dos dados brutos (dados do sequenciamento de cada novo genoma) com o genoma final fechado no processo de montagem de cada um dos novos genomas. No Artemis, verificou-se as regiões de mudança de fase de leitura ao longo de cada CDS e no CLC utilizou-se o mapeamento dos dados brutos (dados do sequenciamento da linhagem correspondente) contra a fita montada para verificar a conservação nucleotídica naquela determinada posição. Após tal verificação, realizou-se a

inserção, deleção ou troca do nucleotídeo de acordo com a necessidade utilizando-se o Artemis. Posteriormente, a cada correção de *frameshift*, uma nova verificação no UNIPROT foi realizada com o objetivo de validar esse procedimento.

4.6. Depósito dos genomas

Todas as sequências nucleotídicas dos genomas estudados foram depositadas no NCBI, juntamente com a anotação conforme o protocolo do NCBI (COMPLETE GENOME SUBMISSION GUIDE, 2016).

4.7. Genômica Comparativa

4.7.1. Pangenoma

Para realizar as análises de pangenoma foi utilizado o centro de recursos de bioinformática PATRIC[6]. Neste, os genomas foram carregados e posteriormente anotados utilizando a ferramenta de anotação disponível em *SERVICES* > *GENOME ANNOTATION*. Após esta etapa, o pangenoma foi realizado utilizando o *Protein Family Sorter*. Este pode ser acessado em *TOOLS* > *Protein Family Sorter* na página do PATRIC como visto na Figura 6. Nesta etapa selecionaram-se os seis genomas aqui estudados e o tipo de família PLfams foi selecionado, como recomendado pelos autores para análises de pangenoma (Figura 7) (MAO *et al*., 2015). Estas análises foram realizadas para os seis genomas, e as mesmas análises foram realizadas separando estes organismos por biovar.

[6] Disponível em: <https://www.patricbrc.org>. Acesso em: 12 de outubro, 2016.

Figura 6 – Tela para acessar a ferramenta *Protein Family Sorter* no PATRIC.

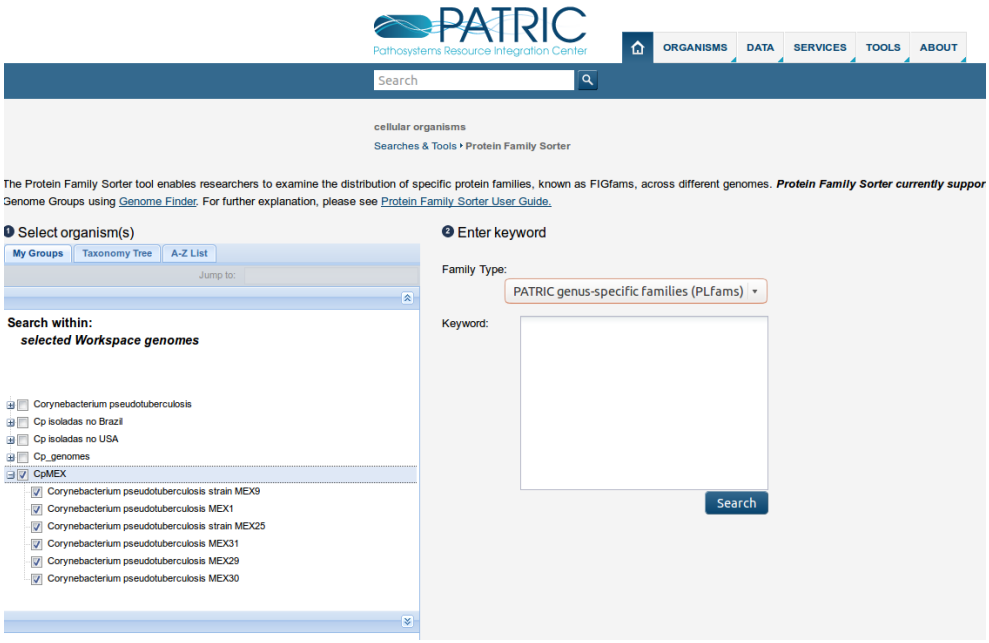


Figura 7 – Opções selecionadas no *Protein Family Sorter*.

4.7.2. Representação circular dos genomas

Para a visualização circular dos genomas foi utilizado o *software* BRIG (ALIKHAN *et al*., 2011). Para tal as sequências genômicas no formato genbank foram adicionadas ao mesmo. Por ter genomas dos dois biovares de *C. pseudotuberculosis*, gerou-se uma imagem com o genoma da linhagem MEX1 como referência para visualização das características do biovar Ovis e uma com a linhagem MEX30 como referência para visualização das características

do biovar Equi. Com a referência inserida, os demais genomas foram adicionados e posteriormente executou-se a comparação por BLASTn, e através dessa comparação gerou-se a imagem circular dos genomas analisados.

4.7.3. Filogenômica

Para as análises de filogenômica utilizou-se o *software* Gegenees (ÅGREN *et al*., 2012.), usando como entrada os arquivos .fna, os quais contém a sequência completa de nucleotídeos de cada genoma utilizado neste trabalho. Após o carregamento dos arquivos de entrada criou-se um projeto com os parâmetros *default* que são os recomendados pelos autores para análise com genomas bacterianos e utilizando-se o método de alinhamento BLASTn. Após esta etapa, o Gegenees realizou a fragmentação do genoma e fez alinhamento todos contra todos dos fragmentos gerados dos seis genomas. A partir destes alinhamentos o *software* criou uma imagem do gráfico *heatmap*. Os resultados foram exportados na extensão .nexus, a qual foi utilizada para criação da árvore no SeaView (GOUY *et al*., 2010), através do método de alinhamento *Neighbor-joining* (NJ), e o FigTree (PYBUS *et al*., 2000; SUCHARD & RAMBAUT, 2009) foi utilizado para edição da árvore.

4.7.4. Ilhas de patogenicidade

A análise de PAIs foi realizada utilizando o *software* Gipsy (SOARES *et al.*, 2015) para as seis linhagens de interesse. Esta análise é realizada para um genoma de cada vez, utilizando-se o gbk da linhagem de interesse e o gbk de uma linhagem próxima, porém não patogênica, que serve como referência na análise. Neste caso foi utilizado o organismo *Corynebacterium glutamicum* ATCC 13032, assim como utilizado por Soares e colaboradores (2013) em sua análise de ilhas de patogenicidade de *C. pseudotuberculosis*. Este software prediz as ilhas genômicas através de oito passos que identificam as seguintes características: desvio de assinatura genômica (conteúdo CG e uso de códons); presença de fatores de virulência, metabolismo, resistência à antibióticos e simbiose; genes de tRNAs flanqueadores; presença de genes

codificadores de transposases; e ausência das regiões identificadas como ilhas genômicas na linhagem de referência. Todos os parâmetros utilizados na análise são os recomendados pelos autores e toda a execução do programa ocorreu de acordo com o manual do mesmo.

4.7.5. Predição de alvos para drogas

As análises de alvos para drogas foram realizadas através do PATRIC, utilizando o *Specialty Genes Search*. Este pode ser acessado em *TOOLS* > *Specialty Genes Search* na página do PATRIC como pode ser observado na Figura 8. Para realizar a predição, foram selecionados: os seis genomas estudados neste trabalho, a opção alvos para drogas e busca por BLASTp e literatura como pode ser observado na Figura 9.

Figura 8 – Tela para acessar a ferramenta *Specialty Genes Search* no PATRIC.

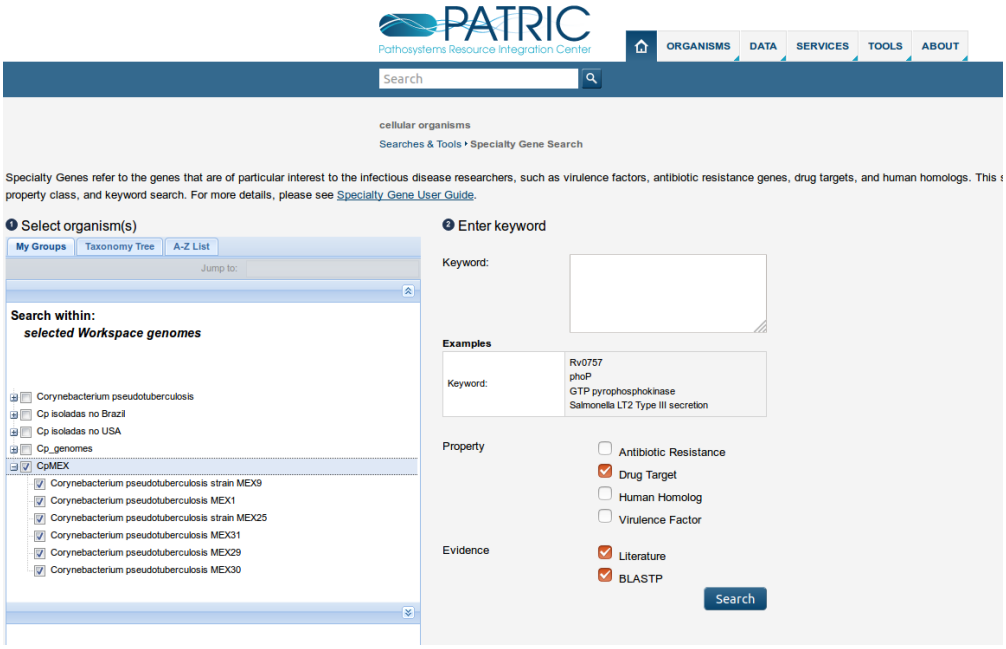


Figura 9 – Opções selecionadas no *Specialty Genes Search*.

5. RESULTADOS E DISCUSSÃO

5.1. Sequenciamento

O sequenciamento gerou os dados apresentados na Tabela 5. Este gerou coberturas médias que variaram de 80,61 vezes a 135,46 vezes o tamanho do genoma o que foi suficiente para montar os genomas pela estratégia *"ab initio"*.

Tabela 5 – Dados provenientes do sequenciamento através da plataforma Ion Torrent com biblioteca *fragment* de 400 pb.

Linhagem	Total de bases	Cobertura	Valor de *Phred*
Cp MEX1	269.476.555	115,30x	27
Cp MEX9	301.599.842	129,03x	27
Cp MEX25	231.238.726	98,92x	23
Cp MEX29	316.695.111	135,46x	27
Cp MEX30	190.895.766	80,61x	23
Cp MEX31	290.476.559	122,67x	27

Os dados de todos os seis genomas apresentaram qualidade *Phred* média variando de 23 (*C. pseudotuberculosis* MEX25 e *C. pseudotuberculosis* MEX30) a 27 (*C. pseudotuberculosis* MEX1, *C. pseudotuberculosis* MEX9, *C. pseudotuberculosis* MEX29 e *C. pseudotuberculosis* MEX31) como pode ser visto na Tabela 5.

5.2. Montagem

Os resultados de montagem gerados pelos montadores Newbler, Mira e SPAdes para cada uma das linhagens de *C. pseudotuberculosis* isoladas no México estão apresentados na Tabela 6.

Para montar a *C. pseudotuberculosis* MEX9 foram executadas duas tentativas. A montagem executada com o montador Newbler apresentou o menor número de *contigs,* o maior *contig* e o maior valor de N50. Portanto, foi selecionada para realizar a montagem do genoma, apesar de ter apresentado a

menor fração do genoma 99,21% (tamanho final do genoma: 2337467 bases) contra os 99,64% da tentativa com o montador MIRA.

Tabela 6 - Montagens geradas para as linhagens de *C. pseudotuberculosis* isoladas no México.

Linhagem	Montador	Número de *contigs*	Tamanho do genoma (pb)	Tamanho do menor *contig*	Tamanho do maior *contig*	N50
Cp MEX9	Newbler*	7	2.318.987	5.919	721.828	372.309
	MIRA	16	2.328.960	499	688.177	488.630
Cp MEX1	Newbler	8	2.318.859	5.918	607.469	405.039
	MIRA	8	2.323.088	636	1.320.751	1.320.751
	SPAdes*	6	2.320.650	5.921	722.024	543.202
Cp MEX25	Newbler*	7	2.321.849	5.964	722.269	543.326
	MIRA	10	2.324.784	702	754.740	543.343
Cp MEX29	Newbler*	9	2.319.027	5.909	569.322	367.275
	MIRA	20	2.330.734	580	445.221	192.411
Cp MEX30	MIRA	62	2.393.483	776	244.218	91.364
	Newbler*	33	2.353.023	3.163	280.084	103.276
	SPAdes	529	2.389.462	351	21.673	6.707
Cp MEX31	MIRA	26	2.363.634	568	621.280	186.262
	Newbler	11	2.352.237	5.970	746.561	347.951
	SPAdes*	13	2.350.992	159	746.203	535.978

*Montagem selecionada.

Foram executadas quatro tentativas de montagem para a linhagem MEX1. Analisando o gráfico do QUAST (ANEXO I - Figura adicional 3) a montagem realizada pelo montador MIRA obteve o melhor resultado. Porém, a mesma apresentou um erro de montagem justamente no maior *contig* que apresentou um tamanho de 1.320.751 pbs (ANEXO I - Figura adicional 4) o que foi confirmado pelo gráfico de sintenia do CONTIGuator (ANEXO II - Figura adicional 5), o que inviabilizou o uso desta. Sendo assim, a montagem selecionada foi realizada pelo montador SPAdes (gráfico de sintenia na Figura 10B, na página 39) que apresentou os melhores resultados dentre os outros critérios avaliados.

Para realizar a montagem da linhagem MEX25 foram executadas duas tentativas de montagem. Ambas as montagens executadas apresentaram bons

valores, com exceção do tamanho do menor *contig* da montagem com MIRA que foi consideravelmente menor. Dessa forma, a montagem selecionada foi a que se mostrou mais consistente em todos os critérios no caso a montagem com Newbler, a qual apresentou 99,33% do tamanho do genoma final.

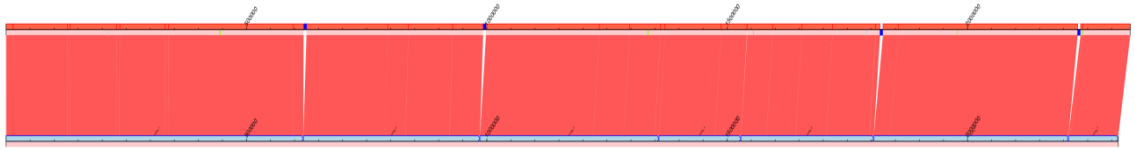
No caso da *C. pseudotuberculosis* MEX29 foram executadas duas tentativas. Analisando os resultados escolheu-se a montagem do *software* Newbler, a qual teve melhores resultados em quase todos os critérios avaliados, com exceção do tamanho do genoma que foi de 2.333.734 no MIRA e 2.319.027 no Newbler.

Foram realizadas três tentativas de montagem para a *C. pseudotuberculosis* MEX30. De acordo com os dados gerados optou-se pela montagem do Newbler que apresentou o menor número de *contigs*, o maior *contig* entre os menores *contigs*, o maior *contig* e o maior valor de N50. Além disso, a ferramenta QUAST confirmou esta escolha como pode ser observado no ANEXO I – Figura adicional 1.

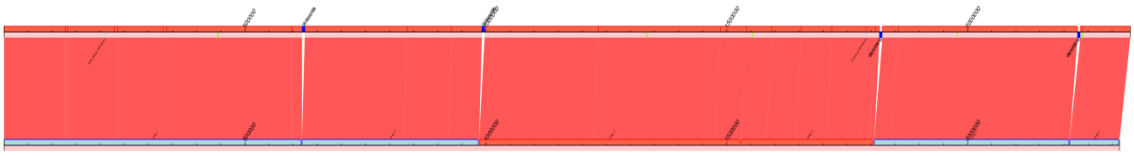
Para a linhagem MEX31, três tentativas de montagem foram executadas. De acordo com os resultados produzidos, a tentativa selecionada para realizar a montagem deste genoma foi o resultado gerado pelo montador SPAdes. Este apresentou uma diferença significativamente maior no valor de N50 e o melhor resultado no gráfico gerado pelo QUAST (ANEXO I - Figura adicional 2), mesmo possuindo o menor *contig* entre os menores *contigs* gerados.

Como já descrito na metodologia, a ordenação dos *contigs* foi realizada no *software* CONTIGuator utilizando um genoma de referência. Para as *C. pseudotuberculosis* biovar Ovis linhagens MEX9, MEX25 e MEX29 o genoma de referência utilizado foi o da *C. pseudotuberculosis* linhagem 29156, como pode ser visto na Figura 10 (A, C e D) respectivamente. Para a *C. pseudotuberculosis* MEX1, também do biovar Ovis, foi utilizado como genoma de referência a *C. pseudotuberculosis* MEX9 (Figura 10B). Pode-se verificar que a similaridade dos alinhamentos confirma a escolha das *C. pseudotuberculosis* 29156 e *C. pseudotuberculosis* MEX9 como referência para a ordenação dos *contigs*.

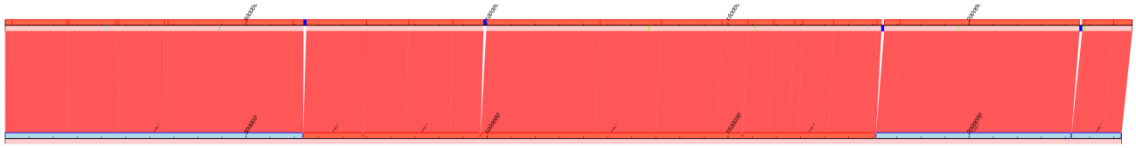
A



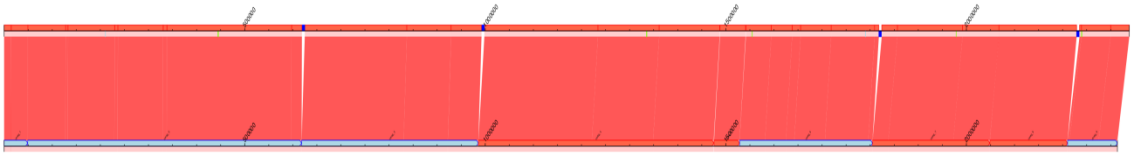
B



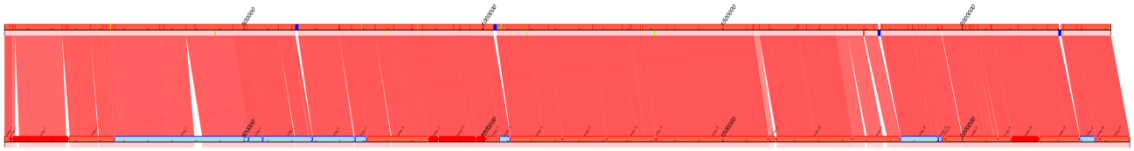
C



D



E



F

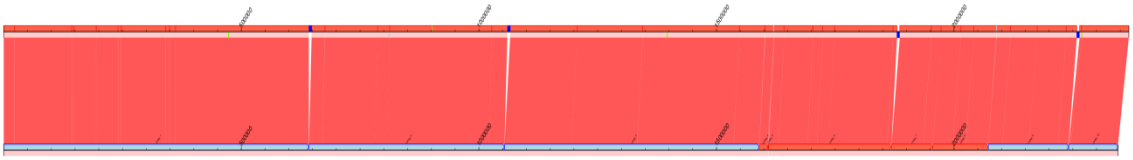


Figura 10 – Gráfico de sintenia do alinhamento dos novos genomas em uma referência utilizando o *software* CONTIGuator, após rodar o *script* MoveDNAA.

(A) Ordenação dos *contigs* do organismo *C. pseudotuberculosis* MEX9 usando como referência o organismo *C. pseudotuberculosis* 29156 (fita superior). (B) Ordenação dos *contigs* do organismo *C. pseudotuberculosis* MEX1 usando como referência o organismo *C. pseudotuberculosis* MEX9 (fita superior). (C) Ordenação dos *contigs* do organismo *C. pseudotuberculosis* MEX25 usando como referência o organismo *C. pseudotuberculosis* 29156 (fita superior). (D) Ordenação dos *contigs* do organismo *C. pseudotuberculosis* MEX29 usando como referência o organismo *C. pseudotuberculosis* 29156 (fita superior). (E) Ordenação dos *contigs* do organismo *C. pseudotuberculosis* MEX30 usando como referência o organismo *C. pseudotuberculosis* 316 (fita superior). (F) Ordenação dos *contigs* do organismo *C. pseudotuberculosis* MEX31 usando como referência o organismo *C. pseudotuberculosis* E19 (fita superior).

Figuras geradas pelo CONTIGuator.

Para realizar o alinhamento e ordenação dos *contigs* gerados para os dados da *C. pseudotuberculosis* MEX30, no *software* CONTIGuator (Figura 10E), utilizou-se como referência a *C. pseudotuberculosis* 316. Já para a *C. pseudotuberculosis* MEX31, os *contigs* foram alinhados e ordenados no mesmo *software* tendo como referência o genoma da *C. pseudotuberculosis* E19 (Figura 10F). Como pode ser visto, o alinhamento apresentado demonstra a similaridade dentre estes organismos.

Em todas as linhagens do biovar Ovis, ao realizar o alinhamento e ordenação dos *contigs*, somente o *contig* que contém o *cluster* ribossomal, composto pelos genes codificadores dos RNAs ribossomais (rRNAs) 23S, 16S e 5S, não alinhou no genoma de referência. Isto ocorreu devido a esta ser uma região muito repetitiva do genoma. Os pontos azuis que podem ser observados na parte superior dos gráficos da Figura 10 indicam o local de cada um dos *clusters* no genoma de referência, o que indica a provável posição destes no novo genoma. A ordenação dos *contigs* da *C. pseudotuberculosis* MEX30 apresentou o mesmo resultado das linhagens do biovar Ovis, somente o *cluster* ribossomal não foi alinhado. Porém, no caso da *C. pseudotuberculosis* MEX31, quatro dos *contigs* gerados não foram alinhados no genoma de referência. Primeiramente, os *clusters* dos rRNAs foram adicionados aos genomas alvos, fechando quatro dos *gaps* encontrados em cada linhagem, após esta etapa foi realizado o fechamento dos *gaps* remanescentes de cada uma das linhagens de acordo com a estratégia observada na Tabela 7.

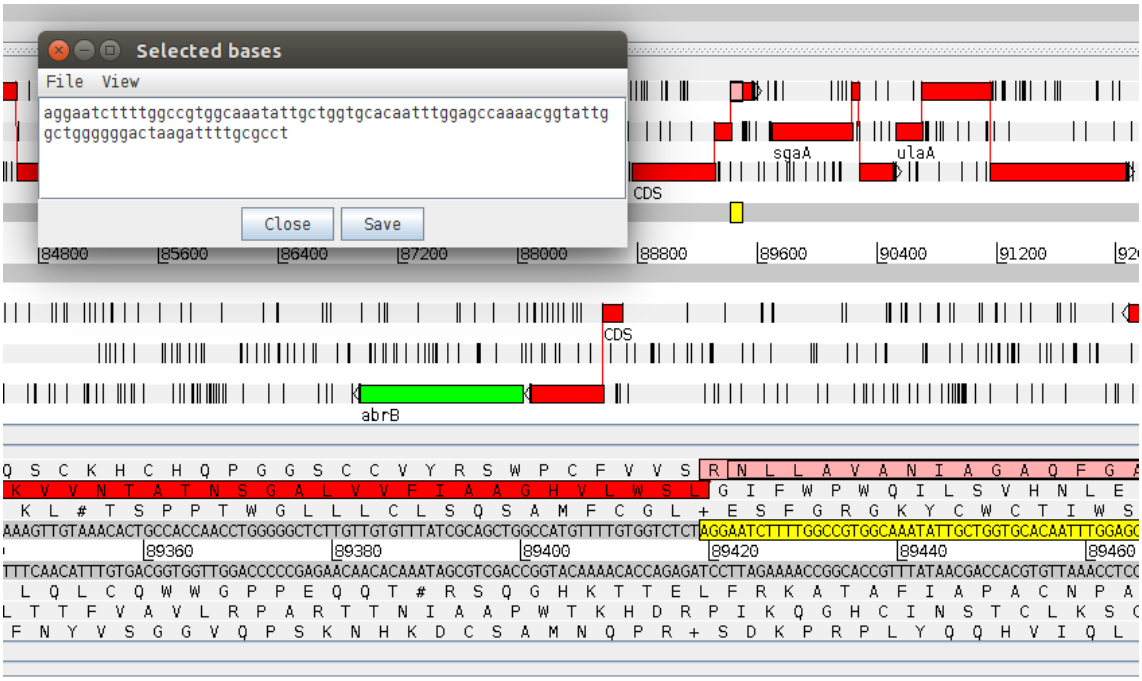
Tabela 7 – Estratégia de fechamento dos *gaps* finais para cada um dos novos genomas.

Linhagem	Estratégia
Cp MEX1	1 *gap* por sobreposição
Cp MEX9	2 *gaps* por alinhamento no CLC
Cp MEX25	2 *gaps* por sobreposição
Cp MEX29	2 *gaps* por sobreposição e 2 *gaps* por alinhamento no CLC
Cp MEX30	15 *gaps* por sobreposição e 14 *gaps* por alinhamento no CLC
Cp MEX31	3 *gaps* por sobreposição e 3 *gaps* por alinhamento no CLC (incluindo os *contigs* que não haviam alinhado)

5.3. Anotação e curadoria

Após a obtenção dos genomas completos passou-se para a fase de anotação e curadoria de *frameshifts*. Foi observado que a maioria dos *frameshifts* ocorreu em regiões de homopolímeros. A Figura 11 apresenta um exemplo que ocorreu no genoma da *C. pseudotuberculosis* MEX30, o que, como discutido anteriormente, é uma característica da plataforma utilizada.

A



B

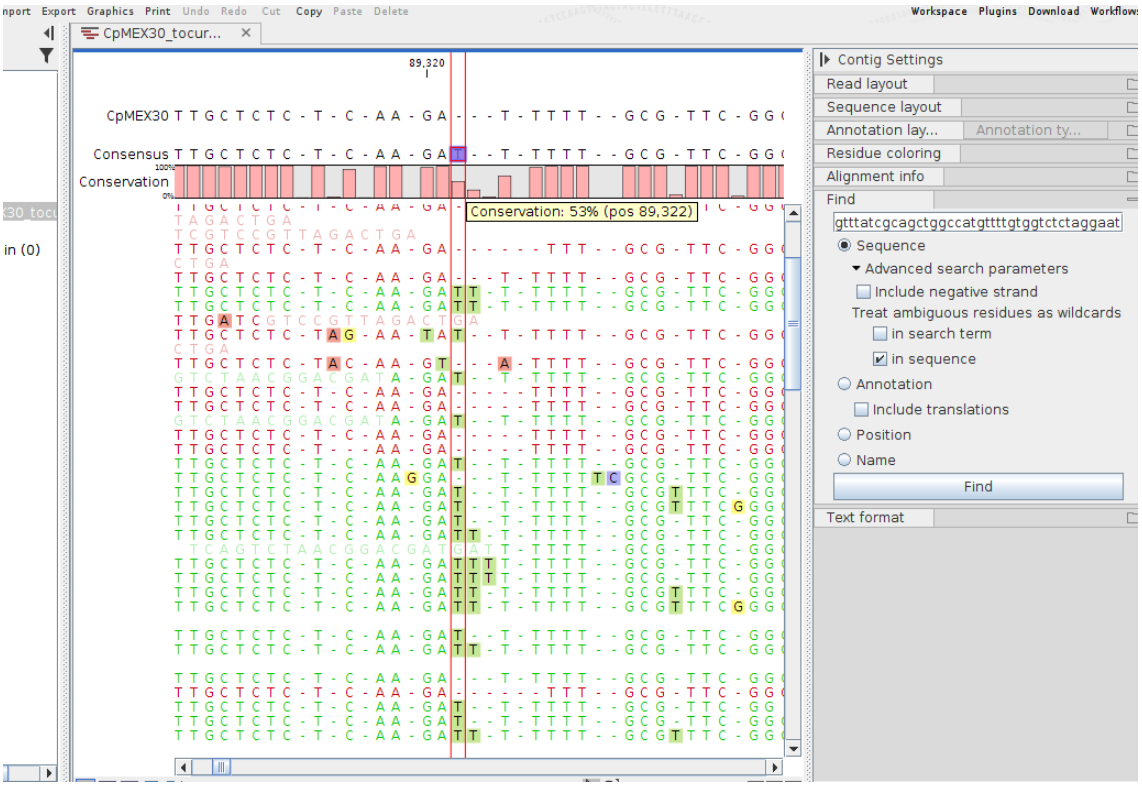


Figura 11 – Homopolímero encontrado durante a curadoria manual da *C. pseudotuberculosis* MEX30.

(A) *Frameshift* encontrado durante a curadoria manual no *software* Artemis. (B) Ao fazer a busca no Artemis identificou-se que havia um aminoácido de Timina (T) faltando nessa região, o qual estava presente na fita *Consensus* com 53% de conservação.

A Tabela 8 apresenta os resultados encontrados nestas etapas para os genomas abordados neste trabalho. Todas as linhagens apresentaram características compatíveis com os demais genomas desta espécie tais como: tamanho do genoma em torno de 2,3 Mb; conteúdo GC de 52%; número de CDSs em torno de 2.000; número de pseudogenes variando de 33 à 92; número de rRNAs igual à 12 para todas as linhagens e número de tRNAs variando de 48 à 51.

Tabela 8 – Resumo dos dados dos genomas após a anotação e curadoria.

Linhagens	Tamanho do genoma (pb)	Conteúdo GC (%)	CDSs	Pseudogenes	rRNAs	tRNAs
Cp MEX1	2.337.090	52,18	2.081	31	12	49
Cp MEX9	2.337.578	52,18	1.996	44	12	49
Cp MEX25	2.337.529	52,18	1.992	45	12	49
Cp MEX29	2.337.866	52,18	2.078	34	12	49
Cp MEX30	2.368.140	52,11	2.053	92	12	51
Cp MEX31	2.367.880	52,09	2.081	61	12	48

5.4. Genômica comparativa

5.4.1. Análises de pangenoma das linhagens de *C. pseudotuberculosis* provenientes do México

O estudo do pangenoma dos organismos permite uma visão geral do repertório genômico de um organismo (SOARES *et al*., 2013) e tem sido considerada uma importante ferramenta nos estudos de genômica comparativa de bactérias (ROULI *et al*, 2015). Para a caracterização do repertório gênico dos organismos estudados, utilizou-se o *Protein Family Sorter* do PATRIC, o qual atribui rapidamente famílias proteicas aos genes dos genomas sendo estudados, utilizando a anotação previamente realizada pelo PATRIC para gerar os *clusters* inicias e então utilizando uma estratégia baseada em k-mers para diferenciar os *clusters*.

Desta forma obteve-se o pangenoma dos seis genomas de *C. pseudotuberculosis* deste estudo, o qual pode ser visto na Figura 12 e apresentou como resultado 2295 genes encontrados nos seis genomas.

Destes, 1903 (Figura 13) formam o genoma central, 49 são genes únicos (Tabela 9) e 343 formam o genoma acessório.

Além do pangenoma dos seis organismos realizou-se um pangenoma para o biovar Equi e um para o biovar Ovis, buscando caracterizar e diferenciar os dois biovares. Assim, a Figura 12 também apresenta uma análise da distribuição gênica por biovar. Na distribuição observou-se que 1974 genes são compartilhados entre os dois biovares, sendo que 71 não estão presentes em todos os genomas. O biovar Ovis possui 164 genes exclusivos e o biovar Equi possui 157.

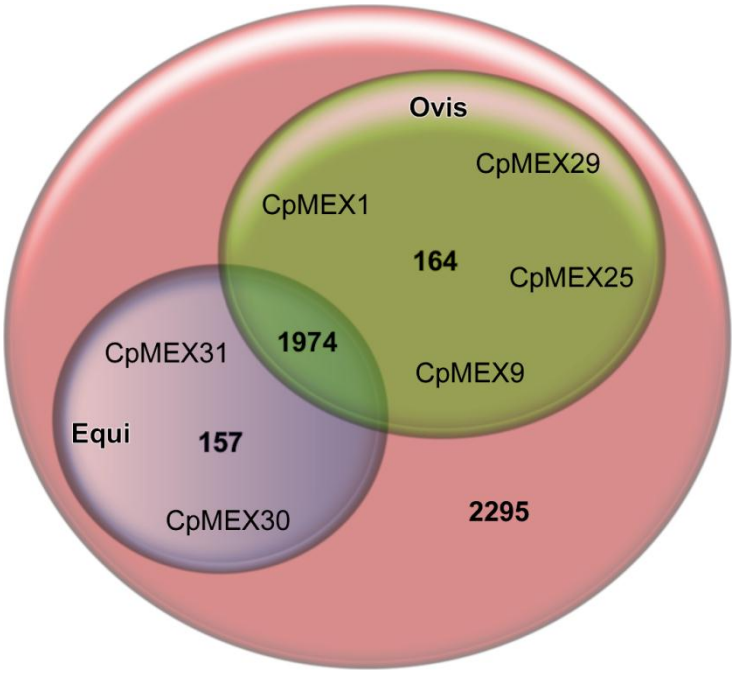


Figura 12 – Pangenoma dos seis genomas de *C. pseudotuberculosis* utilizados neste trabalho separados por biovar.

Além disso, foi encontrado o genoma central do biovar Equi e do biovar Ovis, o do biovar Equi é composto de 2069 genes e o biovar Ovis é composto de 2016 genes. Analisando o genoma central de cada biovar foram identificados genes que aparecem somente a um destes. No genoma central de Ovis há 113 genes que não fazem parte do genoma central compartilhado entre os dois biovares, e no genoma central de Equi há 166 genes. Destes 132 genes formam o genoma central exclusivo de Equi e 101 genes formam o genoma central exclusivo de Ovis. Na Figura 13, é mostrado o genoma central

tanto do biovar Ovis quanto do biovar Equi, destacando-se os valores de genes que estão compartilhados entre o genoma central de todas as linhagens de *C. pseudotuberculosis* estudadas neste trabalho, assim como os genes que pertencem apenas ao genoma central de um biovar ou de outro.

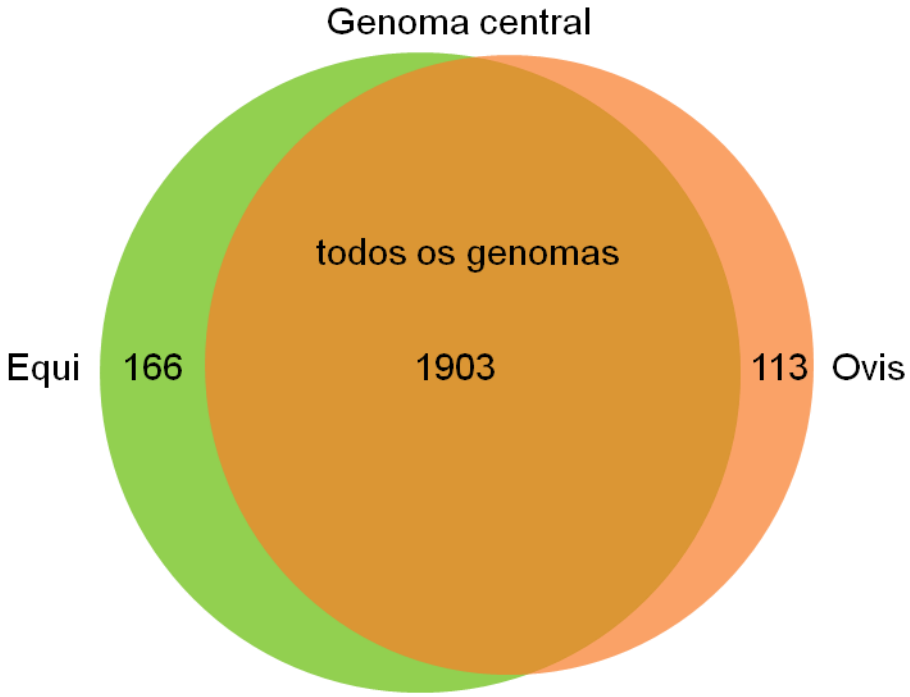


Figura 13 – Diagrama de Venn representando o genoma central das *C. pseudotuberculosis* provenientes do México.

Ao centro o número de genes que constituem o genoma central de todas as *C pseudotuberculosis* estudadas; na parte em verde o número de genes do genoma central do biovar Equi; e na parte em laranja o número de genes do genoma central do biovar Ovis.

O pangenoma exclusivo de cada biovar apresenta os genes que podem diferenciar um biovar do outro. Através da análise destes chegou-se a importantes *clusters* gênicos caracterizados na literatura em outros organismos e que podem estar contribuindo nas características de cada biovar. Os quais serão apresentados abaixo.

Além dos valores de genoma central e de genes exclusivos de cada biovar é possível destacar a quantidade de genes únicos de cada genoma e a quantidade de genes exclusivamente ausentes em cada linhagem estudada, a qual é apresentada na Tabela 9.

Tabela 9 - Quantidade de genes únicos em cada genoma.

Linhagens	Genes únicos
Cp MEX1	7
Cp MEX9	7
Cp MEX25	6
Cp MEX29	4
Cp MEX30	14
Cp MEX31	11

5.4.1.1. Sistemas de restrição de modificação em *C. pseudotuberculosis* biovar Ovis linhagens MEX1, MEX9, MEX25 e MEX29

A Figura 14 mostra o genoma circular dos organismos estudados, a qual utilizou *C. pseudotuberculosis* MEX1 como referência. Este genoma foi utilizado como genoma de referência devido a este ter sido sequenciado utilizando a enzima HiQ *Polymerase*, a qual como demonstrado por Pereira e colaboradores (2016), possui maior qualidade em relação à utilizada no sequenciamento das outras linhagens deste trabalho (*Sequencing Polymerase*). Na figura é possível observar um *cluster* gênico (destacado em preto) presente somente nas linhagens do biovar Ovis. Dentre os genes presentes podemos destacar que a maioria trata-se de genes hipotéticos. Assim como genes de fagos e genes relacionados ao sistema de restrição de modificação (RM) de DNA. Segundo Kobayashi e colaboradores (1999) estes genes relacionados aos sistemas RM poderiam ser elementos genéticos móveis que causam mudanças no genoma, e, em algumas bactérias, a interação desses elementos com o genoma pode ser uma força impulsionadora para a evolução do genoma. Um *Cluster* de genes transferido horizontalmente pode formar regiões conhecidas como ilhas genômicas (GEI), as quais contribuem na plasticidade genômica e evolução bacteriana, e na sobrevivência e virulência de patógenos (HOCHHUT *et al.*, 2004; SOARES *et al*., 2015), como possivelmente os fagos identificados no genoma nesta região.

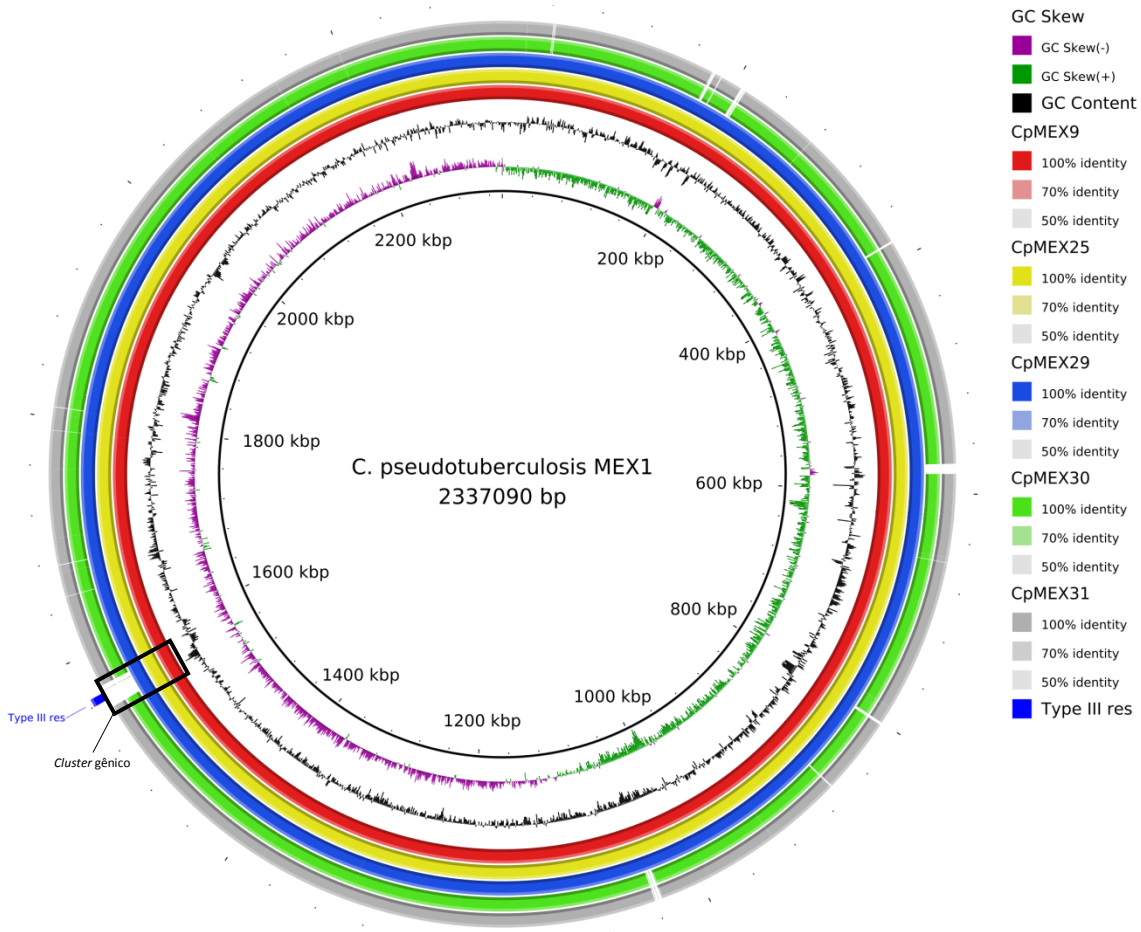


Figura 14 – Visualização circular dos genomas dos organismos estudados tendo como genoma de referência a *C. pseudotuberculosis* MEX1 que pertence ao biovar Ovis.

O retângulo em preto destaca a região do bloco gênico de genes associados à metilação que estão presentes nas linhagens do biovar Ovis e ausentes nas linhagens do biovar Equi.

Figura gerada pelo BRIG.

Existente em muitos procariotos, a metilação de DNA é uma alteração química do mesmo, a qual envolve a transferência de um grupo metil do S-adenosilmetionina para o DNA (WION & CASADESÚS, 2006). A metilação pode estar ativa de duas formas, a enzima de DNA metiltransferase (MTase) associada a uma endonuclease de restrição (REase) (formando os chamados sistemas RM) ou a MTase atuando independente da REase. A MTase atuando independente da REase regula a expressão gênica e também pode estar envolvida na estabilidade do cromossomo, reparo de erros, replicação e virulência (CASADESÚS & LOW, 2006; ZHU *et al*., 2015; BLOW *et al*., 2016).

Nos sistemas RM a MTase metila o genoma da própria bactéria (transferindo um grupo metila) em prováveis sítios de ação da REase evitando

a clivagem do mesmo, e a REase degrada o DNA invasor (BLOW *et al*., 2016). Estes sistemas podem ser classificados em quatro tipos (I, II, III e IV), os quais, juntamente com as subunidades correspondentes, podem ser visualizados na Tabela 10. O tipo I necessita de duas subunidades R, duas subunidades M e uma subunidade S para ser ativo, onde o sítio de clivagem do DNA exógeno é variável, a metilação de duas subunidades R e uma subunidade S para ser ativo. O tipo II possui proteínas independentes, uma para restrição e uma para metilação e, o sítio de clivagem é fixo. O tipo III também possui subunidades e sítio de clivagem variável, precisa de duas subunidades R e duas subunidades M para ser ativo, e duas subunidades M para a metilação ser ativa. Já o tipo quatro possui somente a enzima de restrição e é específico para clivagem de material genético de bacteriófagos com DNA altamente modificado (SUZUKI, 2012).

Tabela 10 - Propriedades gerais dos quatro tipos de sistemas RM e de metilação.

Sistema RM	Restrição (RE)		Metilação (MT)	
	Maquinaria	Sítio de clivagem	Maquinaria	Nucleobase
Tipo I	R_2M_2S	Variável	M_2S	^{6m}A
Tipo II	RE	Fixo	MT	^{6m}A,^{5m}C, ^{4m}C
Tipo III	R_2M_2	Variável	M_2	^{6m}A
Tipo IV	RE	Variável		

O tipo I compreende as subunidades R, M e S. A metilação é realizada pelas subunidades M e S. O tipo III compreende as subunidades R e M. A subunidade M cataliza a metilação sozinha. O tipo II compreende duas proteínas independentes, RE e MT. O tipo IV compreende apenas o sistema RE e restringe o DNA com modificações heterólogas. A metilação normalmente produz ^{6m}A,^{5m}C ou ^{4m}C.

Fonte: Adaptado de SUZUKI (2012).

Shell e colaboradores (2013) demonstraram que em *Mycobacterium tuberculosis* linhagem H37Rv a metilação exerce importante papel para a sobrevivência desta bactéria em um ambiente com pouco oxigênio.

Ambos os tipos I e III de DNA metiltransferase foram encontrados nas quatro linhagens de *C. pseudotuberculosis* do biovar Ovis estudadas, nas quais o tipo I possivelmente não está ativo, pois a subunidade S não foi identificada, já o tipo III está completo e, portanto possivelmente estaria ativo.

As DNA metiltransferases associadas aos sistemas de restrição de modificação do DNA tipo III estão sujeitas a alterações aleatórias reversíveis em regiões repetitivas de DNA, o que pode ativar ou desativar sua expressão. No caso da bactéria patogênica *Neisseria meningitidis* tem se mostrado que isso afeta o seu fenótipo através da expressão de diferentes proteínas de membrana, resposta ao estresse e outros componentes metabólicos. Isto pode afetar a sua capacidade de invasão, de transporte e de adaptação a diferentes ambientes (SRIKHANTA *et al*., 2009; SEIB *et al*., 2011; TAN *et al*., 2016). Além disso, pode alterar sua resistência antimicrobiana, ao estresse oxidativo e a antibióticos, assim como influenciar na formação de biofilme e na sobrevivência durante a colonização e invasão do hospedeiro (TAN *et al*., 2016).

Em outro estudo, Kwiatek e colaboradores (2015) inativaram o gene codificante da metiltransferase tipo III de *Neisseria gonorrhoeae*, os resultados sugeriram que a mesma pode estar envolvida na patogenicidade (invasão e adesão ao hospedeiro e regulação do biofilme) deste organismo.

5.4.1.2. CRISPR-Cas em *C. pseudotuberculosis* biovar Equi linhagens MEX30 e MEX31

A Figura 15 mostra o genoma circular dos organismos estudados, com a *C. pseudotuberculosis* MEX30 como referência. Na figura é possível observar o *cluster* gênico (destacado em preto) presente somente nas linhagens do biovar Equi, no qual está contido os genes responsáveis pela redução de nitrato. A identificação destes genes confirma a diferenciação entre os biovares (BIBERSTEIN *et al*., 1971). Na busca por outro gene capaz de diferenciar os dois biovares chegou-se ao *cluster* que contém o sistema CRISPR-Cas, destaque em azul na mesma figura.

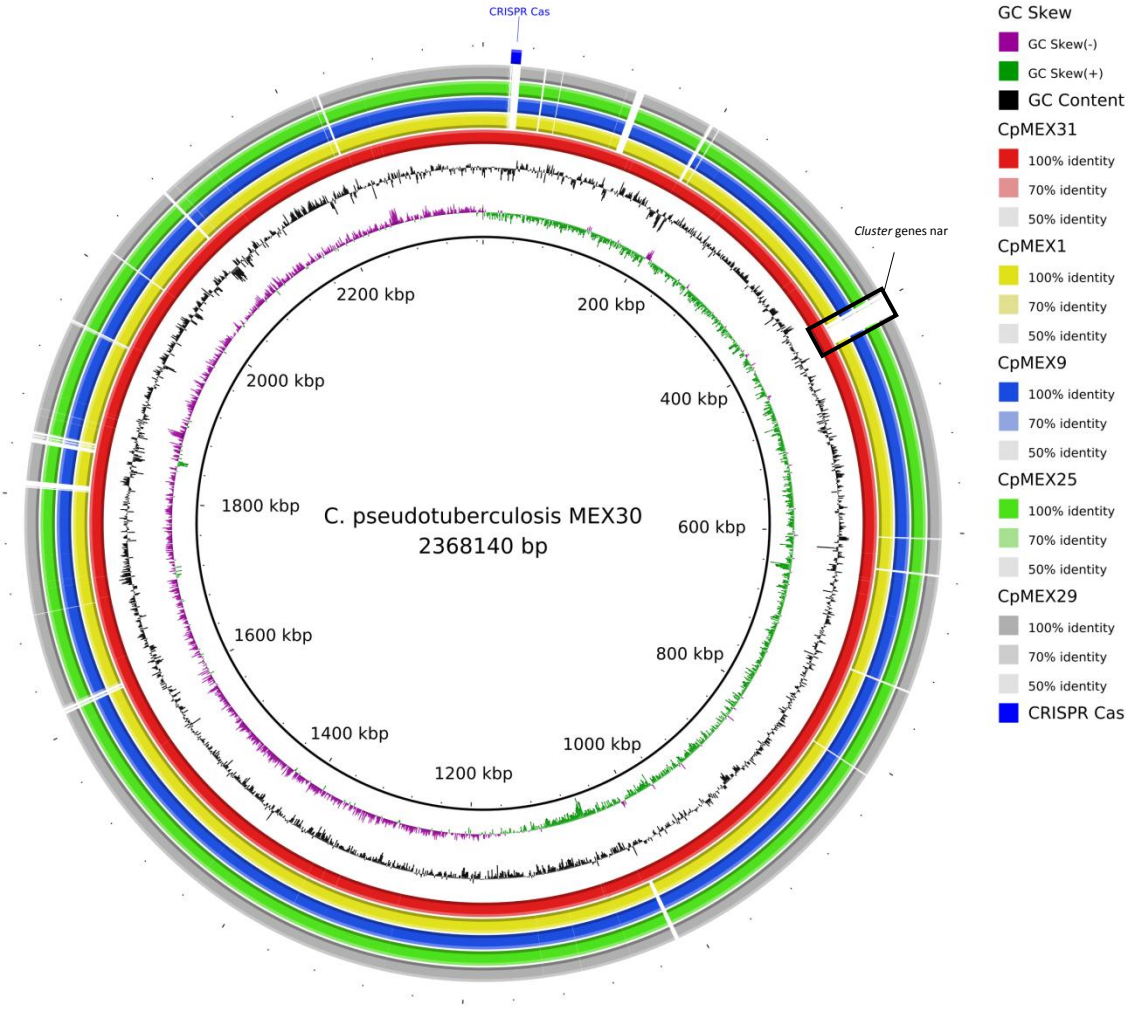


Figura 15 - Visualização circular dos genomas dos organismos estudados tendo como genoma de referência a *C. pseudotuberculosis* MEX30 que pertence ao biovar Equi.

O retângulo em preto destaca a região do bloco gênico nar. Em azul, genes associados ao sistema CRISPR-Cas que estão presentes nas linhagens do biovar Equi e ausentes nas linhagens do biovar Ovis.

Figura gerada pelo BRIG.

O *cluster* do sistema CRISPR-Cas encontrado possui um *locus* com tamanho 10,9 Kb com provável organização em operon. Este *locus* é composto por oito genes Cas organizados na seguinte ordem: Cas2, Cas1, Cas3, Cse3, Cse1, Cse2, Cse4 e Cas5e (Figura 16). O *locus* possivelmente pertence ao tipo I-E, similar ao descrito por SANGAL e colaboradores (2013) em *Corynebacterium diphtheriae*.

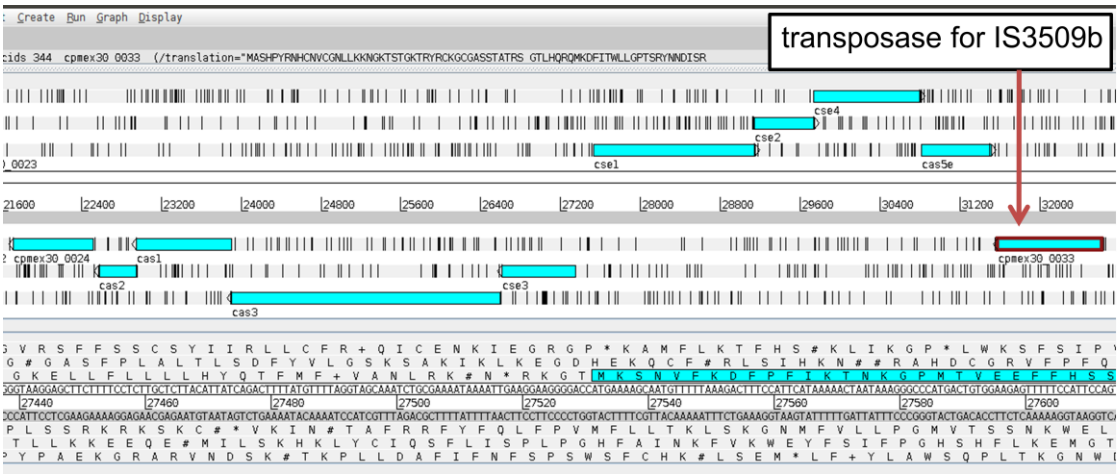


Figura 16 – *Locus* do provável operon de genes cas.

O gene Cas1 atua como DNA endonuclease com provável participação na integração do *spacer* DNA no CRISPR cassete. Assim, ele participa no reconhecimento e na integração do material genético externo no módulo CRISPR. O gene Cas2 corresponde a uma das quatro famílias proteicas encontradas no genoma de procariotos que está associada aos elementos CRISPR e sua atividade pode estar associada a uma endoribonuclease ou endodeoxiribonuclease. A helicase Cas3 corresponde a uma proteína associada aos elementos repetidos de CRISPR que ocorrem em associação com o sistema Cas1 e Cas2 (MAKAROVA *et al*., 2015).

O gene Cse 1 é também conhecido como Cas8, a qual contém características de polimerase-*like*, e também pode contribuir para a proteção do próprio DNA. O gene Cse2 é também conhecido como subunidade pequena e é uma *α-helical* e é de grande importância para a montagem do complexo (TSUI & LI, 2015). O gene Cse3 é também conhecido como Cas6e, é uma proteína catalítica, possivelmente uma derivação divergente do gene Cas6, que cliva o pré-crRNA formando o crRNA (CRISPR RNA) (MAKAROVA *et al*., 2011). O gene Cse4 é também conhecido como cas7, o qual participa da degradação de RNA co-transcricional (MAKAROVA *et al*., 2015). O gene Cse5e é também conhecido como Cas5 (MAKAROVA *et al*., 2011) e pode estar envolvido na clivagem do pré-RNA (NAM *et al*., 2012).

Na constituição do *locus* de CRISPR foi identificado a transposase IS3509b, a qual encontra-se localizada *downstream* em relação aos demais constituintes desta região do genoma. Ainda em relação à localização do *locus*

de CRISPR-Cas foram identificados quatro genes. E em relação às análises de predição de aminoácidos os resultados indicam que essas regiões codificantes correspondem a proteínas hipotéticas descritas na literatura em diferentes organizações de sistemas de CRISPR-Cas em diversos gêneros bacterianos.

Na análise comparativa dos genomas de *C. pseudotuberculosis* das linhagens do biovar Ovis foi observado que este *locus* não encontra-se conservado, no qual foi confirmada somente a presença do CRISPR-Cas5e. Desta forma esta região é diferencial dentre as linhagens do organismo *C. pseudotuberculosis* provenientes do México. Porém, a transposase IS3509b foi identificada em todas as linhagens selecionadas para este estudo. Os resultados encontrados sugerem um perfil genético diferencial entre as linhagens, o qual foi obtido com as análises das proteínas hipotéticas que compõe este *locus*.

De acordo com MAKAROVA e colaborados (2015) em ensaios de classificação evolutiva do sistema CRISPR-Cas este módulo estará completo mediante a identificação dos elementos Cas1 e Cas2 no *locus*, sendo assim é possível sugerir a presença de um *locus* completo nas *C. pseudotuberculosis* MEX30 e MEX31.

Em estudos prévios, MAKAROVA e colaboradores (2015) descreveram a presença de proteínas hipotéticas, isto é sem função definida na constituição do *locus* CRISPR-Cas em diversos gêneros bacterianos. Este dado foi confirmado no presente estudo, no qual proteínas hipotéticas participam da organização do *locus* CRISPR-Cas nas linhagens de interesse. No estudo comparativo foi identificado que não há conservação dessa região entre as diferentes linhagens de *C. pseudotuberculosis* isoladas no México.

Estudos prévios relativos aos sistemas de defesa de bactérias e archeas descrevem a presença de pequenas sequências repetidas do genoma (elementos móveis – sequências de inserção) associadas ao *locus* de CRISPR-Cas (POURCEL *et al*., 2005). No presente trabalho as análises de predição de sequências das linhagens MEX30 e MEX31 foram identificadas transposases IS3509b na composição do *locus* em estudo. No estudo comparativo com as diferentes linhagens provenientes do México foi observada a presença deste elemento móvel em todas as linhagens bacterianas. A presença deste elemento móvel pode estar associada ao processo de transferência horizontal

de genes em que a aquisição e a perda de elementos genéticos é constante e resulta na diferença genética entre os microorganismos (SANGAL *et al*., 2013). Assim, as diferenças encontradas nos *locus* CRISPR-Cas das linhagens provenientes do México podem estar associadas a estes elementos móveis.

5.4.1.3. Genes únicos

Os genes exclusivos de cada linhagem com função conhecida são apresentados na Tabela 11. Dos 49 genes únicos encontrados na análise de pangenoma nas seis linhagens poucos têm função conhecida, e as linhagens de *C. pseudotuberculosis* MEX29 e MEX31 somente apresentaram genes hipotéticos. Porém, em um estudo de transcriptômica (PINTO *et al*., 2014) e um de proteômica (DA SILVA, 2015), nos quais *C. pseudotuberculosis* foram replicadas em diferentes condições, foram observados que vários genes/proteínas hipotéticos(as) apresentaram alto valor de expressão gênica, o que demonstra a necessidade da caracterização destas proteínas.

Tabela 11 - Genes únicos com função conhecida encontrados no PATRIC dos genomas de *C. pseudotuberculosis* deste trabalho.

Linhagem	Gene	Produto
Cp MEX1	fadF	Fe-S oxidoreductase
Cp MEX9	-	ATP/GTP-binding protein
Cp MEX25	-	L-asparaginase (EC 3.5.1.1)
Cp MEX30	-	Hypothetical protein YggS, proline synthase co-transcribed bacterial homolog PROSC
Cp MEX30	arcD	Arginine/ornithine antiporter ArcD

5.4.2. Filogenômica

As análises filogenômicas apresentaram um agrupamento dos organismos de acordo com o biovar (Figuras 17 e 18), o que havia sido demonstrado em outros trabalhos com *C. pseudotuberculosis* (SOARES *et al*., 2013; Oliveira *et al*., 2016). Além disso, pode-se destacar a similaridade dos genomas de acordo com o hospedeiro. *C. pseudotuberculosis* MEX1 e *C.*

pseudotuberculosis MEX9 foram isolados de cabra, *C. pseudotuberculosis* MEX25 e *C. pseudotuberculosis* MEX29 foram isoladas de ovelha e *C. pseudotuberculosis* MEX30 e *C. pseudotuberculosis* MEX31 foram isoladas de cavalo.

Linhagem	1	2	3	4	5	6
1: CpMEX1	100	99.9	99.8	99.7	93.6	93.7
2: CpMEX9	99.9	100	99.8	99.8	93.6	93.7
3: CpMEX25	99.7	99.8	100	100	93.6	93.8
4: CpMEX29	99.7	99.8	99.9	100	93.6	93.7
5: CpMEX30	92.4	92.4	92.4	92.4	100	99.8
6: CpMEX31	92.5	92.6	92.6	92.6	99.8	100

Figura 17 – Alinhamento gerado pelo *software* Gegenees.

Retângulo em azul destaca o agrupamento do biovar Ovis. Retângulo em vermelho destaca o agrupamento do biovar Equi.

Figura gerada pelo Gegenees.

No ANEXO III podem ser vistas as famílias proteicas de genes com função conhecida encontradas somente no biovar Ovis (Tabela adicional 1) e no biovar Equi (Tabela adicional 2). Estas famílias protéicas exclusivas de cada biovar podem estar contribuindo para as diferenças encontradas na árvore filogenômica da Figura 18.

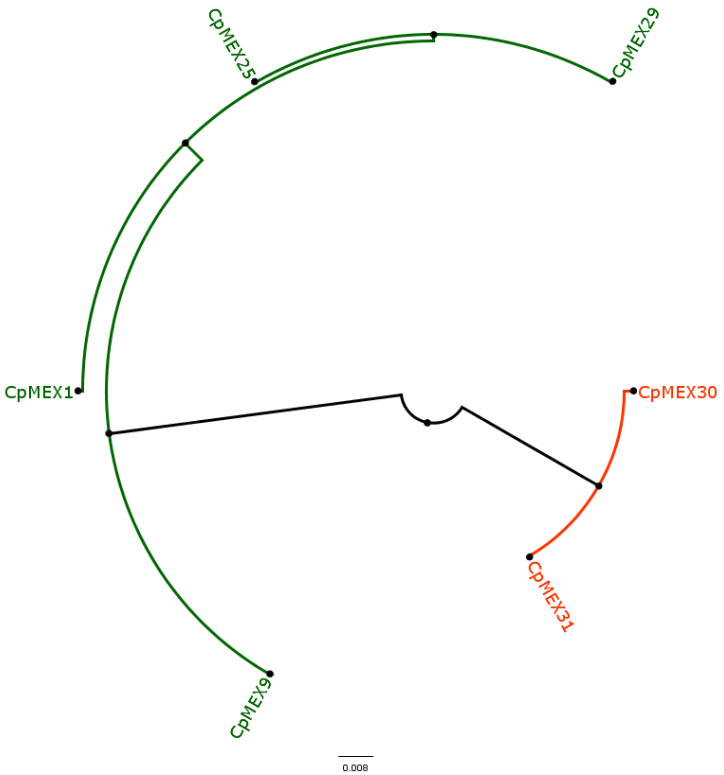


Figura 18 – Árvore filogenômica para os biovares ovis e equi da espécie *C. pseudotuberculosis*.

A partir do método quantitativo *Neighbor Joining* foi determinada a árvore usando os genomas das linhagens descritas no estudo, onde foi possível observar os grupos separados pelos biovares. Em verde estão destacadas as linhagens pertencentes ao biovar Ovis e em vermelho estão destacadas as linhagens pertencentes ao biovar Equi.

Figura gerada pelo FigTree.

Como já descrito por SONGER e colaboradores (1988), os biovares podem ser caracterizados e diferenciados por ensaios bioquímicos a partir do metabolismo de nitrato. Contudo, nossos resultados foram capazes, também, de diferenciar as linhagens aqui estudadas a partir do seu conteúdo genético, por métodos filogenéticos. A distribuição dos biovares feita por filogenômica destacou dois clados que se mostram divergentes na árvore, cada um deles pertencente a um biovar. É possível que essa divergência evolutiva entre os biovares possa ser interpretada como um evento de especiação biológica, mais especificamente a anagênese, discutida recentemente por Oliveira e colaboradores (2016). Este trabalho apontou que as linhagens do biovar Ovis são mais bem estabilizadas geneticamente que as linhagens do biovar Equi, devido à análise nos ramos internos, onde foi possível observar que tais ramos apresentam variação de tamanho dependendo da linhagem. Diferentemente,

as linhagens do biovar Ovis apresentam um padrão no tamanho dos ramos, onde foi possível interpretar uma estabilidade entre esses genomas. O mesmo trabalho apontou que possivelmente esse tipo de variação nos ramos seria devido ao grau de mutações de transição e transversão, no qual avalia a diferença de nucleotídeos purina por purina ou pirimidina por pirimidina (transição) e purina por pirimidina (transversão).

5.4.3. Ilhas de patogenicidade

Para identificar as possíveis Ilhas de Patogenicidade (PAIs) das linhagens de interesse deste trabalho utilizou-se o *software* GIPSy (SOARES *et. al.*, 2015), e os resultados gerados por este foram visualizados através do *software* BRIG. Na tabela 12 pode ser observada a quantidade de PAIs e GEIs preditas para cada uma das linhagens de *C. pseudotuberculosis* estudadas neste trabalho. As GEIs são ilhas genômicas encontradas durante o processo de identificação de PAIs que não apresentaram fatores de virulência suficientes para serem consideradas PAIs.

Tabela 12 – Quantidade de PAIs e GEIs preditos pelo GIPSy com relação aos fatores de virulência.

Linhagem	PAIs	GEIs
Cp MEX1	14	5
Cp MEX9	15	5
Cp MEX25	13	5
Cp MEX29	13	5
Cp MEX30	11	9
Cp MEX31	7	10

Pode-se observar que todas as linhagens pertencentes ao biovar Ovis apresentam entre 13-15 PAIs e cinco GEIs. Tanto o resultado de PAIs como o resultado de GEIs contrasta com os resultados encontrados no biovar Equi, o qual apresentou menos PAIs (11 e 7) e mais GEIs (9 e 10). Com relação à localização e tamanho as PAIs preditas, pôde-se notar que há certa conservação dentro do biovar. No biovar Ovis identificou-se oitos PAIs semelhantes, podendo-se destacar que todas as PAIs e GEIs das linhagens

MEX25 e MEX29, linhagens isoladas de ovelha, são muito semelhantes e se apresentam nas mesmas regiões do genoma. As linhagens do biovar Equi apresentaram essa semelhança na maioria das PAIs e GEIs preditas.

Com o objetivo de representar o biovar Ovis, as PAIs e GEIs preditas para a linhagem MEX1 são apresentadas na Figura 19 e no ANEXO V - Tabela adicional 3, e para representar o biovar Equi as PAIs e GEIs preditas para a linhagem MEX30 foram apresentadas na Figura 20 e na ANEXO V - Tabela adicional 4. A representação visual dos resultados das outras linhagens é apresentada nas figuras no ANEXO IV: Figura adicional 6 (MEX9), Figura adicional 7 (MEX25), Figura adicional 8 (MEX29) e Figura adicional 9 (MEX31). Entre as PAIs preditas, foram observadas divergências nas regiões em que estas foram preditas, e estas serão destacadas a seguir.

Na Figura 19, as regiões em que foram preditas as PAIs 4, 5, 6, 8 e 13 nas *C. pseudotuberculosis* MEX1 apresentam uma região divergente em relação às linhagens do biovar Equi.

Analisando a região que compreende a PAI4 nos seis genomas, pode-se observar que esta foi predita como PAI2 na linhagem MEX9 e PAI1 nas demais e que há conservação dos genes com função conhecida, assim como pequena variação no número de proteínas hipotéticas (apenas as linhagens do biovar Equi variam das demais). Com relação às PAIs 5, 6, 8 e 13 que podem ser vistas na Figura 19, os genes destas regiões estão conservados no biovar Ovis, e a correspondência dessas PAIs nas outras linhagens deste biovar pode ser observada na Tabela 13.

Tabela 13 – Correspondência entre as PAIs preditas nas linhagens do biovar Ovis.

Cp MEX1	*Cp* MEX9	*Cp* MEX25 e MEX29
PAI5	PAI3	PAI2
PAI6	PAI4	PAI3
PAI8	PAI5	PAI4
PAI13	PAI11	PAI10

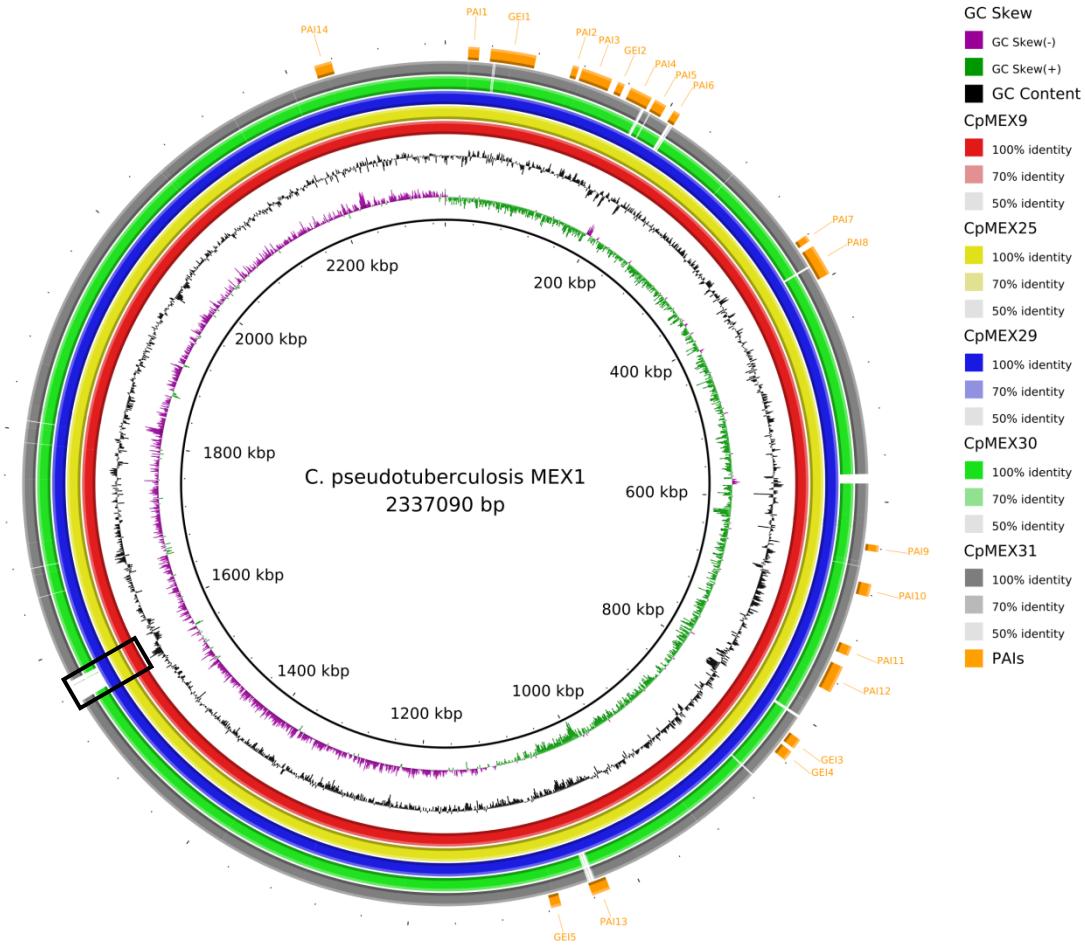


Figura 19 - Visualização circular dos genomas dos organismos estudados tendo como genoma de referência a *C. pseudotuberculosis* MEX1 (ao centro) e *C. pseudotuberculosis* MEX31 (mais externo).

Em laranja as PAIs e GEIs preditas pelo GIPSy para a *C. pseudotuberculosis* MEX1. O retângulo preto destaca a região do *cluster* gênico do sistema RM.

Figura gerada pelo BRIG.

Nos organismos *C. pseudotuberculosis* MEX9, MEX25 e MEX29 (ANEXO IV - Figuras adicionais 6, 7 e 8 respectivamente) foi identificada uma GEI chamada GEI4 na região do *cluster* gênico associado ao sistema RM, o que contribui com a hipótese deste *cluster* ter sido adquirido por transferência horizontal. Esta GEI não foi encontrada na linhagem MEX1, mesmo contendo o *cluster* gênico do sistema RM na mesma região (região destacada com um retângulo preto na Figura 19).

Além disso, na *C. pseudotuberculosis* MEX1 pode ser observada a predição da PAI1 (Figura 19) que contém o gene *pld* e o *cluster* gênico *fagABC* e o gene *fagD*. Isto corresponde aos resultados encontrados por Ruiz e

colaboradores (2011) nas linhagens *C. pseudotuberculosis* 1002 e C231, no qual estes genes foram encontrados na PAI descrita como PiCp1.

Nesta mesma região a linhagem MEX31 apresenta a GEI1 (ANEXO IV - Figura adicional 9), na qual além dos genes citados acima foi encontrado o *cluster* associado ao sistema CRISPR-Cas. Isso vai de encontro com a hipótese deste ter sido adquirido por transferência horizontal, porém na outra linhagem do biovar Equi (MEX30) a mesma não foi predita como GEI e foi observado a ausência de duas proteínas hipotéticas nesta região. Esta região pode ser observada no retângulo preto da Figura 20.

Na PAI2 predita na linhagem MEX1 foram identificados os genes *mntA*, *mntB*, *mntC* e *mntD*, os quais estão associados a aquisição de manganês e pertencem a família ABC de transportadores. Em bactérias Gram-positivas o gene *mntA* codifica uma lipoproteína e em bactéria Gram-negativas este codifica uma proteína responsável pela entrega de manganês ao complexo permease, o gene *mntB* codifica uma subunidade ATPase, esta catalisa a decomposição de trifosfato de adenosina (ATP), e os produtos dos genes *mntC* e *mntD* são proteínas de membrana que medeiam a importação de cátion (JENSEN & JENSEN, 2014). Assim como no processo de captação de ferro, a captação de outros metais como o manganês é essencial para o crescimento e viabilidade bacteriana, porém o excesso destes pode ser tóxico para a bactéria, sendo importante manter a homeostase de metais para sobrevivência celular. O operon *mntABCD* é regulado pela proteína *MntR* (regulador de transporte de manganês) (PANDEY *et al*., 2015). O Manganês pode também atuar no metabolismo e vias regulatórias associadas ao estresse oxidativo (PAPP-WALLACE & MAGUIRE, 2006). Além disso, a aquisição de manganês é importante para virulência em vários patógenos como *Streptococcus pneumoniae, Yersinia pestis* e *Brucella abortus* (BERRY & PATON, 1996; BEARDEN & PERRY, 1999; ANDERSON *et al*., 2009). Em um estudo com *Mycobacterium tuberculosis* PANDEY e colaboradores (2015) inativaram o operon *mntABCD,* o que reduziu o seu crescimento sob condição de redução de manganês.

A predição da PAI3 e da PAI6 da MEX30 (Figura 20) e da PAI12 MEX1 identificou um *locus* que contém os genes *oppA, oppB, oppC* e *oppD*, os quais são organizados em operon (HOGARTH & HIGGINS, 1983) e pertencem a

família de transportadores ABC que por sua vez utilizam a energia da hidrólise de ATP para fazer o transporte de peptídeos, lipídeos e sacarídeos pela membrana plasmática (BRAIBANT *et al*., 2000; MONNET, 2003). O produto destes genes está envolvido no processo de nutrição celular, através do transporte de peptídeos, e em bactérias Gram-positivas pode funcionar como um mecanismo de sinalização celular, tendo participação em mecanismos de controle transcricional de genes envolvidos em virulência (SAMEN *et al*., 2004). O produto do gene *oppA* faz a captura dos peptídeos, dos genes *oppB* e *oppC* forma o canal transmembrana para o transporte de peptídeos e, do gene *oppD* é responsável pela hidrólise de ATP e assim, fornecer energia para o transporte dos peptídeos (HIRON *et al*., 2007; MORAES *et al*., 2014). Um estudo em *C. pseudotuberculosis* mostrou que a ausência do gene *oppD* levou a um atraso na capacidade de adesão e viabilidade da bactéria à membrana celular em um modelo murino (MORAES *et al*., 2014).

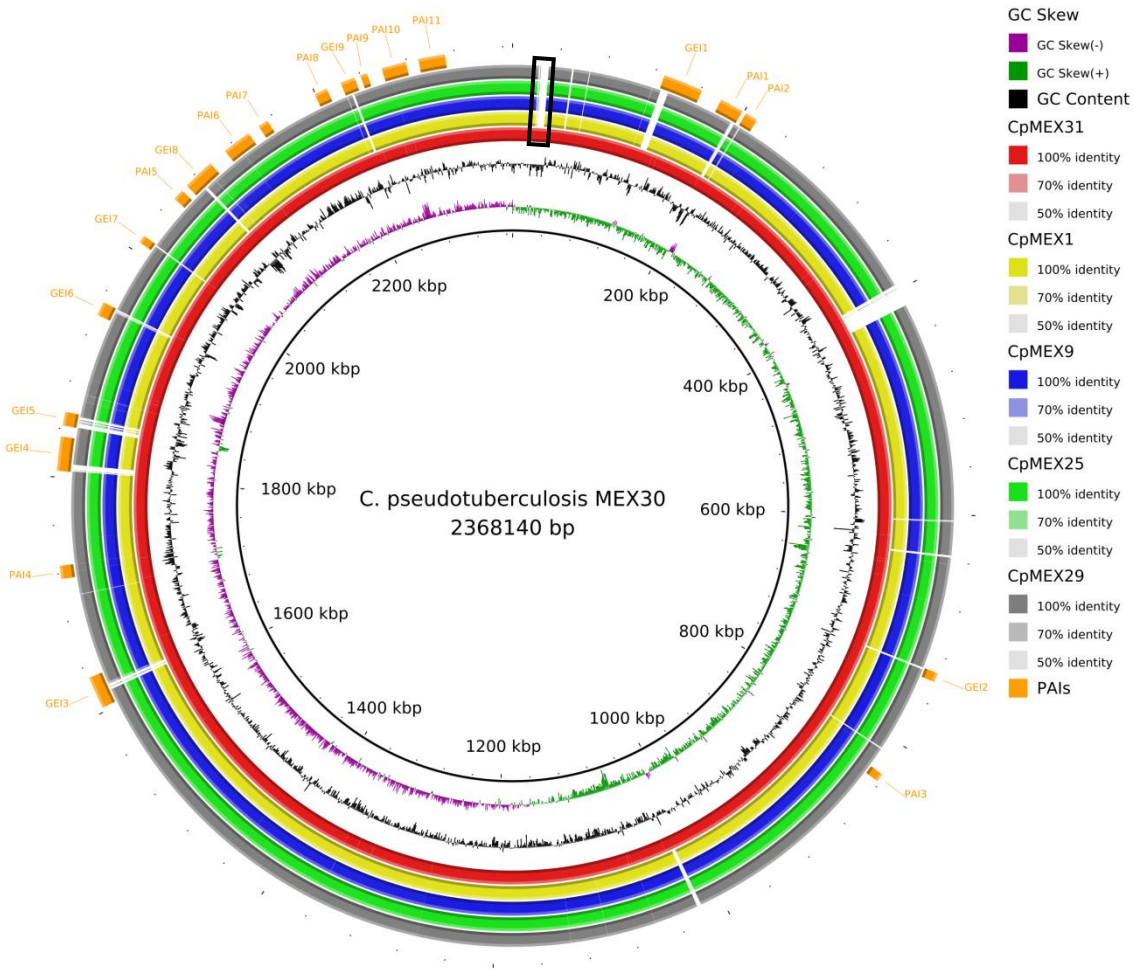


Figura 20 - Visualização circular dos genomas dos organismos estudados tendo como genoma de referência a *C. pseudotuberculosis* MEX30 (ao centro) e *C. pseudotuberculosis* MEX29 (mais externo).

Em laranja as PAIs e GEIs preditas pelo GIPSy para a *C. pseudotuberculosis* MEX30. O retângulo preto destaca a região do *cluster* gênico do sistema CRISPR-Cas.

Figura gerada pelo BRIG.

5.4.4. Alvos preditos para drogas

A Tabela 14 apresenta o possível alvo para drogas encontrado pelo *Specialty Genes Search* do PATRIC.

O gene *nrdF2* foi encontrado como candidato para cinco das seis linhagens estudadas, apesar deste estar presente nas seis linhagens, o mesmo não foi predito como candidato na *C. pseudotuberculosis* MEX30. O produto deste gene é uma subunidade da enzima redutora de ribonucleotídeos (*ribonucleotide reductase* (RNR)), a qual é classificada em três classes (I, II, III). A classe I é dependente de oxigênio e pode ser dividida em classe Ia e Ib,

em ambas a enzima responsável pela redução de ribonucleotídeos para a síntese de DNA é possui uma subunidade maior (*alpha*) e uma subunidade menor (*beta*). Na classe Ia a subunidade maior é codifica pelo gene *nrdA* e a subunidade menor pelo gene *nrdB*, na classe Ib estas são codificadas pelos genes *nrdE* e *nrdF*, respectivamente (KOLBERG *et al*., 2004).

RNRs classe II são codificados pelos genes *nrdJ* ou *RTPR*, estes são encontrados em bactérias que podem sobreviver em condições aeróbicas ou anaeróbicas. Utiliza o cobalto como cofator e juntamente com uma cisteína faz a redução de ribonucleotídeos. A classe III também é composta de uma subunidade maior, codificada pelo gene *nrdD*, e uma subunidade menor, codificada pelo gene *nrdG*, e é encontrada em alguns bacteriófagos e bactéria anaeróbicas estritas ou facultativas (KOLBERG *et al*., 2004).

O gene alvo predito pertence a classe Ib, pois além do gene *nrdF2* os organismos possuem o gene *nrdE*, o qual foi predito como *frameshift* na linhagem MEX30, o que pode ter influenciado para o gene *nrdF2* não ter sido apontado como possível candidato. Estes genes já foram encontrados em várias espécies de *Mycobacterium* como *tuberculosis*, *leprae* e *ulcerans* (MOWA, 2008) e mesmo não estando arranjados na forma de operon, aparentemente são regulados conjuntamente em resposta ao aumento da depleção de oxigênio (DAWES *et al*., 2003). Além disso, o gene *nrdF2* é importante para o crescimento de *M. tuberculosis*, *C. ammoniagenes* e *glutamicum* sob condições normais *in vitro* (OEHLMANN & AULING, 1999; TORRENTS *et al*., 2003; MOWA, 2008) e tem sido apontado como potencial alvo para o desenvolvimento de vacinas em *M. tuberculosis* (YANG *et al*., 1997; NURBO *et al*., 2007; MOWA, 2008).

Tabela 14 – Alvo para drogas predito pelo *Specialty Genes Search*.

Linhagem	Gene	Produto	Organismo	*Subject Coverage*	*Query Coverage*	Identidade %
Cp MEX1	nrdF2	Ribonucleotide reductase of class Ib (aerobic), beta subunit (EC 1.17.4.1)	*M. leprae*	96	92	80
Cp MEX9	nrdF2	Ribonucleotide reductase of class Ib (aerobic), beta subunit (EC 1.17.4.1)	*M. leprae*	96	92	80
Cp MEX25	nrdF2	Ribonucleotide reductase of class Ib (aerobic), beta subunit (EC 1.17.4.1)	*M. leprae*	96	92	80
Cp MEX29	nrdF2	Ribonucleotide reductase of class Ib (aerobic), beta subunit (EC 1.17.4.1)	*M. leprae*	96	92	80
Cp MEX31	nrdF2	Ribonucleotide reductase of class Ib (aerobic), beta subunit (EC 1.17.4.1)	*M. leprae*	96	92	80

6. CONCLUSÕES

Este trabalho apresentou o primeiro estudo e caracterização *in silico* de seis linhagens de *C. pseudotuberculosis* isoladas no México. As informações contribuíram para um melhor conhecimento biológico do organismo e abriu perspectivas para novos estudos relacionados ao desenvolvimento de diagnósticos, vacinas e drogas para *C. pseudotuberculosis*, um patógeno de interesse médico-veterinário que causa danos financeiros ao agronegócio no Brasil e no mundo.

Os organismos deste estudo mostraram semelhanças no tamanho do genoma (~2,3 Mb) e conteúdo GC (52%), assim como os outros organismos já sequenciados desta espécie. Todas as linhagens deste estudo foram depositadas no NCBI.

O estudo de pangenoma com o *Protein Family Sorter* do PATRIC apontou os valores do pangenoma, genoma central e acessório, e de genes únicos. Apresentando também a quantidade de genes exclusiva de cada biovar, o que possibilitou a identificação dos *clusters* gênicos do sistema RM no biovar Ovis e CRISPR-Cas no biovar Equi.

A análise filogenômica com o Gegenees e o FigTree mostrou um alinhamento por biovar e, também uma similaridade genômica sutilmente maior entre os isolados do mesmo hospedeiro. As linhagens MEX25 e MEX29, ambas isoladas de ovelha, são as mais próximas filogenomicamente e também o GIPSy predisse o mesmo número de PAIs e GEIs nas mesmas regiões. Tais informações podem ser utilizadas em futuros estudos para possivelmente compreender a *C. pseudotuberculosis* a nível de hospedeiro.

Mediante ao nosso resultado, a partir da arvore filogenômica acreditamos estar de acordo com o modelo evolutivo discutido Oliveira e colaboradores (2016), onde possivelmente essas linhagens podem estar apresentando um evento de especiação biológica, no qual futuramente possam vir a se destacar como possíveis espécies a partir do evento de anagênese. Foi possível concluir, também, que os biovares apresentam características genéticas exclusivas e consecutivamente em seu conteúdo proteico sendo, assim, capaz de distinguir as linhagens por biovares.

Com relação à predição de PAIs, observou-se uma maior conservação das PAIs em cada biovar. Também foram encontrados operons conservados previamente descritos na literatura pertencentes à superfamília de transportadores ABC que estão relacionados a virulência em patógenos.

Além disso, este trabalho identificou um alvo para drogas, o gene *nrdF2* já descrito anteriormente em organismos do gênero *Mycobacterium*, o que pode levar ao estudo deste em *C. pseudotuberculosis*.

7. PERSPECTIVAS

Este trabalho tem como perspectivas:

- Analisar os resultados já obtidos a partir da anotação e curadoria manual de pangenoma com o *software* BPGA (*Bacterial Pan Genome Analysis tool*) (CHAUDHARI *et al.*, 2016).
- Comparar os resultados de pangenoma obtidos a partir do BPGA e do *Protein Family Sorter* das seis linhagens de *C. pseudotuberculosis* deste trabalho, e todos os outros genomas desta espécie disponíveis no NCBI. Além disso, comparar as ferramentas utilizadas.
- Utilizar o *software* GIPSy para novas análises de PAIs e GEIs com a anotação manual.
- Detalhar os resultados obtidos nas análises de genômica comparativa em relação às PAIs e GEIs preditas neste trabalho e avaliar os alvos envolvidos com virulência e patogenicidade.
- Avaliar a diferença entre a anotação manual e anotação automática e destacar o impacto da curadoria manual nas análises de genômica comparativa.

8. REFERÊNCIAS

ALGAMMAL, Abdelazeem M. Molecular Characterization and Antibiotic Susceptibility of *Corynebacterium pseudotuberculosis* Isolated from Sheep and Goats Suffering from Caseous Lymphadenitis. **Zagazig Veterinary Journal** (Zag. Vet. J.), v. 44, n. 1, 2016.

ALIKHAN, Nabil-Fareed *et al*. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. **BMC genomics**, v. 12, n. 1, p. 1, 2011.

ALTSCHUL, Stephen F. *et al*. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403-410, 1990.

AMBARDAR, Sheetal *et al*. High Throughput Sequencing: An Overview of Sequencing Chemistry. **Indian Journal of Microbiology**, p. 1-11, 2016.

ANDERSON, Eric S. *et al*. The manganese transporter MntH is a critical virulence determinant for *Brucella abortus* 2308 in experimentally infected mice. **Infection and immunity**, v. 77, n. 8, p. 3466-3474, 2009.

ÅGREN, Joakim *et al*. Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. **PloS one**, v. 7, n. 6, p. e39107, 2012.

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, 2012.

BARBA, Marta *et al*. Experimental transmission of *Corynebacterium pseudotuberculosis* biovar *equi* in horses by house flies. **Journal of Veterinary Internal Medicine**, v. 29, n. 2, p. 636-643, 2015.

BARBOSA, Eudes *et al*. On the limits of computational functional genomics for bacterial lifestyle prediction. **Briefings in functional genomics**, p. elu014, 2014.

BARCELLOS, Fernando Gomes *et al*. Evidence of horizontal transfer of symbiotic genes from a *Bradyrhizobium japonicum* inoculant strain to indigenous *Diazotrophs sinorhizobium (ensifer) fredii* and *Bradyrhizobium elkanii* in a Brazilian Savannah soil. **Applied and environmental microbiology**, v. 73, n. 8, p. 2635-2643, 2007.

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

BARH, Debmalya *et al*. A novel comparative genomics analysis for common drug and vaccine targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogens. **Chemical biology & drug design**, v. 78, n. 1, p. 73-84, 2011b.

BEARDEN, Scott W.; PERRY, Robert D. The Yfe system of *Yersinia pestis* transports iron and manganese and is required for full virulence of plague. **Molecular microbiology**, v. 32, n. 2, p. 403-414, 1999.

BERGLUND, Eva C.; KIIALAINEN, Anna; SYVÄNEN, Ann-Christine. Next-generation sequencing technologies and applications for human genetic history and forensics. **Investigative Genetics**, v. 2, n. 1, p. 1, 2011.

BERRY, Anne M.; PATON, James C. Sequence heterogeneity of PsaA, a 37-kilodalton putative adhesin essential for virulence of *Streptococcus pneumoniae*. **Infection and immunity**, v. 64, n. 12, p. 5255-5262, 1996.

BILLINGTON, Stephen J. *et al*. Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*. **FEMS microbiology letters**, v. 208, n. 1, p. 41-45, 2002.

BIBERSTEIN, E. L.; KNIGHT, H. D.; JANG, S. Two biotypes of *Corynebacterium pseudotuberculosis*. **Veterinary Record**, v. 89, n. 26, p. 691-692, 1971.

BLOW, Matthew J. *et al*. The epigenomic landscape of prokaryotes. **PLoS Genet**, v. 12, n. 2, p. e1005854, 2016.

BRAIBANT, Martine; GILOT, Philippe; CONTENT, Jean. The ATP binding cassette (ABC) transport systems of *Mycobacterium tuberculosis*. **FEMS microbiology reviews**, v. 24, n. 4, p. 449-467, 2000.

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.

CAMACHO, Christiam *et al*. BLAST+: architecture and applications. **BMC bioinformatics**, v. 10, n. 1, p. 1, 2009.

CARVER, Tim *et al*. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. **Bioinformatics**, v. 28, n. 4, p. 464-469, 2012.

CASADESÚS, Josep; LOW, David. Epigenetic gene regulation in the bacterial world. **Microbiology and molecular biology reviews**, v. 70, n. 3, p. 830-856, 2006.

CHAN, Cheong Xin; RAGAN, Mark A. Next-generation phylogenomics. **Biology Direct**, v. 8, n. 1, p. 1, 2013.

CHAUDHARI, Narendrakumar M.; GUPTA, Vinod Kumar; DUTTA, Chitra. BPGA-an ultra-fast pan-genome analysis pipeline. **Scientific Reports**, v. 6, 2016.

CHEN, Yen-Chun *et al*. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. **PloSone**, v. 8, n. 4, p. e62856, 2013.

CHEVREUX, Bastien *et al*. Genome sequence assembly using trace signals and additional sequence information. In: **German conference on bioinformatics**. 1999. p. 45-56.

CHEVREUX, Bastien *et al*. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. **Genome research**, v. 14, n. 6, p. 1147-1159, 2004.

CHIKHI, Rayan; RIZK, Guillaume. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. **Algorithms for Molecular Biology**, v. 8, n. 1, p. 1, 2013.

CHIN, Francis YL; LEUNG, Henry CM; YIU, S. M. Sequence assembly using next generation sequencing data—challenges and solutions. **Science China Life Sciences**, v. 57, n. 11, p. 1140-1148, 2014.

CHIU, Charles; MILLER, Steve. Next-Generation Sequencing. **Persing D, Tenover F, Hayden R, Ieven M, Miller**, v. 461, 2016.

CLCBIO. Em: <http://www.clcbio.com/>. Acesso em: 12 de abril, 2016.

COLLETT, M. G.; BATH, G. F.; CAMERON, C. M. *Corynebacterium pseudotuberculosis* infections. Infections diseases of livestock with special reference to Southern Africa. **Oxford University Press**, v. 2, p. 1387-1395, 1994.

COMPLETE GENOME SUBMISSION GUIDE. Em: <http://www.ncbi.nlm.nih.gov/genbank/genomesubmit/>. Acesso em: 09 de maio, 2016.

COSTA, Daniela Arruda, D.Sc., Universidade Federal de Viçosa, julho de 2015. **Genômica comparativa de linhagens de Saccharomyces e Kluyveromyces de interesse biotecnológico**. Tese de doutorado. Programa de Pós-Graduação em Microbiologia Agrícola da UFV, Viçosa (MG), 2015.

DA SILVA, Wanderson Marques**. Estudo do genoma funcional de *Corynebacterium pseudotuberculosis* através de diferentes estratégias proteômicas**. Tese de Doutorado. Programa de Pós-Graduação em Genética da UFMG, Belo Horizonte (MG), 2015.

DAVIS, James J. *et al*. PATtyFams: Protein Families for the Microbial Genomes in the PATRIC Database. **Frontiers in microbiology**, v. 7, 2016.

DAWES, Stephanie S. *et al*. Ribonucleotide reduction in *Mycobacterium tuberculosis*: function and expression of genes encoding class Ib and class II ribonucleotide reductases. **Infection and immunity**, v. 71, n. 11, p. 6124-6131, 2003.

DE RESENDE, Vívian D'afonseca Da Silva. **Genômica estrutural de *Corynebacterium pseudotuberculosis* e estudos de gênômica comparativa com espécies do genêro**. Dissertação de mestrado. Programa de Pós-Graduação em Genética da UFMG, Belo Horizonte (MG), 2011.

DIAS, Larissa Maranhão. **Estudo de genômica comparativa de *Corynebacterium pseudotuberculosis* linhagem 226 (biovar ovis)**. Dissetação de Mestrado. Programa de Pós Graduação em Biotecnologia da UFPA, Belém (PA), 2015.

DOBRINDT, U. *et al*. Toxin genes on pathogenicity islands: impact for microbial evolution. **International journal of medical microbiology**, v. 290, n. 4, p. 307-311, 2000.

DOBRINDT, Ulrich; HACKER, Jörg. Whole genome plasticity in pathogenic bacteria. **Current opinion in microbiology**, v. 4, n. 5, p. 550-557, 2001.

DORELLA, Fernanda Alves *et al*. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. **Veterinary research**, v. 37, n. 2, p. 201-218, 2006.

EISEN, Jonathan A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. **Genome research**, v. 8, n. 3, p. 163-167, 1998.

FLEISCHMANN, Robert D. *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. **Science**, v. 269, n. 5223, p. 496-512, 1995.

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. 1, 2011.

GIRE, Stephen K. *et al*. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. **Science**, v. 345, n. 6202, p. 1369-1372, 2014.

GOLD. Em: <https://gold.jgi.doe.gov/statistics>. Acesso em: 31 de março, 2016.

GOUY, Manolo; GUINDON, Stéphane; GASCUEL, Olivier. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. **Molecular biology and evolution**, v. 27, n. 2, p. 221-224, 2010.

GUEDES, Maria T. *et al*. Infecção por *Corynebacterium pseudotuberculosis* em Equinos: aspectos microbiológicos, clínicos e preventivos1. **Pesq. Vet. Bras**, v. 35, n. 8, p. 701-708, 2015.

GUREVICH, Alexey *et al*. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, p. btt086, 2013.

HEAD, Steven R. *et al*. Library construction for next-generation sequencing: overviews and challenges. **Biotechniques**, v. 56, n. 2, p. 61, 2014.

HIRON, Aurelia *et al*. Only one of four oligopeptide transport systems mediates nitrogen nutrition in *Staphylococcus aureus*. **Journal of bacteriology**, v. 189, n. 14, p. 5119-5129, 2007.

HOCHHUT, Bianca; DOBRINDT, Ulrich; HACKER, Jörg. Pathogenicity islands and their role in bacterial virulence and survival. In: **Concepts in Bacterial Virulence**. Karger Publishers, 2004. p. 234-254.

HOGARTH, Barbara G.; HIGGINS, Christopher F. Genetic organization of the oligopeptide permease (opp) locus of Salmonella typhimurium and Escherichia coli. **Journal of bacteriology**, v. 153, n. 3, p. 1548-1551, 1983.

HUSEMANN, Peter. **Bioinformatic approaches for genome finishing**. 2011. Tese de Doutorado. Bielefeld, Univ., Diss., 2011.

JEBER, Z. K. H. *et al*. Influence of *Corynebacterium pseudotuberculosis* infection on level of acute phase proteins in goats. **BMC veterinary research**, v. 12, n. 1, p. 1, 2016.

JENSEN, Amornrat Naranuntarat; JENSEN, Laran T. Manganese Transport, Trafficking and Function in Invertebrates. **Manganese in Health and Disease**, p. 1, 2014.

KAUR, Ritesh; MALIK, Chander Parkash. NEXT GENERATION SEQUENCING: A REVOLUTION IN GENE SEQUENCING. **CIBTech Journal of Biotechnology** 2 (4): 1-20. 2013.

KILCOYNE, Isabelle *et al*. Frequency of *Corynebacterium pseudotuberculosis* infection in horses across the United States during a 10-year period. **Journal of the American Veterinary Medical Association**, v. 245, n. 3, p. 309-314, 2014.

KOBAYASHI, Ichizo *et al*. Shaping the genome–restriction–modification systems as mobile genetic elements. **Current opinion in genetics & development**, v. 9, n. 6, p. 649-656, 1999.

KOLBERG, Matthias *et al*. Structure, function, and mechanism of ribonucleotide reductases. **Biochimica et Biophysica Acta (BBA)-Proteins and Proteomic**s, v. 1699, n. 1, p. 1-34, 2004.

KRIZOVA, Lenka; NEMEC, Alexandr. A 63 kb genomic resistance island found in a multidrug-resistant *Acinetobacter baumannii* isolate of European clone I from 1977. **Journal of antimicrobial chemotherapy**, p. dkq223, 2010.

KWIATEK, Agnieszka *et al*. Type III Methyltransferase M. NgoAX from Neisseria gonorrhoeae FA1090 Regulates Biofilm Formation and Interactions with Human Cells. **Frontiers in microbiology**, v. 6, 2015.

LANDER, Eric S. *et al*. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860-921, 2001.

LECUIT, Marc; ELOIT, Marc. The potential of whole genome NGS for infectious disease diagnosis. **Expert review of molecular diagnostics**, v. 15, n. 12, p. 1517-1519, 2015.

LIU, Lin *et al*. Comparison of next-generation sequencing systems. **BioMed Research International**, v. 2012, 2012.

LIU, Tsunglin *et al*. Optimizing information in next-generation-sequencing (NGS) reads for improving de novo genome assembly. **PloSone**, v. 8, n. 7, p. e69503, 2013.

LOMAN, Nicholas J. *et al*. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. **Nature Reviews Microbiology**, v. 10, n. 9, p. 599-606, 2012.

LOMAN, Nicholas J.; PALLEN, Mark J. Twenty years of bacterial genome sequencing. **Nature Reviews Microbiology**, 2015.

LUCIANI, Fabio; BULL, Rowena A.; LLOYD, Andrew R. Next generation deep sequencing and vaccine design: today and tomorrow. **Trends in biotechnology**, v. 30, n. 9, p. 443-452, 2012.

MAKAROVA, Kira S. *et al*. Evolution and classification of the CRISPR–Cas systems. **Nature Reviews Microbiology**, v. 9, n. 6, p. 467-477, 2011.

MAKAROVA, Kira S. *et al*. An updated evolutionary classification of CRISPR-Cas systems. **Nature Reviews Microbiology**, 2015.

MAO, Chunhong *et al*. Curation, integration and visualization of bacterial virulence factors in PATRIC. **Bioinformatics**, v. 31, n. 2, p. 252-258, 2015.

MARDIS, Elaine R. A decade/'s perspective on DNA sequencing technology. **Nature**, v. 470, n. 7333, p. 198-203, 2011.

MARDIS, Elaine R. Next-generation sequencing platforms. **Annual review of analytical chemistry**, v. 6, p. 287-303, 2013.

MARIANO, Diego César Batista. **SIMBA: uma ferramenta Web para gerenciamento de montagens de genomas bacterianos**. Dissertação de mestrado. Programa de Pós-Graduação em Bioinformática da UFMG, Belo Horizonte (MG), 2015.

MARINIER, Eric; BROWN, Daniel G.; MCCONKEY, Brendan J. Pollux: platform independent error correction of single and mixed genomes. **BMC bioinformatics**, v. 16, n. 1, p. 1, 2015.

MARKOWITZ, Victor M. *et al*. IMG ER: a system for microbial genome annotation expert review and curation. **Bioinformatics**, v. 25, n. 17, p. 2271-2278, 2009.

MCKUSICK, Victor A.; RUDDLE, Frank H. A new discipline, a new name, a new journal. **Genomics**, v. 1, n. 1, p. 1-2, 1987.

MEYER, Folker; OVERBEEK, Ross; RODRIGUEZ, Alex. FIGfams: yet another set of protein families. **Nucleic acids research**, v. 37, n. 20, p. 6643-6654, 2009.

MÉDIGUE, Claudine; MOSZER, Ivan. Annotation, comparison and databases for hundreds of bacterial genomes. **Research in microbiology**, v. 158, n. 10, p. 724-736, 2007.

MILLER, Jason R.; KOREN, Sergey; SUTTON, Granger. Assembly algorithms for next-generation sequencing data. **Genomics**, v. 95, n. 6, p. 315-327, 2010.

MONNET, V. Bacterial oligopeptide-binding proteins. **Cellular and Molecular Life Sciences CMLS**, v. 60, n. 10, p. 2100-2114, 2003.

MORAES, Pablo MRO *et al*. Characterization of the Opp peptide transporter of *Corynebacterium pseudotuberculosis* and its role in virulence and pathogenicity. **BioMed research international**, v. 2014, 2014.

MOWA, Mohube Betty. **Function and expression of class I ribonucleotide reductase small subunit-encoding genes in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis***. 2008. Tese de Doutorado. Faculty of Science, University of the Witwatersrand, Johannesburg.

MUNROE, David J.; HARRIS, Timothy JR. Third-generation sequencing fireworks at Marco Island. **Nature biotechnology**, v. 28, n. 5, p. 426-428, 2010.

MUÑOZ-BUCIO, A. V. *et al*. Identification of *Corynebacterium pseudotuberculosis* isolated from muscular abscesses in two horses: First report in Mexico. **Equine Veterinary Education**, 2016.

NAGARAJAN, Niranjan; POP, Mihai. Sequence assembly demystified. **Nature Reviews Genetics**, v. 14, n. 3, p. 157-167, 2013.

NAM, Ki Hyun et al. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype IC/Dvulg CRISPR-Cas system. **Structure**, v. 20, n. 9, p. 1574-1584, 2012.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE. Em:<https://www.genome.gov/sequencingcosts/>. Acesso em 05 de abril, 2016.

NURBO, Johanna *et al*. Design, synthesis and evaluation of peptide inhibitors of *Mycobacterium tuberculosis* ribonucleotide reductase. **Journal of Peptide Science**, v. 13, n. 12, p. 822-832, 2007.

OCHMAN, Howard; DAVALOS, Liliana M. The nature and dynamics of bacterial genomes. **Science**, v. 311, n. 5768, p. 1730-1733, 2006.

OEHLMANN, Wulf; AULING, Georg. Ribonucleotide reductase (RNR) of *Corynebacterium glutamicum* ATCC 13032–genetic characterization of a second class IV enzyme. **Microbiology**, v. 145, n. 7, p. 1595-1604, 1999.

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

OLIVEIRA, Letícia de Castro. **Análise do potencial probiótico de Lactococcus lactis subsp. lactis NCDO 2118 por meio de genômica comparativa**. Dissertação de mestrado. Programa de Pós-Graduação em Bioinformática da UFMG, Belo Horizonte (MG), 2014.

OXFORD NANOPORE TECHNOLOGIES. Em: <https://nanoporetech.com/products-services/minion-mki>. Acesso em 06 de abril, 2016.

OVERBEEK, Ross *et al*. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). **Nucleic acids research**, v. 42, n. D1, p. D206-D214, 2014.

PACBIO SYSTEMS. Em: <http://www.pacb.com/products-and-services/pacbio-systems/>. Acessoem: 06 deabril, 2016.

PACHECO, Luis Gustavo Carvalho. **DESENVOLVIMENTO DE UM ENSAIO DE PCR-MULTIPLEX PARA IDENTIFICAÇÃO DE ISOLADOS DE *CORYNEBACTERIUM PSEUDOTUBERCULOSIS* E RÁPIDA DETECÇÃO DESSA BACTÉRIA EM AMOSTRAS CLÍNICAS**. Dissertação de mestrado. Programa de Pós-Graduação em Genética da UFMG, Belo Horizonte (MG), 2006.

PACHECO, Luis GC 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, 2007.

PANDEY, Ruchi *et al*. MntR (Rv2788): a transcriptional regulator that controls manganese homeostasis in *Mycobacterium tuberculosis*. **Molecular microbiology**, v. 98, n. 6, p. 1168-1183, 2015.

PAPP-WALLACE, Krisztina M.; MAGUIRE, Michael E. Manganese transport and the role of manganese in virulence. **Annu. Rev. Microbiol**., v. 60, p. 187-209, 2006.

PAULING, Josch *et al*. On the trail of EHEC/EAEC—unraveling the gene regulatory networks of human pathogenic *Escherichia coli* bacteria. **Integrative Biology**, v. 4, n. 7, p. 728-733, 2012.

PEEL, Margaret M. *et al*. Human lymphadenitis due to *Corynebacterium pseudotuberculosis*: report of ten cases from Australia and review. **Clinical Infectious Diseases**, v. 24, n. 2, p. 185-191, 1997.

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, 2016.

PHILIPPE, Hervé *et al*. Phylogenomics. **Annual Review of Ecology, Evolution, and Systematics**, p. 541-562, 2005.

PINTO, Anne Cybelle. **Análise em larga escala da expressão diferencial de *Corynebacterium pseudotuberculosis* em resposta a estresses abióticos**. Tese de doutorado. Programa de Pós-Graduação do Departamento de Microbiologia do Instituto de Ciências Biológicas da UFMG, Belo Horizonte (MG), 2011.

PINTO, Anne Cybelle *et al*. Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses. **BMC genomics**, v. 15, n. 1, p. 1, 2014.

POP, Mihai. Genome assembly reborn: recent computational challenges. **Briefings in bioinformatics**, v. 10, n. 4, p. 354-366, 2009.

POURCEL, C.; SALVIGNOL, G.; VERGNAUD, Gilles. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. **Microbiology**, v. 151, n. 3, p. 653-663, 2005.

PURANIK, Rutika *et al*. A pipeline for completing bacterial genomes using *in silico* and wet lab approaches. **BMC genomics**, v. 16, n. Suppl 3, p. S7, 2015.

PYBUS, Oliver G.; RAMBAUT, Andrew; HARVEY, Paul H. An integrated framework for the inference of viral population history from reconstructed genealogies. **Genetics**, v. 155, n. 3, p. 1429-1437, 2000.

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.

RICHARDSON, Emily J.; WATSON, Mick. The automatic annotation of bacterial genomes. **Briefings in bioinformatics**, v. 14, n. 1, p. 1-12, 2013.

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, 2011.

RUTHERFORD, Kim *et al*. Artemis: sequence visualization and annotation. **Bioinformatics**, v. 16, n. 10, p. 944-945, 2000.

SAMEN, Ulrike **et al**. Relevance of peptide uptake systems to the physiology and virulence of **Streptococcus agalactiae**. **Journal of bacteriology**, v. 186, n. 5, p. 1398-1408, 2004.

SANGAL, Vartul; FINERAN, Peter C.; HOSKISSON, Paul A. Novel configurations of type I and II CRISPR–Cas systems in *Corynebacterium diphtheriae*. **Microbiology**, v. 159, n. 10, p. 2118-2126, 2013.

SANGER, Frederick; NICKLEN, Steven; COULSON, Alan R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences**, v. 74, n. 12, p. 5463-5467, 1977.

SEIB, Kate L. *et al*. A novel epigenetic regulator associated with the hypervirulent *Neisseria meningitidis* clonal complex 41/44. **The FASEB Journal**, v. 25, n. 10, p. 3622-3633, 2011.

SERNOVA, Natalia V.; GELFAND, Mikhail S. Identification of replication origins in prokaryotic genomes. **Briefings in bioinformatics**, v. 9, n. 5, p. 376-391, 2008.

SEYFFERT, Nubia *et al*. Serological proteome analysis of *Corynebacterium pseudotuberculosis* isolated from different hosts reveals novel candidates for

prophylactics to control caseous lymphadenitis. **Veterinary microbiology**, v. 174, n. 1, p. 255-260, 2014.

SHELL, Scarlet S. *et al*. DNA methylation impacts gene expression and ensures hypoxic survival of *Mycobacterium tuberculosis*. **PLoS Pathog**, v. 9, n. 7, p. e1003419, 2013.

SIMPSON, Jared T.; POP, Mihai. The theory and practice of genome sequence assembly. **Annual review of genomics and human genetics**, v. 16, p. 153-172, 2015.

SIMS, David *et al*. Sequencing depth and coverage: key considerations in genomic analyses. **Nature Reviews Genetics**, v. 15, n. 2, p. 121-132, 2014.

SNYDER, E. E. *et al*. PATRIC: the VBI pathosystems resource integration center. **Nucleic acids research**, v. 35, n. suppl 1, p. D401-D406, 2007.

SOARES, Siomar C. *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, 2013.

SOARES, Siomar C. *et al*. GIPSy: genomic island prediction software. **Journal of biotechnology**, 2015.

SONGER, J. Glenn *et al*. Biochemical and genetic characterization of *Corynebacterium pseudotuberculosis*. **American journal of veterinary research**, v. 49, n. 2, p. 223-226, 1988.

SRIKHANTA, Yogitha N. *et al*. Phasevarions mediate random switching of gene expression in pathogenic *Neisseria*. **PLoS Pathog**, v. 5, n. 4, p. e1000400, 2009.

STEIN, Lincoln. Genome annotation: from sequence to biology. **Nature reviews genetics**, v. 2, n. 7, p. 493-503, 2001.

STOTHARD, Paul; WISHART, David S. Circular genome visualization and exploration using CGView. **Bioinformatics**, v. 21, n. 4, p. 537-539, 2005.

STOTHARD, Paul; WISHART, David S. Automated bacterial genome analysis and annotation. **Current opinion in microbiology**, v. 9, n. 5, p. 505-510, 2006.

SUCHARD, Marc A.; RAMBAUT, Andrew. Many-core algorithms for statistical phylogenetics. **Bioinformatics**, v. 25, n. 11, p. 1370-1376, 2009.

SUZUKI, Hirokazu. **Host-mimicking strategies in DNA methylation for improved bacterial transformation**. INTECH Open Access Publisher, 2012.

SUTHERLAND, S. S.; HART, R. A.; BULLER, N. B. Genetic differences between nitrate-negative and nitrate-positive *C. pseudotuberculosis* strains using restriction fragment length polymorphisms. **Veterinary microbiology**, v. 49, n. 1, p. 1-9, 1996.

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.

TAN, Aimee *et al*. Distribution of the type III DNA methyltransferases modA, modB and modD among *Neisseria meningitidis* genotypes: implications for gene regulation and virulence. **Scientific reports**, v. 6, 2016.

TAUCH, Andreas; SANDBOTE, Jasmin. The family *Corynebacteriaceae*. In: **The Prokaryotes**. Springer Berlin Heidelberg, 2014. p. 239-277.

TETTELIN, Hervé *et al*. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 39, p. 13950-13955, 2005.

THERMO FISHER SCIENTIFIC. Em: <https://www.thermofisher.com/br/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence.html>. Acesso em: 17 de abril, 2016.

TIWARI, Sandeep *et al*. *C. pseudotuberculosis* Phop confers virulence and may be targeted by natural compounds. **Integrative Biology**, v. 6, n. 11, p. 1088-1099, 2014.

TORRENTS, Eduard; ROCA, I.; GIBERT, I. *Corynebacterium ammoniagenes* class Ib ribonucleotide reductase: transcriptional regulation of an atypical genomic organization in the nrd cluster. **Microbiology**, v. 149, n. 4, p. 1011-1020, 2003.

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, 2010.

TROST, Eva *et al*. Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. **BMC genomics**, v. 12, n. 1, p. 1, 2011.

TSUI, Tsz Kin Martin; LI, Hong. Structure principles of CRISPR-Cas surveillance and effector complexes. **Annual review of biophysics**, v. 44, p. 229-255, 2015.

TUMAPA, Sarinna *et al*. *Burkholderia pseudomallei* genome plasticity associated with genomic island variation. **BMC genomics**, v. 9, n. 1, p. 1, 2008.

UNIPROT CONSORTIUM *et al*. Reorganizing the protein space at the Universal Protein Resource (UniProt). **Nucleic acids research**, p. gkr981, 2011.

VAN DIJK, Erwin L. *et al*. Ten years of next-generation sequencing technology. **Trends in genetics**, v. 30, n. 9, p. 418-426, 2014.

VENTER, J. Craig *et al*. The sequence of the human genome. **science**, v. 291, n. 5507, p. 1304-1351, 2001.

VERNIKOS, George *et al*. Ten years of pan-genome analyses. **Current opinion in microbiology**, v. 23, p. 148-154, 2015.

VINCENT, Antony T. *et al*. Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. **Journal of Microbiological Methods**, 2016.

VERLI, Hugo *et al*. Bioinformática da Biologia à flexibilidade molecular. **Porto Alegre, Brasil**, v. 1, 2014.

WAHEED, Yasir *et al*. Development of global consensus sequence and analysis of highly conserved domains of the HCV NS5B protein. **Hepatitis monthly**, v. 12, n. 9, 2012.

WATTAM, Alice R. *et al*. PATRIC, the bacterial bioinformatics database and analysis resource. **Nucleic acids research**, p. gkt1099, 2013.

WION, Didier; CASADESÚS, Josep. N6-methyl-adenine: an epigenetic signal for DNA–protein interactions. **Nature Reviews Microbiology**, v. 4, n. 3, p. 183-192, 2006.

YANG, Fude *et al*. Characterization of two genes encoding the *Mycobacterium tuberculosis* ribonucleotide reductase small subunit. **Journal of bacteriology**, v. 179, n. 20, p. 6408-6415, 1997.

ZHU, Lingxiang *et al*. Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. **Nucleic acids research**, p. gkv1498, 2015.

ANEXO I - FIGURAS DO QUAST

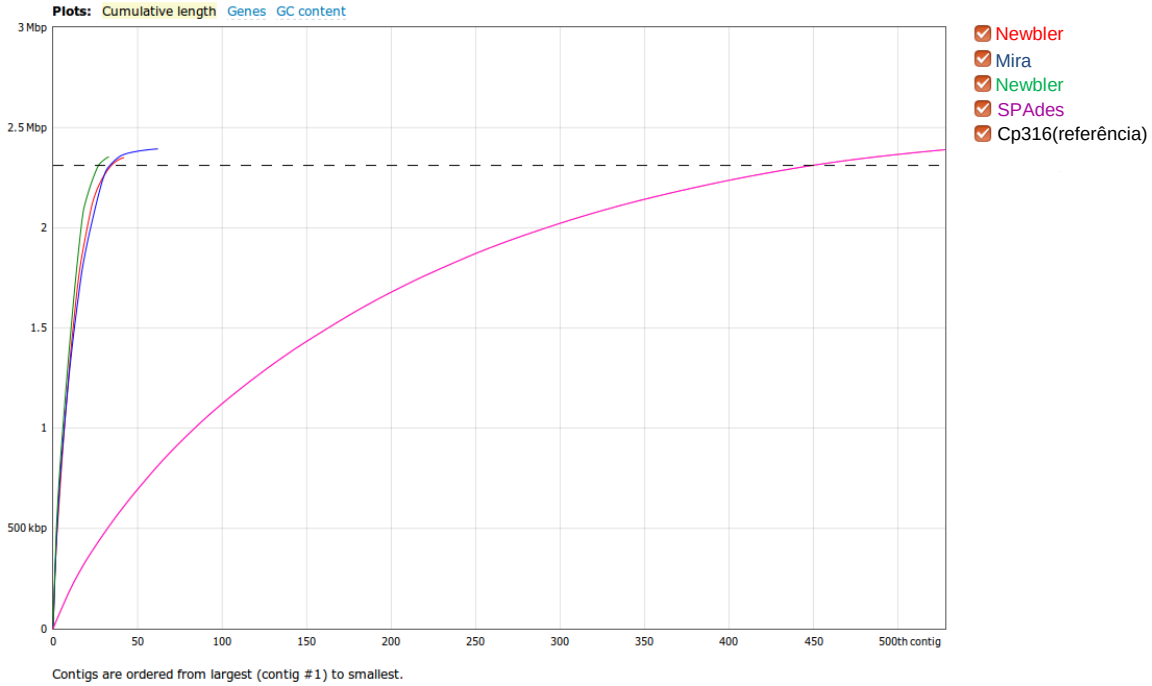


Figura adicional 1 - Gráfico de avaliação da qualidade da montagem da *C. pseudotuberculosis* MEX30 de acordo com o genoma de referência.

Figura gerada pelo QUAST.

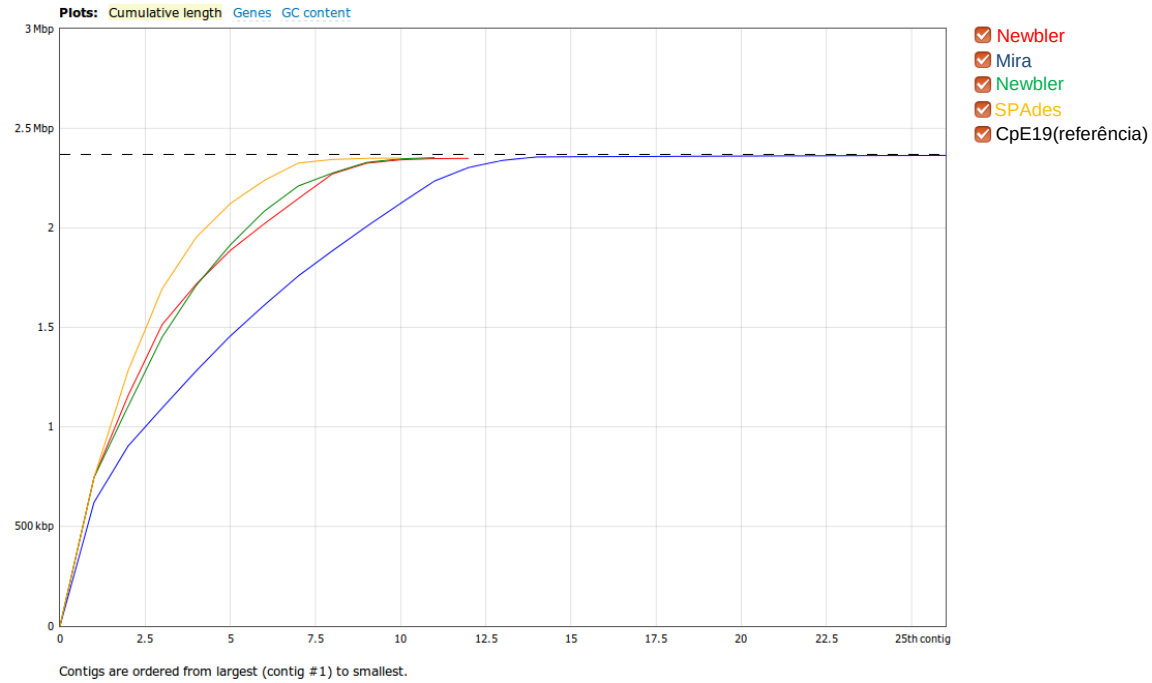


Figura adicional 2 - Gráfico de avaliação da qualidade da montagem da *C. pseudotuberculosis* MEX31 de acordo com o genoma de referência.

Figura gerada pelo QUAST.

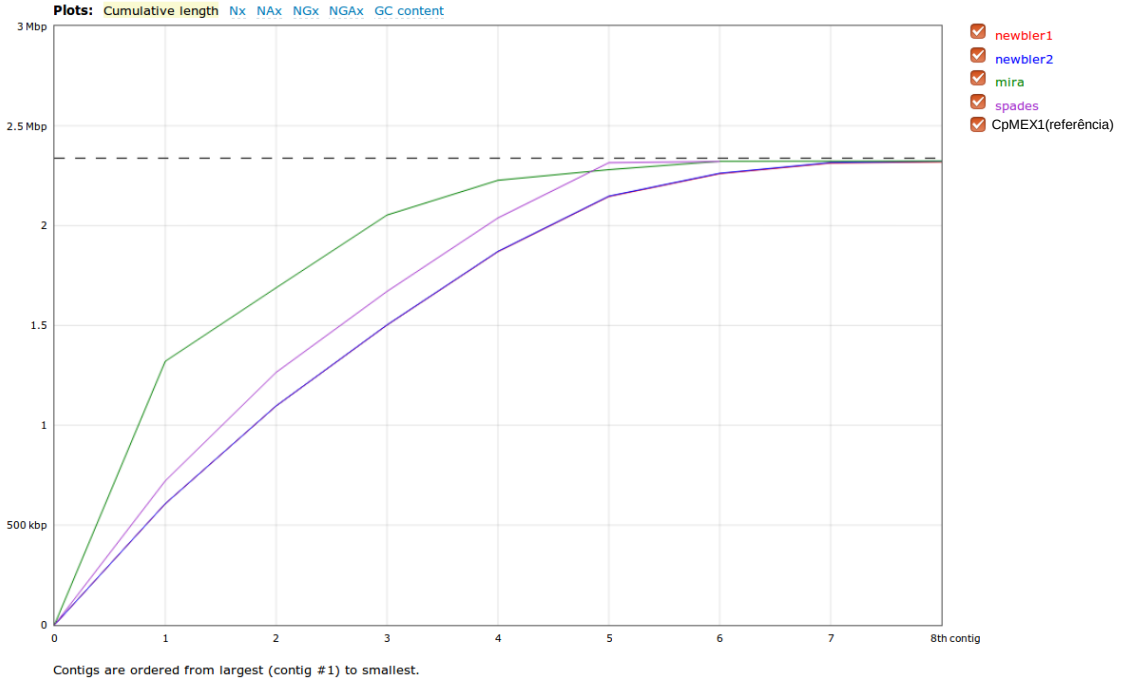


Figura adicional 3 - Gráfico de avaliação da qualidade da montagem da *C. pseudotuberculosis* MEX1 de acordo com o genoma de referência.

Figura gerada pelo QUAST.

QUAST report

01 September 2016, Thursday, 17:11:28

All statistics are based on contigs of size >= 0 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs.)

Reference size: 2 337 578 bp , G+C content: 52.18 %

Worst Median Best Show heatmap

Statistics without reference	newbler1	newbler2	mira	spades
# contigs	8	8	8	6
Largest contig	607 469	607 727	1 320 751	722 024
Total length	2 318 859	2 321 408	2 323 088	2 320 650
Total length (>= 1000 bp)	2 318 859	2 321 408	2 321 803	2 320 650
Total length (>= 10000 bp)	2 312 941	2 315 433	2 321 803	2 314 729
Total length (>= 50000 bp)	2 312 941	2 315 433	2 279 928	2 314 729
Misassemblies				
# misassemblies	0	0	1	0
Misassembled contigs length	0	0	1 320 751	0
Mismatches				
# mismatches per 100 kbp	6.99	7.5	7.06	8.06
# indels per 100 kbp	9.620	10.08	8.83	14.44
# N's per 100 kbp	0	0	0.04	0
Genome statistics				
Genome fraction (%)	99.2	99.291	99.314	99.268
Duplication ratio	1	1	1.001	1
NGA50	405 039	405 469	598 631	543 202

Figura adicional 4 - Gráfico de avaliação da qualidade da montagem da *C. pseudotuberculosis* MEX1 de acordo com o genoma de referência.

Figura gerada pelo QUAST.

ANEXO II – FIGURA DO CONTIGUATOR

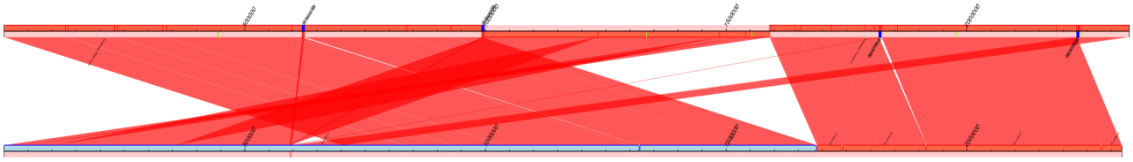


Figura adicional 5 - Gráfico de sintenia do genoma *C. pseudotuberculosis* MEX9 (fita superior) contra os *contigs* gerados pelo montador MIRA para os dados da *C. pseudotuberculosis* MEX1, confirmando a inversão e a transposição.

Figura gerada pelo CONTIGuator.

ANEXO III – TABELAS DE GENES EXCLUSIVOS DE CADA BIOVAR

Tabela adicional 1 – Famílias proteicas com função conhecida encontradas somente no biovar Ovis.

Gene	Produto
metF	5,10-methylenetetrahydrofolate reductase (EC 1.5.1.20)
-	Mobile element protein
corA	Magnesium and cobalt transport protein CorA
-	Type III restriction-modification system methylation subunit (EC 2.1.1.72)
-	L-asparaginase (EC 3.5.1.1)
-	ATP/GTP-binding protein
pstA	Phosphate transport system permease protein PstA (TC 3.A.1.7.1)
-	Putative oxidoreductase
-	Putative transmembrane protein
-	Putative fimbrial subunit
-	DNA double-strand break repair Rad50 ATPase
-	ABC transporter, ATP-binding protein
-	Phage protein
gluD	glutamate transporter permease protein GluD
fadF	Fe-S oxidoreductase
pip	Proline iminopeptidase (EC 3.4.11.5)
-	Zn-ribbon-containing, possibly RNA-binding protein and truncated derivatives
-	Selenoprotein O and cysteine-containing homologs
nanH	Sialidase (EC 3.2.1.18)
-	ABC transporter, ATP-binding protein
nanE	N-acetylmannosamine-6-phosphate 2-epimerase (EC 5.1.3.9)
-	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72)
res	Type III restriction-modification system StyLTI enzyme res (EC 3.1.21.5)
-	Putative helicase
-	Copper(I) chaperone CopZ
sigK	RNA polymerase sigma-70 factor
-	Hypothetical protein YggS, proline synthase co-transcribed bacterial homolog PROSC
-	Phage protein
-	putative RNA methyltransferase
metX	Homoserine O-acetyltransferase (EC 2.3.1.31)
nudF	ADP-ribose pyrophosphatase (EC 3.6.1.13)

-	Type III restriction enzyme, res subunit:DEAD/DEAH box helicase, N-terminal
-	FIG131328: Predicted ATP-dependent endonuclease of the OLD family

Tabela adicional 2 - Famílias proteicas com função conhecida encontradas somente no biovar Equi.

Gene	Produto
moaC	Cyclic pyranopterin monophosphate synthase accessory protein
-	Putative transposase
-	Molybdopterin synthase sulfur carrier subunit
mutT2	7,8-dihydro-8-oxoguanine-triphosphatase
molB	Molybdenum transport system permease protein ModB (TC 3.A.1.8.1) / Molybdenum transport ATP-binding protein ModC (TC 3.A.1.8.1)
-	Hypothetical protein YggS, proline synthase co-transcribed bacterial homolog PROSC
echA	Putative hydrolase
mhpC	Molybdopterin molybdenumtransferase (EC 2.10.1.1)
moaA	Cyclic pyranopterin phosphate synthase (MoaA) (EC 4.1.99.18)
narK	Nitrate/nitrite transporter NarK
narT	Nitrate/nitrite transporter NarT
-	Mobile element protein
-	Bifunctional deaminase-reductase domain protein
molA	Molybdenum ABC transporter, periplasmic molybdenum-binding protein ModA (TC 3.A.1.8.1)
cas1	CRISPR-associated protein Cas1
-	Mobile element protein
-	UPF0061 protein YdiU
-	CRISPR-associated protein, Cse4 family
corA	Magnesium and cobalt transport protein CorA
-	UPF0225 protein YchJ
-	Phytoene dehydrogenase and related proteins
moaE	Molybdopterin adenylyltransferase (EC 2.7.7.75)
-	Molybdenum cofactor guanylyltransferase (EC 2.7.7.77)
mrpf1	Na(+) H(+) antiporter subunit F
-	putative membrane protein
moeB	Molybdopterin-synthase adenylyltransferase (EC 2.7.7.80)
dnaQ2	DNA polymerase III epsilon subunit (EC 2.7.7.7)
cas3	CRISPR-associated helicase Cas3
-	Zn-ribbon-containing, possibly RNA-binding protein and truncated derivatives
hsdM	Type I restriction-modification system, DNA-methyltransferase subunit M (EC

	2.1.1.72)
	Respiratory nitrate reductase gamma chain (EC 1.7.99.4)
	Respiratory nitrate reductase alpha chain (EC 1.7.99.4)
	Respiratory nitrate reductase delta chain (EC 1.7.99.4)
	Respiratory nitrate reductase beta chain (EC 1.7.99.4)
arcD	Arginine/ornithine antiporter ArcD
	rRNA methylase

ANEXO IV – FIGURAS DO BRIG

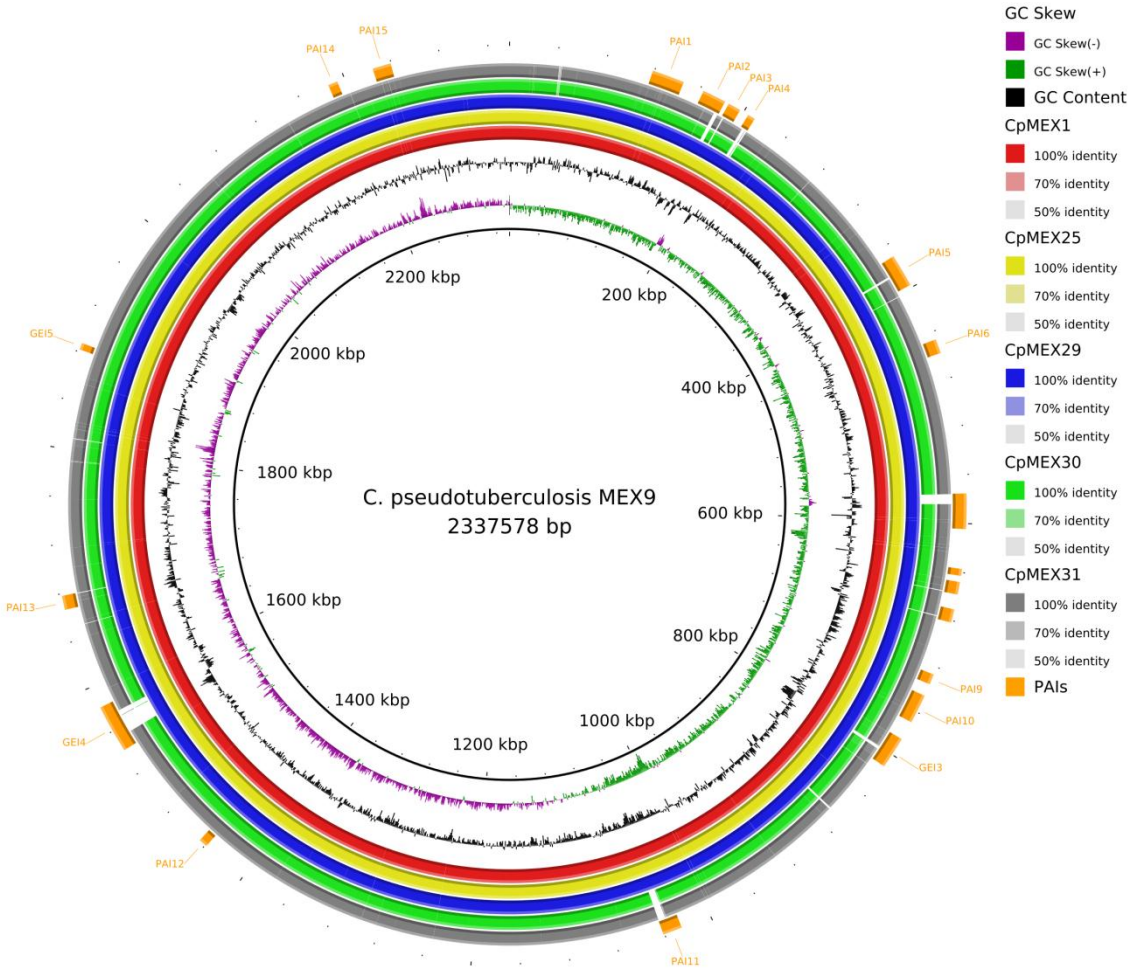


Figura adicional 6 - Visualização circular dos genomas dos organismos estudados tendo como genoma de referência a *C. pseudotuberculosis* MEX9 (ao centro) e *C. pseudotuberculosis* MEX31 (mais externo).

Em laranja as PAIs e GEIs preditas pelo GIPSy para a *C. pseudotuberculosis* MEX9.

Figura gerada pelo BRIG.

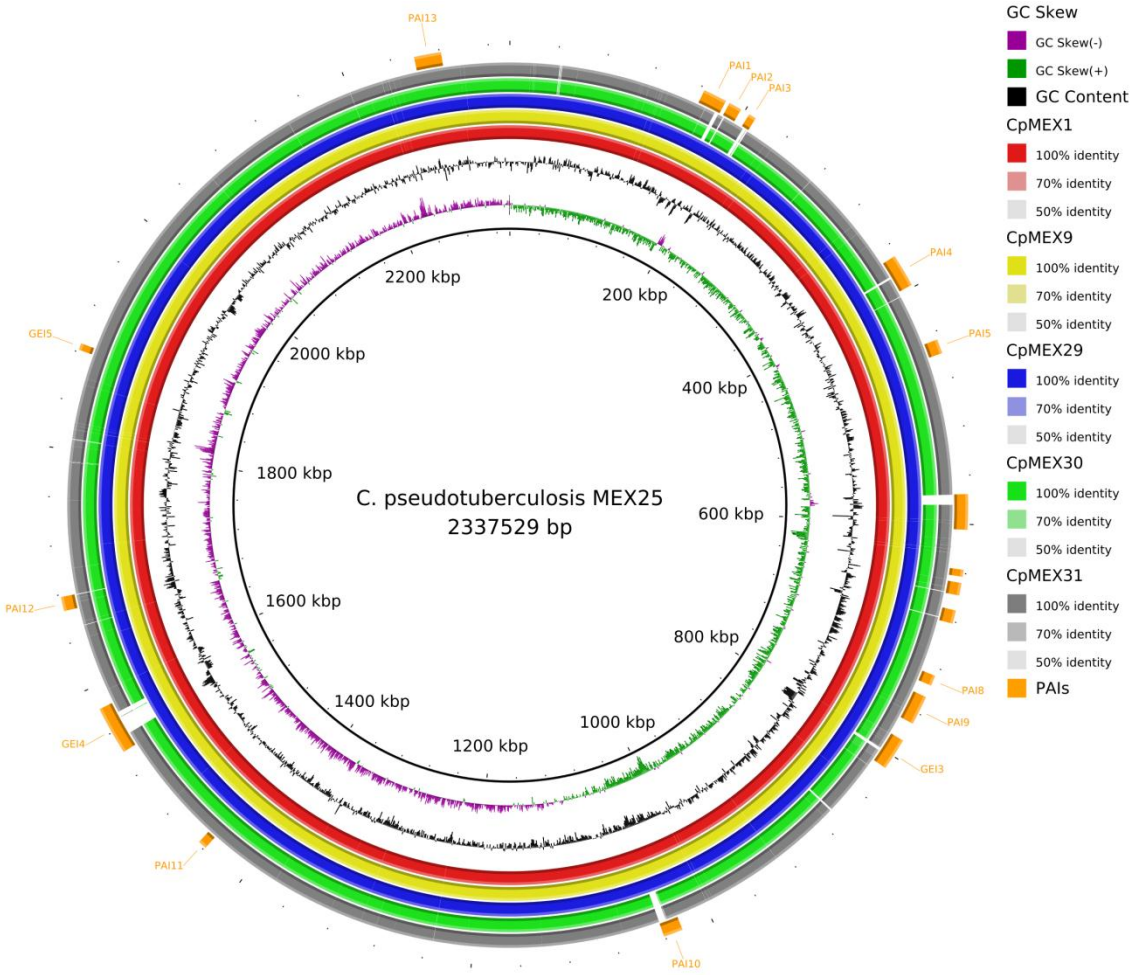


Figura adicional 7 - Visualização circular dos genomas dos organismos estudados tendo como genoma de referência a *C. pseudotuberculosis* MEX25 (ao centro) e *C. pseudotuberculosis* MEX31 (mais externo).

Em laranja as PAIs e GEIs preditas pelo GIPSy para a *C. pseudotuberculosis* MEX25.

Figura gerada pelo BRIG.

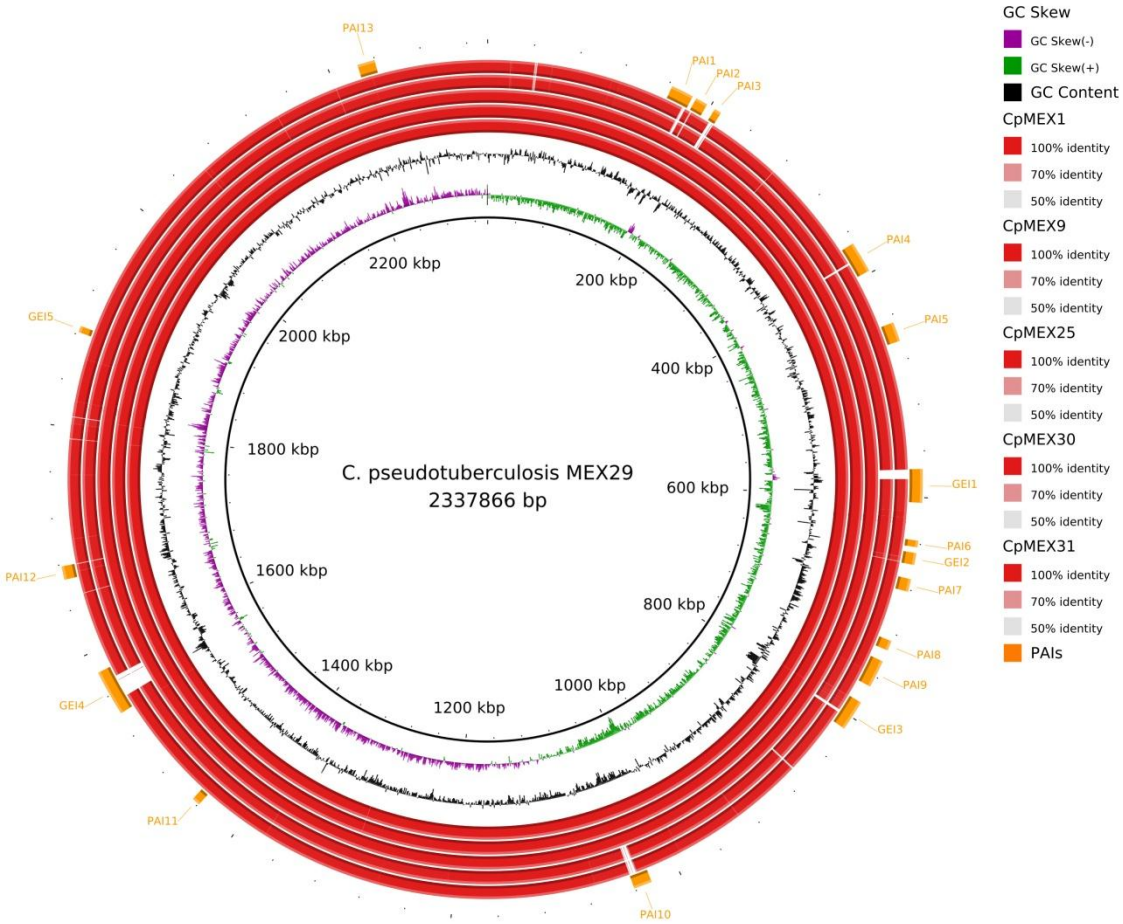


Figura adicional 8 - Visualização circular dos genomas dos organismos estudados tendo como genoma de referência a *C. pseudotuberculosis* MEX29 (ao centro) e *C. pseudotuberculosis* MEX31 (mais externo).

Em laranja as PAIs e GEIs preditas pelo GIPSy para a *C. pseudotuberculosis* MEX29.

Figura gerada pelo BRIG.

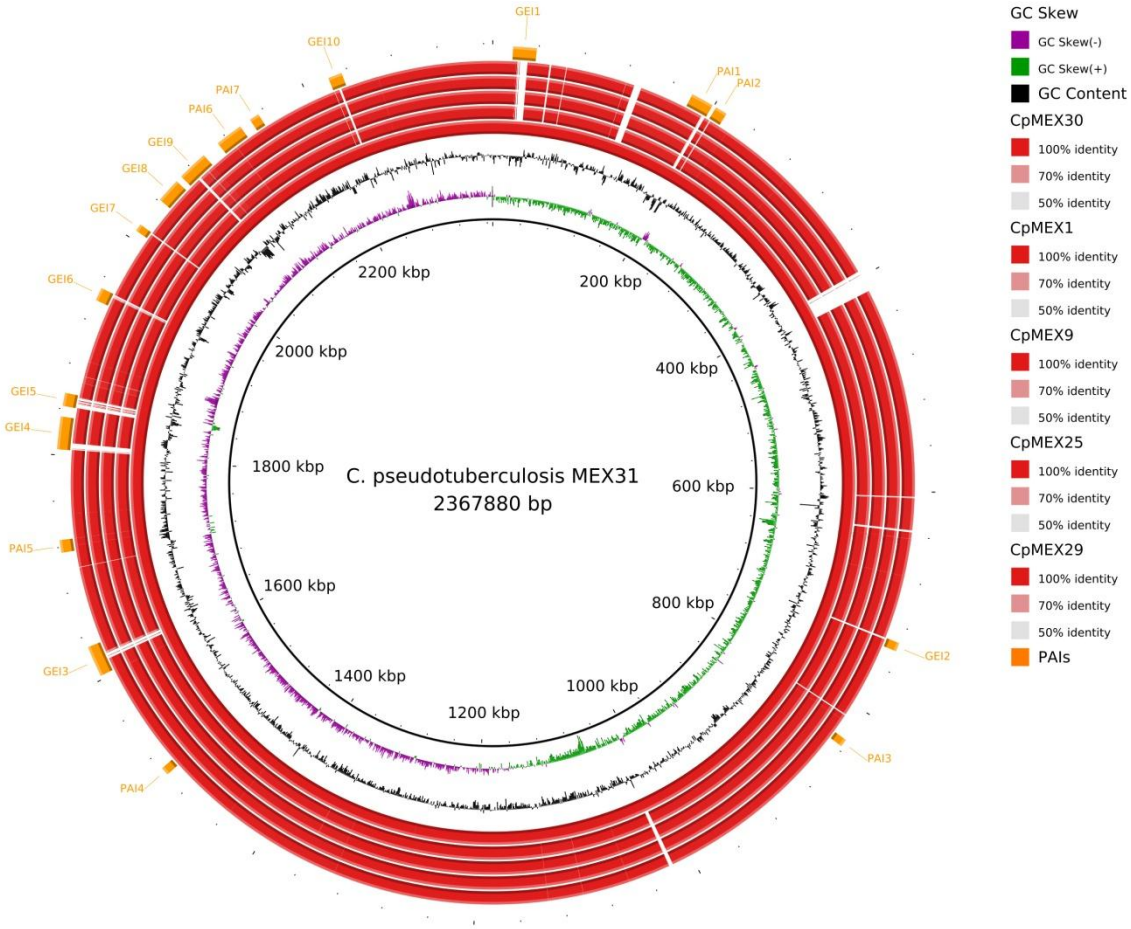


Figura em anexo 9 - Visualização circular dos genomas dos organismos estudados tendo como genoma de referência a *C. pseudotuberculosis* MEX31 (ao centro) e *C. pseudotuberculosis* MEX29 (mais externo).

Em laranja as PAIs e GEIs preditas pelo GIPSy para a *C. pseudotuberculosis* MEX31.

Figura gerada pelo BRIG.

ANEXO V - TABELAS DOS GENES PRESENTES NAS PAIS

Tabela adicional 3 – Genes encontrados nas PAIs preditas da *C. pseudotuberculosis* MEX1.

PAI	Gene / Locus tag	Produto
PAI1	cpmex1_0025	transposase for IS3509b
	Pld	Phospholipase D
	fagC	Putative iron-siderophore uptake system ATP-binding component
	fagB	ABC-type Fe3+-siderophore transport system, permease component
	fagA	ABC-type Fe3+-siderophore transport system, permease component
	fagD	Putative iron-siderophore uptake system exported solute-binding component
PAI2	mntA	ABC-type metal ion transport system, periplasmic component
	mntB	Manganese ABC transporter, ATP-binding protein SitB
	mntC	Manganese/zinc/iron transport system
	mntD	Manganese/zinc/iron transport system permease protein
	troA	Periplasmic zinc-binding protein troA
	pgpB	Phosphatidylglycerophosphatase B
PAI3	qorA	Quinone oxidoreductase
	-	tRNA-Ser-TGA
	pat	Putative phenylalanine aminotransferase
	-	tRNA-Ser-GCT
	-	tRNA-Arg-ACG
	sdcS	Sodium-dependent dicarboxylate transporter sdcS
	glcT	PtsGHI operon antiterminator
	nagE	PTS system N-acetylglucosamine-specific
	cpmex1_0133	PTS N-acetylmuramic acid transporter subunit
	cpmex1_0134	PTS N-acetylmuramic acid transporter subunit IIBC
	norB	Nitric-oxide reductase, cytochrome b-containing subunit I
	-	tRNA-Arg-ACG
	pdxR	HTH-type pyridoxine biosynthesis transcriptional
	pdxS	Pyridoxal 5'-phosphate synthase subunit PdxS
	pdxT	Pyridoxal 5'-phosphate synthase subunit PdxT
	tyrA	Prephenate dehydrogenase
	tadA	tRNA-specific adenosine deaminase

	-	tRNA-Ser-CGA
	cpmex1_0143	putative membrane protein
	Tgt	Queuine tRNA-ribosyltransferase
	gltX	Glutamyl-tRNA synthetase
	gntP	Gluconate permease
	idnK	Gluconokinase (EC 2.7.1.12)
	idnD	Zinc-binding alcohol dehydrogenase
	idnO	Gluconate 5-dehydrogenase
	-	tRNA-Ser-GGA
PAI4	deoR	Deoxyribonucleoside regulator DeoR (transcriptional repressor)
	deoD	Purine nucleoside phosphorylase (EC 2.4.2.1)
	cpmex1_0172	MFS-type drug efflux transporter
	deoC1	Deoxyribose-phosphate aldolase (EC 4.1.2.4)
	pmmB	Probable phosphomannomutase pmmB (EC 5.4.2.8)
	sdrC	Serine-aspartate repeat-containing protein
	sdrD	Serine-aspartate repeat-containing protein D
	cpmex1_0180	putative membrane protein
	lysC	Aspartate kinase
PAI5	potA1	Spermidine/putrescine import ATP-binding protein PotA
	potC	Spermidine/putrescine transport system permease PotC
	potD	Spermidine/putrescine-binding periplasmic protein
	glpT	Glycerol-3-phosphate transporter
	mprA	Response regulator
	senX1	Sensor-like histidine kinase
PAI6	cpmex1_0203	Putative secreted protein
	yhaP	Protein yhaP
	yhaQ	ABC transporter ATP-binding protein YhaQ
	cpmex1_0208	Transposase, Mutator family
	-	tRNA-Pro-CGG
PAI7	Tuf	Translation elongation factor Tu
	cpmex1_0356	Alkaline shock protein 23
PAI8	oppA1	Oligopeptide-binding protein OppA
	oppDF1	Oligopeptide transport ATP-binding protein
	oppC1	Oligopeptide transport system permease OppC
	oppB1	Oligopeptide transport system permease OppB
	rplN	50S ribosomal protein L14
	rplX	50S ribosomal protein L24
	rplE	50S ribosomal protein L5
	sdaC	Serine transporter

	sdaA	L-serine dehydratase 1
	cpmex1_0380	L-asparaginase (EC 3.5.1.1)
	cpmex1_0381	putative phosphatase
	rpsH	30S ribosomal protein S8
	rplF	50S ribosomal protein L6
	rplR	50S ribosomal protein L18
	rpsE	30S ribosomal protein S5
	rpmD	50S ribosomal protein L30
	rplO	50S ribosomal protein L15
	malE1	Maltotriose-binding protein
	malF1	Maltose transport system permease protein
	malG1	Maltose transport system permease
	traX	Protein traX
	nirC	FNT family formate-nitrite transporter
	cpmex1_0394	Chromosome segregation ATPases
	malK	Glycerol-3-phosphate-transporting ATPase
PAI9	cpmex1_0625	putative membrane protein
	mazG	MazG nucleotide pyrophosphohydrolase
	uhpT	Sugar phosphate antiporter
PAI10	yvrC	ABC transporter substrate-binding protein
	cpmex1_0664	Cobalamin/Fe3+-siderophores ABC transporter, permease
	potA2	Spermidine/putrescine import ATP-binding protein PotA
	Glf	UDP-galactopyranose mutase
	glpT1	Glycerol-3-phosphate transporter
	glpQ2	Glycerophosphoryl diester phosphodiesterase (EC 3.1.4.46)
PAI11	cpmex1_0702	Aldehyde dehydrogenas
	-	tRNA-Gly-CCC
	Dcd	Deoxycytidine triphosphate deaminase
	udgA	UDP-glucose 6-dehydrogenase
	cpmex1_0707	FIG00732228: membrane protein
	lysS1	Lysyl-tRNA synthetase
PAI12	cpmex1_0723	ABC transporter domain-containing permease component
	cpmex1_0724	Iron(III) dicitrate transport system permease fecD
	phuC	Iron(III) dicitrate transport permease-like protein yusV
	merR	putative mercury resistance operon regulator MerR
	fadF	Fe-S oxidoreductase
	pitB	Phosphate permease
	oppA	Oligopeptide-binding protein oppA
	oppB	Oligopeptide transport system permease OppB

	oppC	Oligopeptide transport system permease OppC
	oppD	Oligopeptide ABC transporter ATP-binding protein
	cpmex1_0739	putative permease
	srtA1	Fimbrial associated sortase-like protein
	cpmex1_0741	potential surface-anchored protein
	cpmex1_0742	Putative surface-anchored membrane protein
	cpmex1_0743	Putative surface-anchored membrane protein
	cpmex1_0744	Putative surface-anchored membrane protein
PAI13	cpmex1_0982	Iron(III) dicitrate transport system, periplasmic iron-binding protein FecB (TC 3.A.1.14.1)
	cpmex1_0983	Putative iron transport system membrane protein
	cpmex1_0984	Iron(III) dicitrate transport system permease protein FecD (TC 3.A.1.14.1)
	cpmex1_0985	Ferric enterobactin transport ATP-binding protein FepC (TC 3.A.1.14.2)
	cpmex1_0991	Putative surface-anchored membrane protein
	-	tRNA-Pro-GGG
PAI14	-	-
	cpmex1_2103	Putative ATP-dependent DNA helicase
	cpmex1_2104	putative membrane protein
	cpmex1_2105	Aspartate aminotransferase (EC 2.6.1.1)
	cpmex1_2106	Glutamine-dependent 2-keto-4-methylthiobutyrate transaminase

Tabela adicional 4 – Genes encontrados nas PAIs preditas da *C. pseudotuberculosis* MEX30.

PAI	Gene / Locus tag	Produto
PAI1	(Os dados estão na Tabela X, PAI4 da MEX1)	
PAI2	potA1	Spermidine/putrescine import ATP-binding protein
	potC	Spermidine/putrescine transport system permease
	potD	Spermidine/putrescine-binding periplasmic
	glpT	Glycerol-3-phosphate transporter
	mprA1	Response regulator
	senX3A	Sensor-like histidine kinase
PAI3	mutT1	MutT-like protein
	oppA	Oligopeptide-binding protein
	oppB	Oligopeptide transport system permease protein oppB
	oppC	Oligopeptide transport system permease protein oppC
	oppD	Oligopeptide transport ATP-binding protein OppD

	sprX	Trypsin-like serine protease
PAI4	cpmex30_1658	Proline iminopeptidase
	cpmex30_1659	Pyridoxal phosphate enzyme, YggS family
	cpmex30_1660	UPF0033 protein YeeD
	cpmex30_1664	Membrane protein
	cpmex30_1667	ABC transporter substrate-binding protein
	cpmex30_1668	MFS-type drug efflux transporter
	cpmex30_1669	oligopeptide ABC transporter
	cpmex30_1670	ABC transporter ATP-binding protein
PAI5	cpmex30_2003	putative secreted protein
	cpmex30_2004	putative membrane protein
	yxlG	ABC transporter permease YxlF
	yxlF	ABC transporter, ATP-binding protein
	yxlE	protein YxlE
	cpmex30_2008	FIG00732228: membrane protein
	padR	Transcriptional regulator, PadR family
PAI6	cpmex30_2057	Putative surface-anchored membrane protein
	cpmex30_2058	potential surface-anchored protein
	srtA1	Sortase A, LPXTG specific
	cpmex30_2060	putative integral membrane protein
	oppD	Oligopeptide ABC transporter ATP-binding protein
	oppC	Oligopeptide transport system permease OppC
	oppB	Oligopeptide transport system permease OppB
	oppA	Oligopeptide-binding protein oppA
	pitB	Phosphate permease
	fadF	Fe-S oxidoreductase
	merR	putative mercury resistance operon regulator MerR
	phuC	Iron(III) dicitrate transport permease-like protein yusV
	cpmex30_2072	Iron(III) dicitrate transport system permease fecD
	cpmex30_2073	ABC transporter domain-containing permease
PAI7	lysS1	Lysyl-tRNA synthetase (class II) (EC 6.1.1.6)
	cpmex30_2086	FIG00732228: membrane protein
	udgA	UDP-glucose 6-dehydrogenase (EC 1.1.1.22)
	dcd	Deoxycytidine triphosphate deaminase (EC 3.5.4.30) (dUMP-forming)
	-	tRNA-Gly-CCC
	cpmex30_2090	Succinate-semialdehyde dehydrogenase [NAD] (EC 1.2.1.24); Succinate-semialdehyde dehydrogenase [NADP+] (EC 1.2.1.79)
PAI8	cpmex30_2126	LPxTG domain-containing protein

	glpQ2	Glycerophosphoryl diester phosphodiesterase (EC 3.1.4.46)
	glpT1	Glycerol-3-phosphate transporter
	glf	UDP-galactopyranose mutase (EC 5.4.99.9)
	potA2	Spermidine/putrescine import ATP-binding protein PotA
	cpmex30_2132	ABC-type transporter, permease component
	yvrC	ABC transporter substrate-binding lipoprotein
PAI9	uhpT	Hexose phosphate transport protein UhpT
	cpmex30_2169	putative membrane protein
PAI10	-	5S rRNA ## 5S ribosomal RNA
	-	LSU rRNA ## 23S rRNA, large subunit ribosomal RNA
	-	SSU rRNA ## 16S rRNA, small subunit ribosomal RNA
	cpmex30_2187	ABC transporter, ATP-binding protein
	gntR	Transcriptional regulator, GntR family
	cpmex30_2189	putative protein (2G313) / putative protein (2G313)
	pspA1	Phage shock protein A
	trxA2	Thioredoxin
	ctpA2	Cation transport protein
	cpmex30_2195	Lead, cadmium, zinc and mercury transporting ATPase (EC 3.6.3.3) (EC 3.6.3.5); Copper-translocating P-type ATPase (EC 3.6.3.4)
	cpmex30_2196	MFS family major facilitator transporter
	dnaB	Replicative DNA helicase (DnaB) (EC 3.6.4.12)
	cpmex30_2198	Glycoside hydrolase 15-related protein
PAI11	cpmex30_2214	DedA protein
	dps	DNA protection during starvation protein
	cpmex30_2216	Endonuclease VIII
	cpmex30_2217	NAD(P)H-hydrate epimerase (EC 5.1.99.6) / ADP-dependent (S)-NAD(P)H-hydrate dehydratase (EC 4.2.1.136)
	cpmex30_2219	Conserved integral membrane protein
	leuS	Leucyl-tRNA synthetase (EC 6.1.1.4)
	cpmex30_2222	antimicrobial peptide ABC transporter ATPase
	cpmex30_2223	ABC transporter, permease protein
	chrS	two-component system histidine kinase ChrS
	hrrA	Hemoglobin-dependent two component system response regulator HrrA