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

Marcus Vinicius Canário Viana

Patogenômica de *Corynebacterium silvaticum*

Belo Horizonte

2022

Marcus Vinicius Canário Viana

Patogenômica de *Corynebacterium silvaticum*

Versão final

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

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo
Coorientador: Prof. Dr. Sandeep Tiwari

Belo Horizonte
2022

043 Viana, Marcus Vinicius Canário.
Patogenômica de Corynebacterium silvaticum [manuscrito] / Marcus Vinicius Canário Viana. – 2022.
133 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo. Coorientador: Prof. Dr. Sandeep Tiwari.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Corynebacterium. 3. Noxas. 4. Genômica. I. Azevedo, Vasco Ariston de Carvalho. II. Tiwari, Sandeep. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

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

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

"Patogenômica de *Corynebacterium silvaticum*"

Marcus Vinicius Canário Viana

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Vasco Ariston de Carvalho Azevedo - Orientador
UFMG

Prof. Sandeep Tiwari - Coorientador
UFBA

Prof. José Miguel Ortega
UFMG

Prof. Henrique Cesar Pereira Figueiredo
UFMG

Prof. Rogerio Margis
UFRGS

Prof. Luis Gustavo Carvalho Pacheco
UFBA

Prof. Siomar de Castro Soares
UFTM

Prof. Rodrigo Dias de Oliveira Carvalho
UFBA

Belo Horizonte, 19 de dezembro de 2022.

Documento assinado eletronicamente por **Rogerio Margis**, **Usuário Externo**, em 20/12/2022, às 09:40, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rodrigo Dias de Oliveira Carvalho**, **Usuário Externo**, em 20/12/2022, às 10:18, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Siomar de Castro Soares**, **Usuário Externo**, em 20/12/2022, às 11:30, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Henrique Cesar Pereira Figueiredo**, **Professor do Magistério Superior**, em 20/12/2022, às 15:31, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jose Miguel Ortega**, **Servidor(a)**, em 20/12/2022, às 16:38, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luis Gustavo Carvalho Pacheco**, **Usuário Externo**, em 20/12/2022, às 16:56, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Vasco Ariston de Carvalho Azevedo**, **Professor do Magistério Superior**, em 20/12/2022, às 17:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Sandeep Tiwari**, **Usuário Externo**, em 20/12/2022, às 17:08, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1944897** e o código CRC **369379E0**.

Referência: Processo nº 23072.271407/2022-95 SEI nº 1944897

AGRADECIMENTOS

Agradeço a todos que contribuíram para a minha segunda tese de doutorado, especialmente:

Ao Prof. Dr. Vasco Ariston de Carvalho Azevedo pela orientação e confiança desde 2014.

Às secretárias Fernanda Magalhães Madureira, Marcia Natalia Dutra Ramalho e Silva, e Sheila Cristina Magalhães Madureira Santana pelo suporte com toda a burocracia da Pós-Graduação.

À Profª. Drª. Alice Rebecca Wattam pela colaboração com a pesquisa sobre *Corynebacteirum* desde 2015, quando fui para a Virginia Tech, Estados Unidos da América, para o doutorado sanduíche pelo Programa de Pós-Graduação em Genética da UFMG.

Aos programas de pós-graduação dos quais fui aluno, pela formação e apoio: Programa de Pós-Graduação em Genética, Biodiversidade e Conservação (Universidade Estadual do Sudoeste da Bahia, UESB), Programa de Pós-Graduação em Genética (Universidade Federal de Minas Gerais, UFMG), e Programa Interunidades de Pós-Graduação de Bioinformática (UFMG).

A todos os meus professores, em especial Paulo Roberto Antunes de Mello Affonso que, por mencionar a Bioinformática em uma de suas aulas da graduação em 2008, me influenciou a ter um título de doutorado nesta área.

RESUMO

Corynebacterium silvaticum é uma bactéria gram-positiva patogênica recentemente descrita, pertencente ao grupo da *C. diphtheriae*, um grupo de espécies que possui o potencial de produzir a toxina diftérica (DT). A espécie foi descrita em 2020 como não produtora de toxina diftérica e capaz de infectar porcos selvagens e veados, causando doença similar à linfadenite caseosa causada por *C. pseudotuberculosis* biovar ovis em caprinos e ovinos. As primeiras amostras foram isoladas na Alemanha e Áustria. Em 2014 a espécie foi isolada pela primeira vez de porcos domésticos e em Portugal, classificada erroneamente como *C. pseudotuberculosis*. Pouco se conhece sobre os seus mecanismos de patogenicidade além do fator de virulência fosfolipase D, duas espécies de hospedeiros (porco e veado) ou distribuição geográfica (Alemanha, Áustria, Suíça e Portugal). Este trabalho teve o objetivo de utilizar análises genômicas para investigar a diversidade, os mecanismos de patogenicidade, desenvolver marcadores moleculares para diagnóstico por PCR e proteínas sob seleção positiva. Em comparação com a espécie mais próxima *C. ulcerans*, *C. silvaticum* é mais homogênea, com 75.2% dos genes conservados, e possui oito ilhas genômicas. 18 de 36 fatores de virulência descritos em *Corynebacterium* estão presentes no genoma central. Os genomas formam dois clados, sendo um deles representado pelas linhagens de Portugal e a 05-13 da Áustria que provavelmente produzem DT devido à ausência de um *frameshift* do gene *tox*. Genes marcadores foram encontrados para os dois clados e para amostras de Portugal. Nenhum gene foi identificado como sob seleção positiva. Em resumo, *C. silvaticum* é uma espécie clonal e predita como tendo o potencial de causar zoonose e difteria em humanos, mas a produção de DT deve ser verificada experimentalmente nas linhagens de Portugal e 05-13.

Palavras-chave: *Corynebacterium*, patógeno, genômica.

ABSTRACT

Corynebacterium silvaticum is a recently described pathogenic gram-positive bacterium belonging to the *C. diphtheriae* group, a group of species have the potential to produce diphtheria toxin (DT). The species was described in 2020 as not producing diphtheria toxin and capable of infecting wild pigs and deer, causing disease similar to caseous lymphadenitis caused by *C. pseudotuberculosis* biovar ovis in goats and sheep. The first samples were isolated in Germany and Austria. In 2014, the species was isolated for the first time from domestic pigs and, in Portugal, it was misclassified as *C. pseudotuberculosis*. Little is known about its pathogenicity mechanisms beyond the phospholipase D virulence factor, two host species (pig and deer) or geographical distribution (Germany, Austria, Switzerland and Portugal). This work aimed to use genomic analyzes to investigate diversity, mechanisms of pathogenicity, develop molecular markers for diagnosis by PCR and proteins under positive selection. Compared to the closest species *C. ulcerans*, *C. silvaticum* is more homogeneous, with 75.2% of genes conserved, and has eight genomic islands. 18 of 36 virulence factors described in Corynebacterium are present in the core genome. The genomes form two clades, one of which is represented by lineages from Portugal and 05-13 from Austria that probably produce DT due to the absence of a frameshift of the tox gene. Marker genes were found for both clades and for samples from Portugal. No genes were identified as under positive selection. In summary, *C. silvaticum* is a clonal species and predicted to have the potential to cause zoonosis and diphtheria in humans, but DT production must be experimentally verified in the strains from Portugal and 05-13.

Keywords: *Corynebacterium*, pathogen, genomics.

LISTA DE FIGURAS

Figura 1 - Abscessos caseosos causados por infecção por *Corynebacterium silvaticum* em porco selvagem (Fonte: Einsenberg *et al.* 2014)......18

LISTA DE ABREVIATURAS E SIGLAS

AA	*Amino Acid*
ACT	*Artemis Comparison Tool*
ANI	*Average Nucleotide Identity*
ATCC	*American Type Culture Collection*
BLAST	*Basic Local Alignment Search Tool*
Cb	*Corynebacterium belfantii*
CC	*Clonal Complex /* Complexo Clonal
Cd	*Corynebacterium diphtheriae*
CDD	Conserved Domain Database
CDS	*Coding Sequence*
CLA	*Caseous Lymphadenitis*
COG	*Cluster of Orthologous Groups*
Cp	*Corynebacterium pseudotuberculosis*
Cpequi	*Corynebacterium pseudotuberculosis* biovar equi
Cpovis	*Corynebacterium pseudotuberculosis* biovar ovis
Cr	*Corynebacterium rouxii*
CRISPR	*Clustered Regularly Interspaced Short Palindromic Repeats*
Cul	*Corynebacterium ulcerans*
Cul1	*Corynebacterium ulcerans* lineage 1
Cul2	*Corynebacterium ulcerans* lineage 2
Cs	*Corynebacterium silvaticum*
CWSS	*Cell Wall Sorting Signal*
GO	*Gene Ontology*
DT	*Diphtheria Toxin*
dDDH	*digital DNA:DNA hybridization*
EC	*Enzyme Commission*
ERIC-PCR	*Enterobacterial Repetitive Intergenic Consensus Polymerase Chain Reaction*
FDR	*False Discovery Rate*
FT-IR	*Fourier-transform Infrared Spectroscopy*
GBDP	*Genome Blast Distance Phylogeny*
GGDC	*Genome-to-Genome Distance Calculator*

GI	*Genomic Island*
GWAS	*Genome-Wide Association Studies*
HGT	*Horizontal Gene Transfer*
IS	*Insertion Sequence*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
LCA	Linfadeninte caseosa
LRT	*Likelihood Ratio Test*
MALDI-TOF	*Matrix-Assisted Laser Desorption Ionization-Time of Flight*
MGE	*Mobile Genetic Element*
MHC	Major Histocompatibility Complex
MLSA	*Multilocus Sequence Analysis*
MLST	*Multilocus Sequence Typing*
MS	*Mass Spectrometry*
NCTC	*National Collection of Type Cultures*
NTTB	*Non-Toxigenic but Tox gene-Bearing*
NSS	*Neighbor Similarity Score*
OLC	*Overlap–Layout–Consensus*
OTU	*Operational Taxonomic Units*
PATRIC	*Pathosystems Resource Integration Center*
PHI	*Pairwise Homoplasy Index*
PBIT	*Pipeline Builder for Identification of Targets*
PCR	*Polymerase Chain Reaction*
PLD	*Phospholipase D*
PS	*Positive Selection / Positively Selected*
RASTtk	*Rapid Annotation using Subsystems Technology tool kit*
RT-qPCR	Reverse Transcription quantitative Real-Time *Polymerase Chain Reaction*
SNP	*Single Nucleotide Polymorphism*
ST	*Sequence Type*
SE	Surface Exposed
SSU	*Small Subunit*
tRNA	*Transfer Ribonucleic Acid*
TYGS	*Type (Strain) Genome Server*
VFDB	*Virulence Factors Database*

wgMLST	*Whole Genomic Multilocus Sequence Typing*
WBC	Wild Board Cluster
WGS	*Whole Genome Sequencing*

SUMÁRIO

1 INTRODUÇÃO 13

1.1 A pesquisa sobre *Corynebacterium silvaticum* 13

1.2 Colaboradores 13

1.3 Estrutura da tese 14

2 REVISÃO DE LITERATURA 15

2.1 Biologia de *Corynebacterium silvaticum* 15

2.1.1 Taxonomia e microbiologia 15

2.1.2 Patogenicidade 16

2.1.3 Epidemiologia 18

2.2 Genômica e patogenômica 19

2.2.1 Sequenciamento, montagem e anotação de genomas 19

2.2.2 Identificação de espécies e tipagem 22

2.2.3 Filogenética e filogenômica 23

2.2.4 Pangenômica 23

2.2.5 Análise de seleção positiva 24

2.2.6 Patogenômica 24

3. JUSTIFICATIVA E OBJETIVOS 27

3.1 Justificativa 27

3.2 Objetivo geral 27

3.3 Objetivos específicos 27

4. ARTIGOS 28

4.1 Artigo 1 – Taxonomic classification of strain PO100/5 shows a broader geographic distribution and genetic markers of the recently described *Corynebacterium silvaticum* 28

4.2 Artigo 2 – Analysis of *Corynebacterium silvaticum* genomes from Portugal reveals a single cluster and a clade suggested to produce diphtheria toxin 47

4.3 Artigo 3 – Evidence of episodic positive selection in *Corynebacterium diphtheriae* complex of species and its implementations in identification of drug and vaccine targets .. 75
5. DISCUSSÃO GERAL 96
6. CONCLUSÃO E PERSPECTIVAS 98
REFERÊNCIAS 99
APÊNDICE 105
APÊNDICE A – Linhas de comando 105
APÊNDICE B – Artigos em primeira ou última autoria 108
APÊNDICE C – Artigos em colaboração 111
APÊNDICE D – Trabalhos apresentados em eventos 123
APÊNDICE E – Participação em eventos 125
APÊNDICE F – Organização de eventos 126
APÊNDICE H – Disciplinas ministradas 127
APÊNDICE I – Cursos e estágios ministrados 128
APÊNDICE J – Coorientações 129
APÊNDICE K – Bancas 130
APÊNDICE L – Prêmios 131
APÊNDICE M – Pós-Doutorados 132

1 INTRODUÇÃO

1.1 A pesquisa sobre *Corynebacterium silvaticum*

O Laboratório de Genética Celular e Molecular (LGCM) da Universidade Federal de Minas Gerais e colaboradores pesquisam os mecanismos de patogenicidade, diagnósticos e vacinas de *Corynebacterium* por genômica estrutural e funcional. O primeiro genoma completo sequenciado pelo grupo foi o da C. *pseudotuberculosis* biovar Ovis linhagem 1002, liberada em 2010. Em 2014, o grupo de pesquisa da Drª. Manuela Oliveira, da Universidade Federal de Lisboa, Portugal, publicou o primeiro isolamento de *C. pseudotuberculosis* de porcos. Em 2017, oito linhagens foram enviadas para o LGCM para análise genômica e foi verificado que estas representavam uma nova espécie, filogeneticamente próxima a *C. ulcerans*. A comparação com genomas públicos revelou que as linhagens pertenciam ao grupo até então denominado "Wild Boar Cluster" (WBC) de *C. ulcerans*, que continha linhagens isoladas de porcos selvagens e um veado na Alemanha. Em 2020, o WBC foi reclassificado com a nova espécie *C. silvaticum*. Após a descrição da espécie, o objetivo da presente tese foi alterado de uma sugestão de nova espécie para a análise de plasticidade genômica da espécie, com foco nas características únicas das linhagens de Portugal.

1.2 Colaboradores

Dra. Alice Rebecca Wattam, pesquisadora da University of Virginia, EUA;

Dr. Artur Silva, pesquisador e professor da UFPA, Brasil;

Bruna Ferreira Sampaio Ribeiro, graduada em Medicina Veterinária da UFMG, Brasil;

Dr. Debmalya Barh, professor visitante da UFMG, Brasil;

Dra. Francielly Morais-Rodrigues, doutora em Bioinformática pela UFMG, Brasil;

Dr. Henrique Figueiredo, pesquisador e professor da UFMG, Brasil;

Msc. Janaína Canário Cerqueira, mestre em Bioinformática pela UFMG, Brasil;

Dr. José Henrique Souza Galdino Brandão, doutor em Genética e Biologia Molecular pela UFRGS, Brasil;

Dra. Manuela Oliveira, pesquisadora e professora da Universidade de Lisboa, Portugal;

Dr. Luís Tavares, pesquisador e professor da Universidade de Lisboa, Portugal;

Dra. Manuela Oliveira, pesquisadora e professora da Universidade de Lisboa, Portugal;

Dra. Marcelle Oliveira Almeida, doutora em Genética pela UFMG, Brasil;

Dr. Paulo Luiz Souza Carneiro, pesquisador e professor da UESB, Brasil;

Dr. Preetam Ghosh, pesquisador da Virginia Commonwealth University, EUA;

Msc. Raquel Hurtado, doutoranda do programa de Bioinformática da UFMG, Brasil;

Msc. Rodrigo Profeta, doutorando do programa de Bioinformática da UFMG, Brasil;

Dr. Rommel Ramos, pesquisador e professor da UFPA, Brasil;

Dr. Siomar de Castro Soares, pesquisador e professor da UFTM, Brasil.

1.3 Estrutura da tese

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

A Revisão Bibliográfica aborda a biologia de *C. silvaticum*, e conceitos e análises das áreas de genômica e patogenômica;

O Artigo 1 apresenta a primeira análise genômica de um isolado de *C. silvaticum* de porco doméstico e de Portugal e o pangenoma da espécie;

O Artigo 2 apresenta uma análise genômica com foco em oito linhagens de *C. silvaticum* de porco doméstico e de Portugal;

O Artigo 3 apresenta uma análise de seleção positiva nas seis espécies do complexo da *Corynebacterium diphtheriae*, no qual *C. silvaticum* está incluída;

A Discussão Geral integra resultados dos artigos apresentados;

O Apêndice apresenta as linhas de comando utilizadas e demais atividades e produção científica desenvolvidas durante a tese.

2 REVISÃO DE LITERATURA

2.1 Biologia de *Corynebacterium silvaticum*

2.1.1 Taxonomia e microbiologia

O gênero *Corynebacterium* pertence à família Corynebacteriaceae, ordem Corynebacteriales, classe Actinobacteria, filo Actinobacteria. O gênero possui espécies de bactérias gram-positivas, claviformes, imóveis, catalase positivas, anaeróbias facultativas ou aeróbias. A parede celular possui arabinose e galactose como açúcares principais e pode conter ácido micólico. A espécie tipo do gênero é a *Corynebacterium diphtheriae*. O gênero possui espécies de vida livre, comensais e patogênicas, com espécies de interesse biotecnológico, médico e veterinário (BERNARD; FUNKE, 2015).

Entre espécies patogênicas, o grupo da *C. diphtheriae* é um clado de espécies com o potencial de produzir a toxina diftérica (DT) quando adquirem o gene *tox* de um profago (BERNARD; FUNKE, 2015). Atualmente, seis espécies pertencem a este grupo. *Corynebacterium diphtheriae* infecta humanos e outros animais, causando difteria em humanos quando produtora da DT (BERNARD; FUNKE, 2015). *C. belfantii* é uma reclassificação de *C. diphtheriae* biovar Belfanti (DAZAS et al., 2018) e sinônimo de *C. diphtheriae* subespécie *lausannense* (BADELL et al., 2020). *C. rouxii* é uma reclassificação de linhagens de *C. diphtheriae* biovar Belfanti (BADELL et al., 2020). *C. pseudotuberculosis* infecta diversos mamíferos e possui dois biovares. O biovar ovis possui tropismo por caprinos e ovinos e causa linfadenite caseosa (LCA), podendo infectar humanos, enquanto equi possui tropismo por cavalos, búfalos e outros mamíferos e causa diferentes doenças (DORELLA et al., 2006). *C. ulcerans* possui um amplo espectro de hospedeiros mamíferos, podendo causar diversas doenças, e atualmente é o agente etiológico mais frequente da difteria (HACKER et al., 2016). *C. silvaticum* é uma reclassificação de linhagens atípicas de *C. ulcerans* (DANGEL et al., 2020). *C. pseudotuberculosis*, *C. ulcerans* e *C. silvaticum* são produtoras da fosfolipase D (PLD) (BERNARD; FUNKE, 2015; DANGEL et al., 2020), um dos fatores de virulência do gênero (TAUCH; BURKOVSKI, 2015).

C. silvaticum é a espécie descrita mais recentemente do grupo da *C. diphtheriae* (DANGEL et al., 2020). As primeiras linhagens foram isoladas em 1997 na Alemanha, oriundas de lesões similares à LCA em porcos selvagens (*Sus scrofa*) e em um veado (*Capreolus capreolus*) (CONTZEN et al., 2011; EISENBERG et al., 2014). As linhagens foram descritas como produtoras de PLD e portadoras do gene *tox* mas não produtoras de DT (*non toxigenic*

but tox gene-bearing, NTTB). A identificação taxonômica por testes bioquímicos (API Coryne e VITEK2-compact) resultou em *C. pseudotuberculosis* ou um perfil inválido (CONTZEN et al., 2011; EISENBERG et al., 2014), enquanto a similaridade com *C. ulcerans* foi sugerida com o uso de Espectrometria de Massa (MS) *Matrix-Assisted Laser Desorption Ionization-Time of Flight* (MALDI-TOF), espectroscopia no Infravermelho por Transformada de Fourier (FT-IR) e análises de DNA, como filogenia do gene *rpoB*, (CONTZEN et al., 2011; DANGEL et al., 2019; EISENBERG et al., 2014). As linhagens passaram a ser classificadas como "grupo do porco selvagem" (*wild boar cluster*) de *C. ulcerans*, devido a formação de um clado irmão por análises filogenéticas de diferentes sequências dos genes *tox*, *dtxR* e elementos móveis (DANGEL et al., 2019), além de um perfil próprio de MALFDI-TOFF MS discriminado após o refinamento do método (RAU et al., 2019). Em 2020, o grupo foi descrito como uma nova espécie (DANGEL et al., 2020).

Segundo a descrição, *C. silvaticum* foi isolada de LCA de animais selvagens. As células são gram-positivas, não esporulantes, de formas irregular à bastonete claviforme, com 1,2–2,5 μm de comprimento e 0,4–0,6 μm de largura, crescem em condições aeróbias ou microaeróbias, com crescimento ótimo a 37 °C e pH de 7 a 8. São NTTB, produtoras de PLD e suscetíveis a clindamicina, penicilina e eritromicina. Ao contrário de *C. ulcerans*, possui lipídios não identificados em seu perfil, é negativa para fermentação de glicogênio e susceptível a clindamicina. Possui sequência do gene 16S e perfil de ácidos graxos compatíveis com grupo da *C. diphtheriae*. A filogenia do gene 16S sugere um ancestral comum mais recente com *C. pseudotuberculosis*, enquanto o gene *rpoB* sugere um ancestral comum mais recente com *C. ulcerans*. A filogenia dos genes da DT (*tox*) e do seu regulador (*dtxR*) são mais próximas de *C. diphtheriae*. Pode ser distinguida por FT-IR e MALDI-TOF MS, Identidade Média de Nucleotídeo (ANI), Tipagem de Sequência Multilocus (*Multi Locus Sequence Typing*, MLST) por possuir a sequência tipo (ST) exclusiva ST578, genoma aumentado em relação a *C. ulcerans* (2.55 Mbp) e conteúdo G+C de 54.4%. A linhagem tipo é a KL0182^{T} (DANGEL et al., 2020).

2.1.2 Patogenicidade

Patogenicidade é a capacidade de causar doença. Para colonizar um hospedeiro, patógenos dependem de proteínas e outros componentes necessários para adesão, sobrevivência e crescimento, aquisição de nutrientes, competição com outros microrganismos e evasão do sistema imune. Estes componentes são classificados como fatores de nicho ou de virulência.

Fatores de nicho são necessários para colonizar um nicho ecológico. Exemplos são proteínas de adesão, mobilidade e aquisição de nutrientes, como sideróforos. Estes fatores podem ser encontrados em bactérias da microbiota normal de um indivíduo. Por outro lado, fatores de virulência permitem ao microrganismo ter acesso e sobreviver em locais normalmente não colonizados do hospedeiro e causar danos. Exemplos de fatores de virulência são internalinas, invasinas e toxinas (HILL, 2012; TAUCH; BURKOVSKI, 2015). Em *C. silvaticum*, diversos fatores de nicho e virulência foram identificados por similaridade de sequências de proteínas (DANGEL et al., 2020; MÖLLER et al., 2020), porém apenas o fator de virulência PLD foi verificado experimentalmente como funcional (DANGEL et al., 2020). Esta toxina degrada a esfingomielina da membrana celular do hospedeiro, contribuindo para a disseminação do patógeno e morte de macrófagos, e é o principal fator de virulência de *C. pseudotuberculosis* (D'AFONSECA et al., 2008; MCKEAN; DAVIES; MOORE, 2007).

A manifestação da infecção por *C. silvaticum* descrita até o momento é a LCA (Figura 1) (BUSCH et al., 2019; CONTZEN et al., 2011; EISENBERG et al., 2014; OLIVEIRA et al., 2014). Na LCA, fatores de virulência como a PLD ou de nicho como ácidos micólicos da parede celular causam dermonecrose com inflamação e aumento da permeabilidade vascular. No interior de fagócitos, o patógeno é levado aos nódulos linfáticos e causa a destruição destes. As lesões caseosas, ou piogranulomas, aumentam devido a ciclos repetidos de necrose e encapsulamento, o que resulta em camadas concêntricas ou "em forma de anéis de cebola" (WINDSOR, 2011).

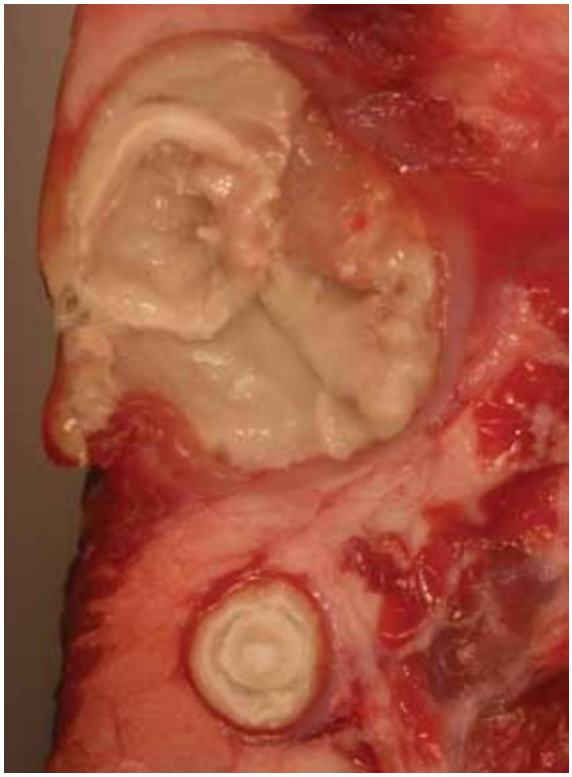


Figura 1 - Abscessos caseosos causados por infecção por *Corynebacterium silvaticum* em porco selvagem (Fonte: Einsenberg *et al.* 2014).

2.1.3 Epidemiologia

A informação sobre os hospedeiros de *C. silvaticum* é limitada, pois a espécie foi isolada apenas de porcos selvagens e um veado na Alemanha (CONTZEN et al., 2011; DANGEL et al., 2019; EISENBERG et al., 2014), porco selvagem na Áustria (BUSCH et al., 2019), e porco doméstico em Portugal (MÖLLER et al., 2020; OLIVEIRA et al., 2014).

A linhagem de porco de Portugal PO100/5 foi identificada como *C. pseudotuberculosis* por teste bioquímico (Api Coryne® *kit*), reação em cadeia da polimerase (*Polymerase Chain Reaction*, PCR) multiplex e eletroforese em campo pulsante (OLIVEIRA et al., 2014). Após sequenciamento genômico e filogenia do gene *rpoB* pelo grupo de pesquisa do professor Vasco Azevedo da UFMG, o genoma foi depositado em 2017 no GenBank como *C. ulcerans* (CP021417.1). Uma análise de *rpoB* na publicação que descreveu *C. silvaticum* (DANGEL et al., 2020) e uma análise genômica de linhagens de *C. ulcerans* (MÖLLER et al., 2020) concluíam que esta linhagem seria da espécie *C. silvaticum*.

Porcos selvagens são reservatórios de patógenos que podem ser transmitidos para animais domésticos e humanos por contato com suas secreções ou consumo de carne (MENG;

LINDSAY; SRIRANGANATHAN, 2009). *C. ulcerans*, por exemplo, pode ser transmitida a partir de porcos para humanos (BERGER et al., 2013; KONIG et al., 2014; SCHUHEGGER et al., 2009). Como a infecção por *C. silvaticum* em porco doméstico já foi documentada fora da Alemanha (OLIVEIRA et al., 2014), a distribuição geográfica pode envolver outras regiões da Europa. Além disso, existe a possibilidade de transmissão zoonótica, caso a espécie seja capaz de infectar humanos.

É importante que a identificação de um patógeno seja realizada de maneira rápida, efetiva e com melhor custo-benefício para se guiar o tratamento correto (BAUER et al., 2014). Neste contexto, ainda não existem marcadores PCR multiplex para *C. silvaticum*. O PCR multiplex é mais rápido e mais acurado que testes baseados em fenótipos como bioquímico e morfológico (BAUER et al., 2014; CHRISTENSEN; OLSEN, 2018a) e, diferente de outros métodos moleculares, não necessita de sequenciamento de DNA, espectrometria ou espectroscopia, como análises de filogenia, MLST, MALDI-TOF MS e FT-IR. Marcadores para PCR multiplex foram desenvolvidos para a identificação de *C. diphtheriae*, *C. ulcerans, C. pseudotuberculosis* e *C. belfantii* (ALMEIDA et al., 2017; BADELL et al., 2019; PACHECO et al., 2007). Recentemente foi desenvolvido um método de identificação baseado em *Enterobacterial Repetitive Intergenic Consensus* PCR (ERIC-PCR) para a diferenciar *C. silvaticum*, *C. pseudotuberculosis* e *C. auriscanis* (RAMOS et al., 2022).

Como *C. silvaticum* foi recentemente descoberta e descrita, pouco se conhece sobre sua biologia. O sequenciamento e análise do seu material genético pode fornecer informações que serão aplicadas para o seu controle, como mecanismos de patogenicidade, marcadores moleculares, vacinas e desenvolvimento de antibióticos.

2.2 Genômica e patogenômica

2.2.1 Sequenciamento, montagem e anotação de genomas

O genoma é o material genético de um indivíduo e a genômica é o estudo da sua estrutura e função. A genômica é uma ciência moderna que surgiu pela convergência de outras ciências, principalmente a Bioquímica, a Estatística e a Ciência da Computação. A Era da Genômica iniciou em 1977, com o sequenciamento do genoma de 5.375 mil pares de bases do Bacteriófago ϕx174. O sequenciamento é a determinação de nucleotídeos em uma sequência de DNA ou RNA, visando conhecer a informação genética contida nessas estruturas (FIETTO; LAMËGO, 2015). O sequenciamento de genomas bacterianos contribui para o estudo da função

de genes, metabolismo, interação com hospedeiro e ambiente, diversidade, evolução, epidemiologia e desenvolvimento de métodos de controle (LOMAN; PALLEN, 2015).

O sequenciamento genômico passou por três revoluções: sequenciamento *shotgun*, de alto rendimento ou massivamente paralelo, e de molécula única (LOMAN; PALLEN, 2015). Devido a limitações das tecnologias de sequenciamento atuais, o DNA genômico (ou RNA do transcriptoma) deve ser fragmentado para gerar bibliotecas de fragmentos que serão usadas para reconstruir a sequência da molécula original, em um processo chamado de montagem. Três diferentes bibliotecas são usadas no sequenciamento massivamente paralelo. Em bibliotecas de fragmento, adaptadores são ligados aos fragmentos de DNA para gerar leituras simples. Em leituras pareadas, o sequenciamento gera pares de leitura de distância conhecida, facilitando o processo posterior de montagem. Em bibliotecas *paired-end*, um par de leituras é gerado a partir das extremidades de um mesmo fragmento. Em bibliotecas *mate-pair*, as leituras são geradas a partir das extremidades, permitindo o uso de fragmentos maiores e distâncias maiores entre as leituras. Isso é obtido unindo suas extremidades, circularizando o fragmento, clivando a região que contém estas pontas unidas e usando-a para gerar a biblioteca pareada. A quantidade de vezes em que um nucleotídeo do genoma é representado nas leituras é chamada cobertura vertical de sequenciamento (NAGARAJAN; POP, 2013).

Como abordado anteriormente, a montagem é a reconstrução da molécula original de DNA ou RNA. Na montagem por referência, as leituras de sequenciamento são mapeadas em um genoma referência e a sequência consenso é extraída. Na montagem *de novo* ou *ab initio*, apenas as sobreposições entre as leituras são usadas, reconstruindo sequências maiores, ou *contigs*. Diferentes algoritmos podem ser usados. O *Overlap–Layout–Consensus* (OLC) identifica pares de leituras que se sobrepõem o suficiente para gerar um grafo que contém um nó para cada leitura e uma aresta para leituras que se sobrepõem. O algoritmo Grafo de *De Bruijn* extrai subsequências consecutivas de tamanho *k* de cada leitura (*k-mers*) para gerar um grafo no qual os nós são os *k-mers* e as arestas são *k-mers* sobrepostos por exatamente *k* - 1 caracteres. Um caminho específico no grafo representa uma montagem. Uma medida da qualidade da montagem é o N50. Para se obter esta medida, os *contigs* são ordenados do maior para o menor, e os seus tamanhos são somados progressivamente. O N50 é tamanho do *contig* cuja soma com o tamanho dos anteriores atinge mais de 50% da montagem total. Quanto maior este valor, melhor a montagem. Os *contigs* são ordenados e orientados para gerar um rascunho do genoma (NAGARAJAN; POP, 2013). As lacunas, ou *gaps*, entre os *contigs* podem ser

fechadas usando *contigs* de montagens feitas por diferentes algoritmos ou por montagem por referência (SOUSA et al., 2019).

A anotação é a descrição da função dos elementos de um genoma. A anotação de sequências codificadoras de proteínas (*Coding Sequence*, CDS) e RNA não codificantes pode ser realizada por estratégia extrínseca ou intrínseca. A extrínseca usa evidência e conhecimento *a priori* de banco de dados e atribui função por similaridade de sequência, sendo apropriada para genomas filogeneticamente relacionados. A intrínseca usa propriedades de DNA codificante e não codificante, como a composição de códons, bases e aminoácidos para criar modelos de probabilidade para identificar genes. *Pipelines* de anotação podem usar uma combinação das duas estratégias (BECKLOFF et al., 2012). A anotação funcional pode ser realizada por busca por de sequências similares em bancos de dados como GenBank para proteínas e ácidos nucleicos (BENSON, 2004), Uniprot para proteínas (WASMUTH et al., 2017) e *Kyoto Encyclopedia of Genes and Genomes* (KEGG) para vias metabólicas (OGATA et al., 1999).

A busca por sequências similares é normalmente realizada por alinhamento. O alinhamento de sequências é a determinação da correspondência entre caracteres de duas (alinhamento par a par) ou mais sequências (alinhamento múltiplo). O alinhamento global procura o melhor alinhamento entre duas sequências inteiras, enquanto o alinhamento local busca o melhor alinhamento entre segmentos das sequências (LESK, 2008). A identidade é a proporção de caracteres idênticos no alinhamento, enquanto similaridade é a proporção de caracteres similares como aminoácidos de mesma carga. Uma matriz de substituição baseada em um modelo biológico é usada para avaliar o melhor alinhamento possível, dando pontuação maior para caracteres idênticos e menor para similares, diferentes ou lacunas. O valor *p* indica a probabilidade de o alinhamento ocorrer ao caso, ou seja, a probabilidade de se obter uma pontuação igual ou maior quando os caracteres de uma das duas sequências são permutados aleatoriamente. O valor *e* indica esta probabilidade para alinhamento com as sequências de um banco de dados (JUNQUEIRA; BRAUN; VERLI, 2014). O algoritmo de alinhamento mais comum é o *Basic Local Alignment Search Tool* (BLAST) (CAMACHO et al., 2009).

Os estudos que comparam sequências de diferentes genomas requerem a identificação de homólogos. Dentre estes, ortólogos se originam de especiação, parálogos de duplicação no mesmo genoma, xenólogos de transferência horizontal (*Horizontal Gene Transfer*, HGT), e onólogos por duplicação de genomas inteiros. A relação de ancestralidade pode ser inferida por

grafos, árvores ou combinações de métodos (meta-métodos) (ALTENHOFF; GLOVER; DESSIMOZ, 2019).

2.2.2 Identificação de espécies e tipagem

A identificação de espécies de bactérias por métodos tradicionais como testes bioquímicos possui acurácia limitada e demanda mais tempo do que os métodos moleculares. A comparação de sequências do gene do rRNA 16S é comumente usada pois o 16S é parte da estrutura do ribossomo e está presente em todas as espécies. Uma proporção de caracteres idênticos, ou identidade, acima de 97 % é esperada para sequência da mesma espécie e acima de 95% para o mesmo gênero. Outros genes podem ser usados quando o 16S de duas espécies não possui mutações suficientes para discriminá-las (CHRISTENSEN; OLSEN, 2018a). A identidade média de nucleotídeo (ANI) alinha fragmentos de DNA de duas amostras e considera uma identidade maior que 95% como amostras da mesma espécie. A identidade abaixo deste valor tornaria a recombinação menos frequente, isolando geneticamente as populações de bactérias. A análise de mais de 90 mil genomas mostrou que 99.8% das comparações conformam com identidade > 95% para a mesma espécie e < 83 entre espécies (JAIN et al., 2018). A hibridização de DNA-DNA é uma técnica onde o DNA de duas amostras é misturado e desnaturado. A renaturação de mais de 70% do DNA é esperada quando são da mesma espécie. A técnica *in silico* hibridização de DNA-DNA digital usando *Genome Blast Distance Phylogeny* (GBDP) foi desenvolvida para substituí-la (MEIER-KOLTHOFF et al., 2013). Exemplos de serviços *online* para classificação taxonômica são o *Type (Strain) Genome Server* (TYGS) (MEIER-KOLTHOFF; GÖKER, 2019) e o *Genome-to-Genome Distance Calculator* (AUCH; KLENK; GÖKER, 2010). Outros métodos de identificação que não usam DNA são MALDI-TOF MS (CLARK et al., 2013) e FT-IR (NOVAIS et al., 2019).

Métodos de tipagem foram desenvolvidos para estudar populações bacterianas. A MLST compara a sequência de genes conservados em linhagens da mesma espécie. Um número é designado para cada alelo e cada perfil de alelo é uma ST. Em um esquema de sete locos, um grupo de STs que compartilham pelo menos cinco alelos forma um complexo clonal (CC). A análise filogenética dos dados do MLST que mostra a evolução dos STs é denominada Análise de Sequência Multilocus (*Multilocus Sequence Analysis*, MLSA). A Tipagem de Sequência Multilocus de Genoma Inteiro (wgMLST) inclui todos os genes conservados estimados ou um conjunto de genes conservados selecionados, para evitar um número muito alto de STs. O

Polimorfismo de Nucleotídeo Único (*Single Nucleotide Polymorphism*, SNP) são alterações de nucleotídeo único nas sequências. Na tipagem por SNP, as leituras de sequenciamento de um genoma são mapeadas para um genoma de referência para identificar posições de SNPs. Os SNPs de diferentes genomas podem ser usados para gerar uma filogenia. Quando o *background* genético de uma espécie é bem caracterizado, um sistema de tipagem pode ser estabelecido para características como sorotipo, resistência a antibióticos ou virulência (CHRISTENSEN; OLSEN, 2018b).

2.2.3 Filogenética e filogenômica

Filogenética descreve relações taxonômicas entre espécies, representando-as como uma árvore de espécies, onde os nós representam ancestrais comuns e as folhas são as amostras analisadas, ou unidades taxonômicas operacionais (*Operational Taxonomic Units*, OTUs). A filogenética possui aplicações na reconstrução da história de sequências e populações, e identificação e classificação de genes, sequências reguladoras e sua homologia. Na filogenia molecular, as sequências são alinhadas e a filogenia é estimada usando um método baseado em distância (matriz de distância), como junção de vizinhos (*Neihgbor-Joining*), ou um método baseado em caracteres, como Máxima Verossimilhança, e os relacionamentos são representados como uma árvore de gene. Os valores de suporte dos nós são calculados por um teste de permutação como o *bootstrap*, que reamostra posições do alinhamento de sequência para gerar pseudo-amostras e calcular a proporção de árvores que possuem um nó específico (CHRISTENSEN; OLSEN, 2018c; YANG; RANNALA, 2012). A filogenômica utiliza sequências genômicas, permitindo separar o sinal filogenético e o ruído, como por exemplo, diferenças entre a árvore de um gene e a árvore da espécie (CHAN; RAGAN, 2013).

2.2.4 Pangenômica

A análise pangenômica descreve o conteúdo do genoma de um grupo de linhagens. O genoma *core* contém os genes que estão presentes em todas as linhagens e controla aspectos básicos da biologia de uma espécie e suas principais características fenotípicas. O genoma dispensável, ou acessório, contém genes que são únicos para uma linhagem (*singletons*) e genes que estão presentes em duas ou mais, mas não em todas as linhagens. Estes genes podem codificar vias bioquímicas e funções suplementares que conferem vantagens seletivas. O

pangenoma é a soma do *core* e do genoma dispensável. Um pangenoma é considerado aberto quando se prevê que cresça com a inclusão de mais genomas na análise. Isto pode ser verificado com a fórmula $n = \kappa * N^{\gamma}$, na qual n é o número de genes, N é o número de genomas, e κ e γ ($\alpha = \gamma - 1$) são parâmetros livres determinados empiricamente. Um valor $\alpha > 1$ indica um pangenoma fechado, onde o sequenciamento de mais genomas não contribuirá significativamente com a identificação de novos genes. Um valor $\alpha < 1$ indica a situação oposta. Como os genes são continuamente trocados dentro e entre espécies bacterianas por HGT, em teoria o pangenoma nunca será totalmente descrito (MEDINI et al., 2005; TETTELIN et al., 2008).

2.2.5 Análise de seleção positiva

Seleção positiva é o aumento de frequência ou fixação de mutações vantajosas com maior probabilidade do que o esperado ao acaso. A maioria dos genes conhecidos sob esse processo estão envolvidos na imunidade e defesa (ANISIMOVA; LIBERLES, 2012). A seleção positiva em sequências codificantes de proteínas pode ser estimada pela proporção ω de substituições não sinônimas (*dN*) e sinônimas (*dS*) ($\omega = dN / dS$). Um valor de $\omega > 1$ sugere seleção positiva, que favorece uma alteração de aminoácidos. Valores de $\omega = 1$ e $\omega < 1$, representam evolução neutra e seleção negativa, que conserva o aminoácido, respectivamente. Testes de razão de verossimilhança (*Likelihood Ratio Tests*, LRTs) baseados em modelos de substituição de códons foram desenvolvidos para estimar a seleção positiva em diferentes níveis. Testes com modelos *site* estimam seleção diversificadora contínua em posições específicas de aminoácidos de todas as linhagens. Os testes com modelos *branch-site* identificam a seleção episódica, exclusiva para posições específicas de aminoácidos em um grupo filogenético predefinido (ANISIMOVA, 2012; ANISIMOVA; LIBERLES, 2012).

2.2.6 Patogenômica

O estudo do genoma de patógenos, ou patogenômica, permite identificar variações e pressões evolutivas que podem estar associadas a adaptação a um estilo de vida. Durante o processo evolutivo, o genoma sofre rearranjos, recombinações, processos de aumento e/ou ganho de função como duplicação, divergência funcional e HGT que podem conferir múltiplas adaptações; além de processos de redução e/ou perda de função como deleção e

pseudogenização. Este potencial de mudança é chamado plasticidade genômica (PALLEN; WREN, 2007; SHEPPARD; GUTTMAN; FITZGERALD, 2018; TOFT; ANDERSSON, 2010). Por exemplo, durante a mudança de estilo de vida livre para parasita o genoma tende a diminuir. A dependência de um hospedeiro causa gargalo populacional, redução do tamanho efetivo, menos oportunidades para a ganho de sequências por HGT, e oferece um ambiente rico em metabólitos dentro do hospedeiro. A consequência é a diminuição da eficiência da seleção, permitindo o acúmulo de mutações deletérias e perda gradual de genes (TOFT; ANDERSSON, 2010). Nem toda variação associada a um hospedeiro é adaptativa e existem estratégias que são usadas para identificar tal variação, diferenciando-as de variações neutras associadas à deriva genética e efeito fundador (SHEPPARD; GUTTMAN; FITZGERALD, 2018). A análise de seleção positiva foi abordada anteriormente (ANISIMOVA; LIBERLES, 2012). A identificação de mutações que ocorrem de maneira independente em linhagens filogenéticas divergentes (homoplasias) indica adaptação, ou evolução convergente. Uma adaptação pode ser inferida pela presença de elementos genéticos móveis como profagos e ilhas genômicas contendo genes com funções relacionadas à colonização, como evasão do sistema imune do hospedeiro e virulência. Os genes necessários para a colonização podem ser transferidos pela microbiota do hospedeiro. A análise de sequência de genes do genoma *core* pode identificar alelos com mutações adaptativas, como um SNP que leva a mudança de aminoácido. A perda de função em vias metabólicas comuns pode identificar a transição do patógeno para intracelular. Análises pangenômicas podem identificar diferenças no genoma acessório de isolados de nichos do hospedeiro diferentes. Os estudos de associação genômica ampla (*Genome-Wide Association Studies*, GWAS) podem associar variantes à fenótipos, como colonização de um hospedeiro, sem a necessidade de uma hipótese a *priori* (SHEPPARD; GUTTMAN; FITZGERALD, 2018). A ecologia do hospedeiro pode influenciar na adaptação do patógeno. A expansão das populações humanas, da agropecuária permite que bactérias associadas a animais selvagens estejam expostas às redes de transmissão mundial, levando a emergência de novas linhagens virulentas e resistentes a antibióticos. A dieta do hospedeiro pode selecionar genes relacionados ao metabolismo para a colonização do trato digestivo (SHEPPARD; GUTTMAN; FITZGERALD, 2018).

Exemplos de aplicações da patogenômica são a identificação de marcadores moleculares para diagnóstico (ALMEIDA et al., 2017; BADELL et al., 2019), de proteínas imunogênicas para o desenvolvimento de vacinas (vacinologia reversa) (RAPPUOLI, 2000; RAPPUOLI et al., 2016) e de proteínas alvo para quimioterápicos (BARH et al., 2011; KUMAR

JAISWAL et al., 2017). Estes dados reduzem o custo e o tempo necessários para o desenvolvimento de métodos de identificação, tratamento e controle.

3. JUSTIFICATIVA E OBJETIVOS

3.1 Justificativa

C. silvaticum é uma espécie de bactéria patogênica recentemente descrita e pouco se conhece sobre seus hospedeiros potenciais e mecanismos de patogenicidade. Não existe um marcador de PCR multiplex para diagnóstico rápido e ainda não existe uma vacina. A análise genômica deste organismo pode contribuir com o entendimento de tais mecanismos e desenvolvimento de métodos de controle, diminuindo o custo e o tempo necessários para validação experimental.

3.2 Objetivo geral

O objetivo deste trabalho é a análise genômica de *C. silvaticum* para a identificação de genes envolvidos em adaptações e candidatos para o desenvolvimento de marcadores moleculares e vacinas.

3.3 Objetivos específicos

Os objetivos específicos foram:

Montar e anotar os genomas de *C. silvaticum* e *C. ulcerans* disponíveis publicamente como leituras de sequenciamento genômico;

Identificar características únicas dos genomas de *C. silvaticum* e das linhagens de Portugal;

Identificar adaptação por presença de genes específicos ou seleção positiva de sequencias codificantes de proteína;

Desenvolver marcadores moleculares para diagnóstico molecular rápido;

Identificar proteínas antigênicas candidatas para o desenvolvimento de vacinas.

4. ARTIGOS

4.1 Artigo 1 – Taxonomic classification of strain PO100/5 shows a broader geographic distribution and genetic markers of the recently described *Corynebacterium silvaticum*

Marcus Vinicius Canário Viana, Rodrigo Profeta, Alessandra Lima da Silva, Raquel Hurtado, Janaína Canário Cerqueira, Bruna Ferreira Sampaio Ribeiro, Marcelle Oliveira Almeida, Francielly Morais-Rodrigues, Manuela Oliveira, Luís Tavares, Henrique Figueiredo, Alice Rebecca Wattam, Debmalya Barh, Preetam Ghosh, Artur Silva, Vasco Azevedo. **Publicado na revista Plos One.** https://doi.org/10.1371/journal.pone.0244210.

Neste capítulo investiguei a taxonomia da linhagem PO100/5 isolada de porco em Portugal. Durante o processo de revisão por pares, outra publicação descreveu a espécie desta linhagem como *C. silvaticum*. O objetivo do manuscrito foi modificado para a análise do pangenoma para a identificação de genes relacionados a adaptações da espécie e marcadores moleculares.

PLOS ONE

RESEARCH ARTICLE

Taxonomic classification of strain PO100/5 shows a broader geographic distribution and genetic markers of the recently described *Corynebacterium silvaticum*

Marcus Vinicius Canário Viana[1,2], Rodrigo Profeta[1], Alessandra Lima da Silva[1], Raquel Hurtado[1], Janaína Canário Cerqueira[1], Bruna Ferreira Sampaio Ribeiro[1], Marcelle Oliveira Almeida[1], Francielly Morais-Rodrigues[1], Siomar de Castro Soares[3], Manuela Oliveira[4], Luís Tavares[4], Henrique Figueiredo[5], Alice Rebecca Wattam[6], Debmalya Barh[7], Preetam Ghosh[8], Artur Silva[2], Vasco Azevedo[1]*

1 Department of Genetics, Ecology and Evolution, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **2** Department of Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil, **3** Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil, **4** Centre for Interdisciplinary Research in Animal Health, Faculty of Veterinary Medicine, University of Lisbon, Lisboa, Portugal, **5** National Reference Laboratory of Aquatic Animal Disease, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **6** Biocomplexity Institute, University of Virginia, Charlottesville, Virginia, United States of America, **7** Institute of Integrative Omics and Applied Biotechnology, Purba Medinipur, West Bengal, India, **8** Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia, United States of America

* vasco@icb.ufmg.br

Citation: Viana MVC, Profeta R, da Silva AL, Hurtado R, Cerqueira JC, Ribeiro BFS, et al. (2020) Taxonomic classification of strain PO100/5 shows a broader geographic distribution and genetic markers of the recently described *Corynebacterium silvaticum*. PLoS ONE 15(12): e0244210. https://doi.org/10.1371/journal.pone.0244210

Editor: Chih-Horng Kuo, Academia Sinica, TAIWAN

Received: April 17, 2020

Accepted: December 4, 2020

Published: December 21, 2020

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico); CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior); Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG); Universidade Federal de Minas Gerais (UFMG); and Pró-Reitoria de Pesquisa da UFMG (PRPQ-UFMG); A. R. Wattam

Abstract

The bacterial strain PO100/5 was isolated from a skin abscess taken from a pig (*Sus scrofa domesticus*) in the Alentejo region of southern Portugal. It was identified as *Corynebacterium pseudotuberculosis* using biochemical tests, multiplex PCR and Pulsed Field Gel Electrophoresis. After genome sequencing and *rpoB* phylogeny, the strain was classified as *C. ulcerans*. To better understand the taxonomy of this strain and improve identification methods, we compared strain PO100/5 to other publicly available genomes from *C. diphtheriae* group. Taxonomic analysis reclassified it and three others strains as the recently described *C. silvaticum*, which have been isolated from wild boar and roe deer in Germany and Austria. The results showed that PO100/5 is the first sequenced genome of a *C. silvaticum* strain from livestock and a different geographical region, has the unique sequence type ST709, and could be could produce the *diphtheriae* toxin, along with strain 05–13. Genomic analysis of PO100/5 showed four prophages, and eight conserved genomic islands in comparison to *C. ulcerans*. Pangenome analysis of 38 *C. silvaticum* and 76 *C. ulcerans* genomes suggested that C. *silvaticum* is a genetically homogeneous species, with 73.6% of its genes conserved and a pangenome near to be closed ($\alpha > 0.952$). There are 172 genes that are unique to *C. silvaticum* in comparison to *C. ulcerans*. Most of these conserved genes are related to nutrient uptake and metabolism, prophages or immunity against them, and could be genetic markers for species identification. Strains PO100/5 (livestock) and KL0182^T (wild

was supported by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201400027C. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

boar) were predicted to be potential human pathogens. This information may be useful for identification and surveillance of this pathogen.

Introduction

The genus *Corynebacterium* from the phylum Actinobacteria has Gram-positive bacteria of biotechnological, veterinary and medical relevance with free, commensal and pathogenic lifestyles. Within pathogenic species, the most prominent species are the nearly exclusively human pathogen *C. diphtheriae* and the zoonotic *C. pseudotuberculosis* and *C. ulcerans*. These three compose the *C. diphtheriae* group, a clade of species that can be lysogenized by phages harboring the diphtheria toxin (DT) gene (*tox*) [1]. Within this group, three new species were recently described. *C. rouxii* [2] and *C. belfantii* are reclassifications of some of the *C. diphtheriae* biovar Belfanti strains [3]. *C. belfantii* is also a synonym of *C. diphtheriae* subspecies *lausannense* [2]. *C. silvaticum* [4] is a reclassification of atypical *C. ulcerans* strains. Strains of *C. silvaticum* were previously described as atypical non-toxigenic but *tox*-gene-bearing (NTTB) strains of *C. ulcerans*, isolated from wild boars and roe deer in Germany and Austria, which caused caseous lymphadenitis similar to *C. pseudotuberculosis* infections [5–8]. This variant, examined using genomics and proteomics, was initially named as the "wild boar cluster" (WBC) of *C. ulcerans* [5–7] and later reclassified as *C. silvaticum* [4].

The strain PO100/5 was isolated from caseous lymphadenitis lesions in a Black Alentejano pig (*Sus scrofa domesticus*) from a swine farm in the Alentejo region of Portugal. It was identified as *Corynebacterium pseudotuberculosis* by both biochemical tests (Api Coryne® kit) and by multiplex PCR and Pulsed Field Gel Electrophoresis [9]. Genome sequencing and *rpoB* phylogeny showed that this strain was closer to *C. ulcerans* and the genome was deposited in GenBank as a strain within this species (accession number CP021417.1). Recently the description of *C. silvaticum* was published and PO100/5 was suggested to be a strain of this species by *rpoB* phylogeny [4], while a genomic analysis of 28 *C. ulcerans* strains suggested that PO100/5, W25 and KL1196 could represent a new species [10]. KL1196 had already been classified as *C. silvaticum* [4].

Pigs and boars are reservoirs of *C. silvaticum* [4–7] and *C. ulcerans* [11–13], and are known to transmit pathogens to humans and other domestic animals [11, 12, 14]. Rapid, simple and reliable identification of this new species is essential for diagnosis, treatment and surveillance [15, 16]. To better understand the taxonomy of PO100/5, we performed a comparative analysis of 34 *C. silvaticum* and 80 *C. ulcerans* genomes, as well as other publicly available genomes from the *C. diphtheriae* group in order to explore the genomic diversity of *C. silvaticum* and to identify molecular markers of this species. We have reclassified PO100/5, established the other three strains recently deposited as *C. silvaticum*, and found both a unique sequence type and genes that can be useful for species classification.

Materials and methods

Genomes, assembly, and annotation

For the taxonomic, phylogenetic and genome plasticity analyses, a total of 120 genomes (S1 File) were selected, including 80 *C. ulcerans*, 34 *C. silvaticum* strains and six type strains from the *C. diphtheriae* group. Assembled genomes were retrieved from the Pathosystems Resource Integration Center (PATRIC) [17], while genomes available as sequencing reads were

assembled in PATRIC using the SPAdes [18] strategy. All genomes were annotated using the Rapid Annotation using Subsystems Technology (RASTtk) pipeline [19] that is available in PATRIC.

Taxonomic analysis

Average Nucleotide Identity (ANI) was estimated using FastANI v1.3 [20]. An automatic genome-based taxonomic analysis was performed using the Type (Strain) Genome Server (TYGS) (https://tygs.dsmz.de) [21]. This pipeline first identifies the closest type strains using MASH [22] for genomic sequences and BLAST [23] for 16S rRNA sequences. It then identifies the 10 closest type strains using Genome Blast Distance Phylogeny (GBDP) [24]. Finally, it clusters species and subspecies using digital DNA:DNA hybridization (dDDH) with a formula that is independent of genome length, being robust against the use of incomplete genomes (formula *d4*) [24]. It uses a threshold of 70 and 79%, respectively [25]. The difference in G+C content is also evaluated and expected to vary no more than 1% within a species [26]. Those analysis were performed for *C. ulcerans* strains from GenBank, using either the assembled genomes or sequencing data to check for misidentification of *C. silvaticum* strains.

Phylogenetic trees of the *rpoB* and *tox* genes were built using the Maximum Likelihood method [27] implemented in MEGA v10.1.6 [28]. The *tox* tree included all sequences from the genomes of *C. ulcerans* and included outgroups from *C. silvaticum* and *C. pseudotuberculosis*. *C. rouxii* was not included due to all sequenced genomes being *tox*- [2]. All trees were visualized using iTOL [29].

Multi Locus Sequence Typing (MLST) was performed using MLSTcheck [30], using the scheme for *C. diphtheriae* and *C. ulcerans* (genes *atpA*, *dnaE*, *dnaK*, *fusA*, *leuA*, *odhA* and *rpoB*) [13]. The Minimum Spanning Tree (MST) generated using goeBURST Full MSLT algorithm was built using PHYLOViZ v2.0 [31].

Genome plasticity analysis

Prophages of PO100/5 were predicted using PHASTER [32]. Genomic islands were predicted using GIPSy v1.1.2 [33], with *C. ulcerans* NCTC7910^{T} and *C. pseudotuberculosis* ATCC19410^{T} used as references. A circular map was generated using BRIG v0.95 [34]. The presence of niche and virulence factors of *Corynebacterium* [35, 36] was verified using PATRIC's Protein Family Sorter and Proteome Comparison tools. Gene neighborhoods were compared with other strains using the Artemis Comparison Tool 17.0.1 [37]. Signal peptide and conserved protein domains were verified using InterProScan [38], while cell wall sorting signal (CWSS) was verified using CW-PRED [39]. Mapping of sequencing reads to sequences of interest was performed using CLC Genomics Workbench v7 [40]. Specific nucleotide sequences in other genomes were identified using BLASTn [41] and GenBank non-redundant (nr) database [42].

The identification of groups of homologous genes (orthogroups) was performed using OrthoFinder v2.3.12 [43]. The output files Orthogroups.tsv and Orthogroups_UnassignedGenes.tsv from OrthoFinder were used as input for pangenome analyses using in-house scripts. The pangenome was represented by all orthogroups and the core genome by orthogroups conserved across all (100%) genomes. The accessory (or dispensable) genome was represented by the genes not conserved across all genomes. Within this subset, singletons were exclusive to a single genome, and shared genome (or dispensable genome minus singletons) are shared between two or more, but not all genomes [44]. To develop molecular markers, we identified subsets of orthogroups conserved and exclusive to a group of genomes (exclusive core). The development of a pangenome was calculated according to Heaps' law fit formula $n = \kappa^{*}N^{\gamma}$, in which n is the number of genes, N is the number of genomes, and κ and γ

($\alpha = \gamma - 1$) are free parameters determined empirically. Heap's law establishes the pangenome as being closed when $\alpha > 1$ ($\gamma < 0$), which means that there is no significant increase with the addition of new genomes. It also defines a pangenome as open when $\alpha \leq 1$ ($0 < \gamma < 1$). The development of core genome and singletons were calculated using least-squares fit of the exponential regression decay $n = \kappa^{*}\exp[-N/\tau] + tg(\theta)$, in which *n* is the number of genes, *N* is the number of genomes, and κ, τ, and tg(θ) are free parameters determined empirically [44]. A functional annotation of genes was performed using the eggNOG-mapper v2 [45].

The pathogenicity of *C. silvaticum* to humans was predicted using PathogenFinder v. 1.1 [46]. The prediction is performed using CD-HIT-2D [47] against a database of protein families associated with human pathogens. Strains PO100/5 and KL0182^{T} were used to represent livestock (domestic pig) and wild boars isolates, respectively.

Results

Taxonomic analysis

ANI results showed that *C. ulcerans* strains PO100/5, 04–13, 05–13 and W25 had identity values ≥ 99.74% with *C. silvaticum* KL0182^{T}, and ≤ 91.03% with *C. ulcerans* NCTC7910^{T} (S2 File). The taxonomic classification using TYGS classified those strains as *C. silvaticum* due to genome and 16S rRNA GBDP trees, dDDh > 70% and G+C content difference > 1% with *C. ulcerans* genomes (S3 to S8 Files). In the *rpoB* phylogeny, the *C. ulcerans* strains PO100/5, 04–13, 05–13 and W25 clustered with *C. silvaticum* KL0182^{T}, while other *C. ulcerans* strains formed two clades (Fig 1). In the phylogenetic tree of *tox* gene, the same four strains (PO100/5, 04–13, 05–13 and W25) also clustered with *C. silvaticum*, and were distinct from the *C. ulcerans*, *C. diphtheriae* and *C. pseudotuberculosis* clusters (Fig 2). MLST analysis classified strains 04–13, 05–13 and W25 as ST578 (53-60-121-70-76-66-57) and identified PO100/5 as having a unique and new sequence type, ST709 (53-60-121-70-76-82-57) where it differed from ST578 in the locus *odhA*. Due to those results those four strains were reclassified for the next analyses, changing the number of *C. silvaticum* genomes from 34 to 38 and *C. ulcerans* genomes from 80 to 76. Three new STs were identified, ST710 and ST711 in *C. ulcerans* lineage 1 and ST712 in lineage 2 (S9 File, Fig 3).

The taxonomic analyses led to additional insights. TYGS classified nine *C. ulcerans* strains (03–8664, 04–7514, 131002, FRC11, KL0349, LSPQ-04227, LSPQ-04228, NCTC8666 and NCTC12077) as being part of a potential new species. These genomes had dDDH values greater than 70% (99.8–75.9%) within them and less than 70% (63 to 67.2%) with other *C. ulcerans* genomes, although the difference in the G+C content was less than 1% (S3 File). In ANI analysis, those nine genomes were more similar to each other than to the other genomes. They had values between 95.52 and 96.57% with *C. ulcerans* NCTC7910^{T}, and ≥ 97.82% when one of them (NCTC12077) was used as reference for the other eight (S2 File). MLST analysis classified them as having the unique sequence types ST335, ST341, ST344 and the new ST ST712 (Fig 3, S9 File). The ANI analysis showed 99.3% identity between *C. diphtheriae lausannense* strains CHUV2995 and *C. belfantii* FRC0043^{T} (S2 File). A further analysis using TYGS classified *C. diphtheriae lausannense* strains CHUV2995^{T} and CMCNS703 as belonging to *C. belfantii*, and as *C. diphtheriae* the non-type strains with genomes deposited in GenBank as *C. belfantii* (https://www.ncbi.nlm.nih.gov/genome/78252/) (S10 File). For this reason, *C. belfantii* 2937 was renamed to *C. diphtheriae* 2937 in Fig 2.

Genome plasticity analysis

The presence of genes encoding 16 niche and virulence factors described in the genus *Corynebacterium* [35, 36] were examined in *C. silvaticum* (Table 1). The genes *rhuM*, *rpb* and *tspA*

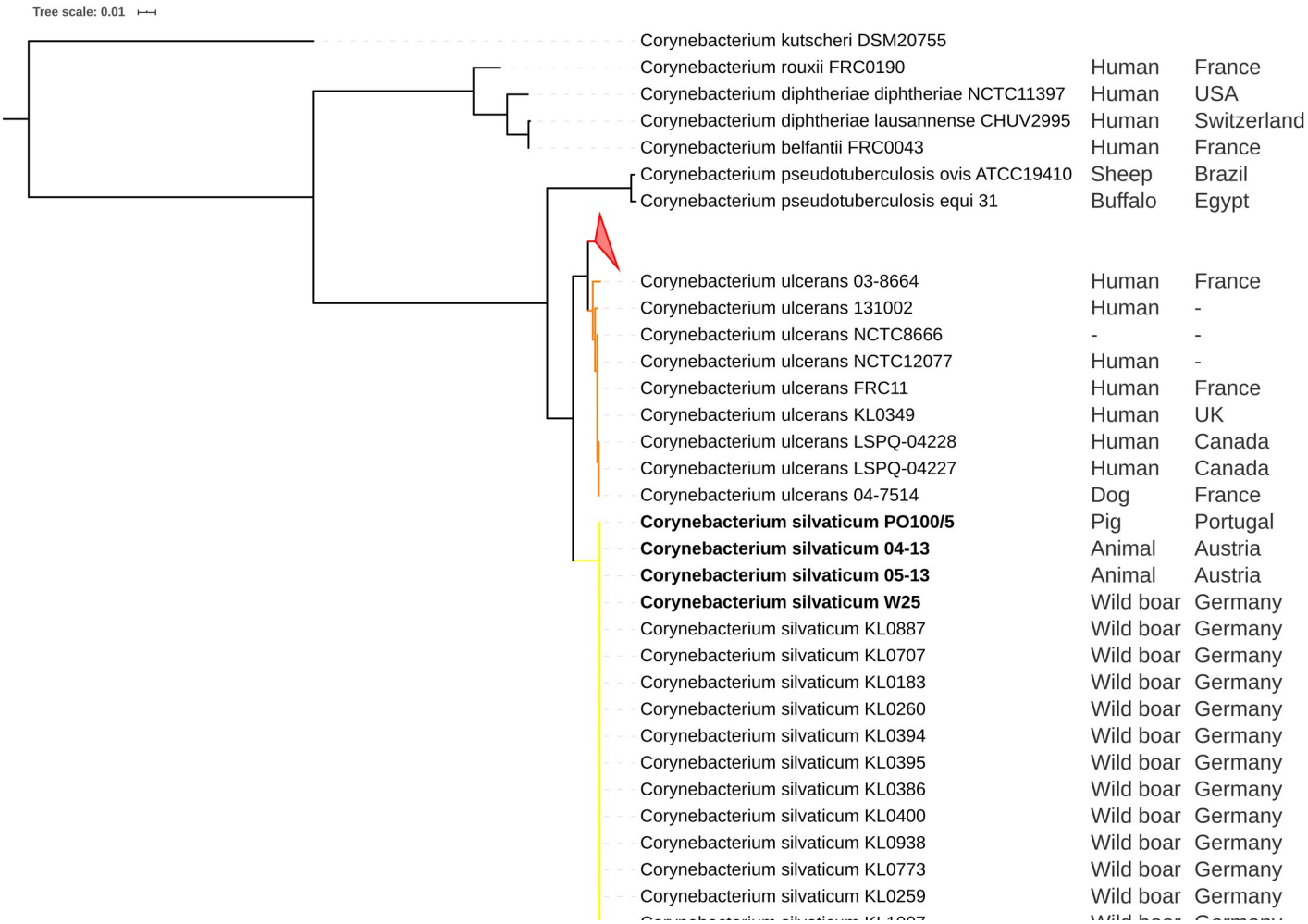


Fig 1. Phylogenetic tree of *rpoB* gene from *Corynebacterium* species. The phylogeny was inferred using the Maximum Likelihood method and the Tamura-Nei (TN93 + G) model implemented in the Mega v10.1.6. The *Corynebacterium ulcerans* strains PO100/5, 04–13, W25 and 05–13 cluster with the *Corynebacterium silvaticum* KL0182^T (yellow). *C. ulcerans* strains form lineage 1 (red, collapsed) and lineage 2 (orange).

https://doi.org/10.1371/journal.pone.0244210.g001

were not found, and all the pilus genes, except *spaB*, were found to be pseudogenized, lacking the signal peptide or CWSS. *C. silvaticum* has the two pilus gene clusters structured as *srtA*, *spaB*C, and *srtB*, *spaD*, *srtC* and *spaEF*, despite fragmentation of pilin genes. Only eight genomes had the *tox* gene (04–13, 05–13, KL0182, KL0884, KL0957, KL1196, PO100/5 and W25). PO100/5 and 05–13 do not have a two bases insertion (GG) after position 44, in a homopolymer of four guanines, that introduces a frameshift (S1 Fig). The mapping of PO100/5 sequencing reads to its assembled genome, showed an insertion of one guanine in the beginning of the homopolymer, in 5% of the reads (S2 Fig). The *tspA* gene was present in all *C. ulcerans* strains, but *rpb* was only found in strain 809, and *rhuM* was found in the 16 strains from Austria, France and Germany that had been isolated from humans, cats and dogs (02–13, FRC58, KL0195, KL0246-cb3, KL0251-cb4, KL0252-cb5, KL0349, KL0387-cb8, KL0475, KL0497, KL0541, KL0547, KL0796, KL0867, KL0880, NCTC12077).

Sixteen and eight genomic islands were predicted by comparing PO100/5 with the reference strains *C. pseudotuberculosis* ATCC19410^T and *C. ulcerans* NCTC7910^T, respectively. No island was detected when it was compared to *C. silvaticum* KL0182^T (Table 2, Fig 4). The

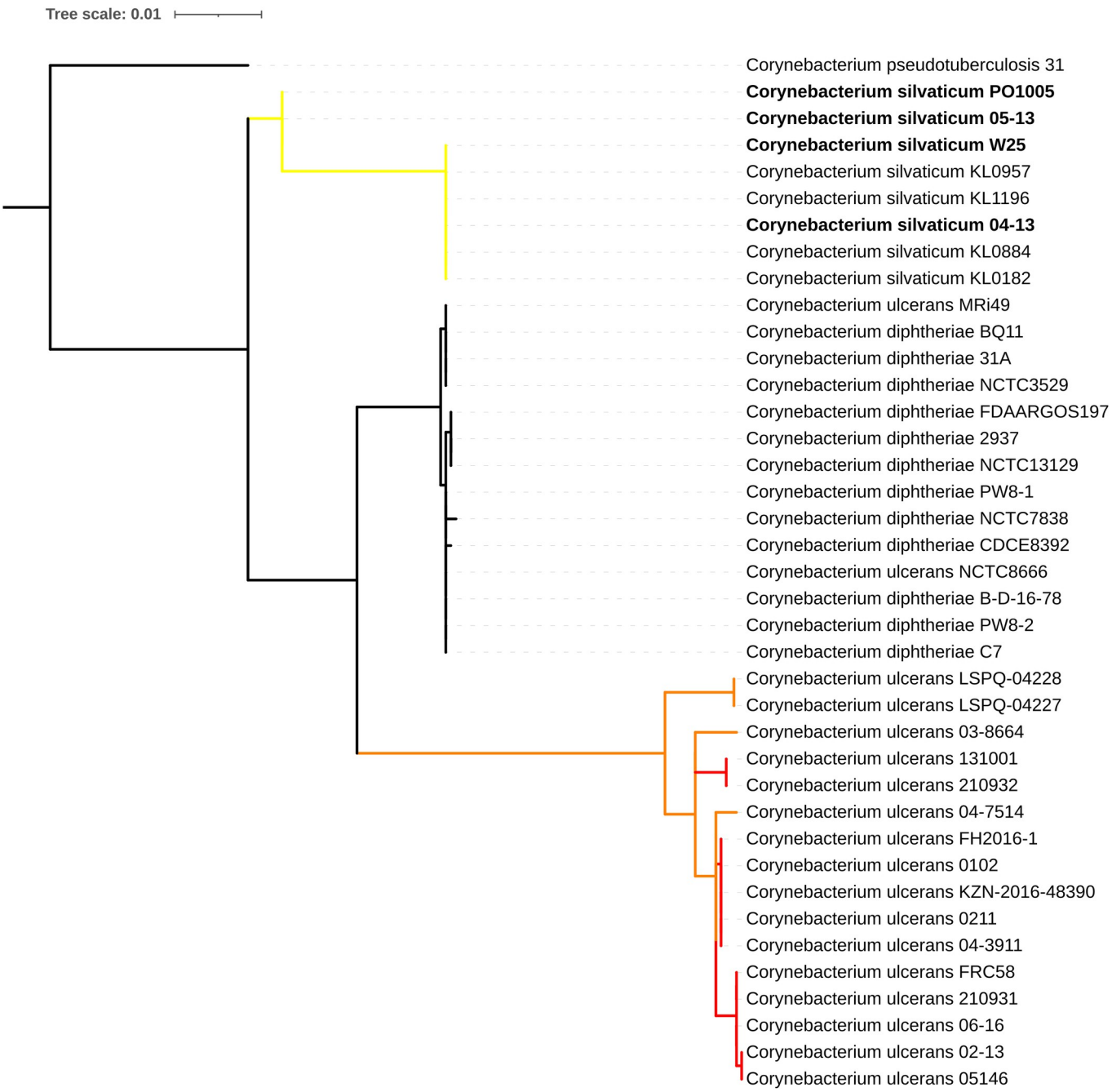


Fig 2. Phylogenetic tree of the *tox* gene from *Corynebacterium* species. The phylogeny was inferred using the Maximum Likelihood method and the Tamura-Nei (T92 + G) model implemented in Mega v10.1.6. The strains PO100/5, 04–13, W25 and 05–13 cluster with *Corynebacterium silvaticum*. Clade colors represent *rpoB* clades of *C. silvaticum* (yellow) and *C. ulcerans* (red and orange).

https://doi.org/10.1371/journal.pone.0244210.g002

genes in the discovered islands are provided in S11 File. They include one complete and three incomplete prophages. Prophage I harbors the *tox* gene and is similar to Gordonia phage Nyceirae (NC_031004.1) (S11 File, Fig 5). BLASTn of the tox^{+} prophage sequence using the GenBank nr database identified the best hits as *C. ulcerans* strains 0102 and 0211, with the same coverage (63%) and identity (92.68%). The best hits with other species were *C.*

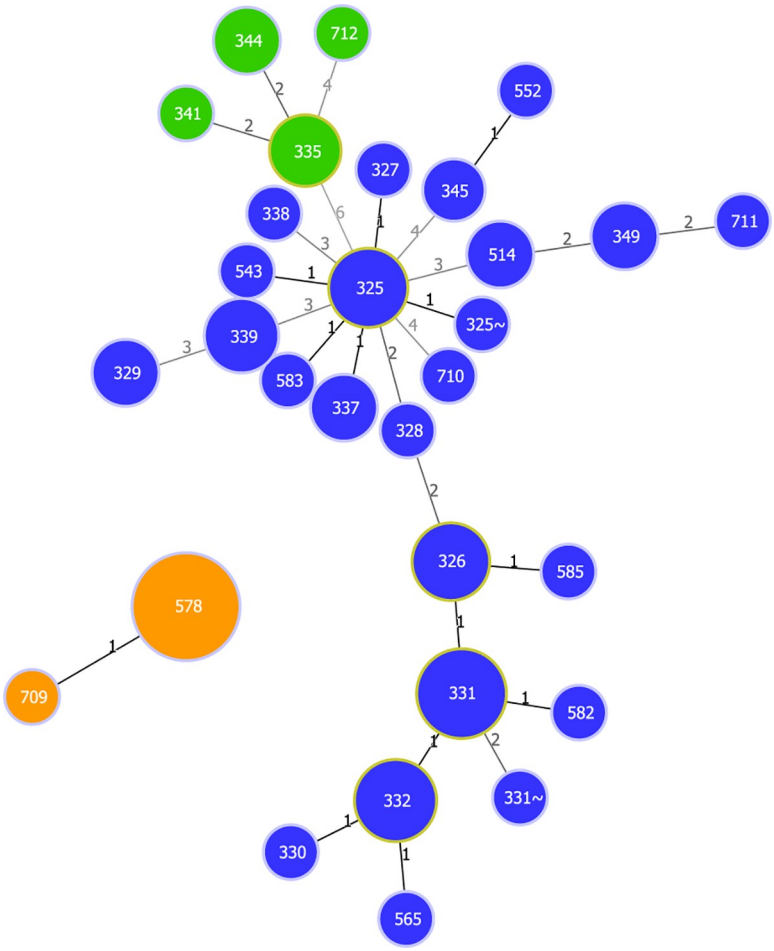


Fig 3. goeBURST diagram for the MLST data set of 38 *Corynebacterium silvaticum* and 76 *C. ulcerans* strains generated using PHILOViZ 2.0. Blue–*C. ulcerans* lineage 1. Green—*C. ulcerans* lineage 2. Orange–*C. silvaticum*. The numbers on the links indicate the number of divergent alleles between STs.

https://doi.org/10.1371/journal.pone.0244210.g003

diphtheriae lausannense (*C. belfantii*) CMCNS703 (37 and 85.95%), *C. diphtheriae* strain B-D-16-78 (41 and 85.86%) and 15 strains of *C. pseudotuberculosis* (14 and 84.94%). Fig 5 shows the alignment of PO100/5 and *C. ulcerans* 0102 *tox*$^{+}$ prophages.

For the pangenome analysis, the number of orthogroups in each subset is shown in Table 3 and S12 File for *C. silvaticum* and *C. ulcerans*. The core genome represented 73.6% and 40% of orthogroups for *C. silvaticum* and *C. ulcerans*, respectively. The pangenome, core genome and singletons development graphs and formulas are shown in Fig 6. Both species had genes conserved in all strains that were absent in the other species, or the exclusive core. In *C. silvaticum*, 172 orthogroups were detected in this subset. They are represented in strain PO100/5 by 174 proteins, 81 of which are located across genomic islands 1, 2, 5, 6, 8, 9, 10, 11, 12 and 14. *C. ulcerans* lineage 2 had a hypothetical protein with 37 amino acids (S12 File). A graph comparing the distribution of Cluster of Homologous Groups (COG) categories of the exclusive core genome of both species is shown in Fig 7.

Finally, PO100/5 (isolated from domestic pig) and KL0182^{T} (wild boar) were predicted to be potential human pathogens by PathogenFinder, with 14 and 13 matches with proteins associated to pathogens, respectively (S13 File).

Table 1. Presence of 16 known niche and virulence factors of *Corynebacterium* in *C. silvaticum*.

Gene	Product	Reference locus tag	Reference protein family	*C. silvaticum* protein family	Manual curation
endoE	Endoglycosidase E (former corynebacterial protease CP40)	CULC809_01974	PLF_1716_00006954	PLF_1716_00006954	Present
cwlH	Cell wall-associated hydrolase	CULC809_01521	PLF_1716_00062893	PLF_1716_00062893	Present
nanH	Sialidase (neuraminidase H)	CULC809_00434	PLF_1716_00002393	PLF_1716_00002393	Present
pld	Phospholipase D	CULC809_00040	PLF_1716_00029465	PLF_1716_00029465	Present
rbp	Shiga-like ribosome-binding protein	CULC809_00177	PLF_1716_00033486	-	Absent
rhuM	RhuM-like protein	CulFRC58_0285	PLF_1716_00026137	-	Absent
rpfI	Resuscitation-promoting factor-interacting protein	CULC809_01133	PLF_1716_00001449	PLF_1716_00001449	Present
spaB	Surface-anchored protein (minor pilus subunit)	CULC809_01980	PLF_1716_00010184	PLF_1716_00010184	Present
spaC	Surface-anchored protein (pilus tip protein)	CULC809_01979	PLF_1716_00004783	PLF_1716_00004783	Pseudogene, no cell wall sorting signal
spaD	Surface-anchored protein (major pilus subunit)	CULC809_01952	PLF_1716_00090862	PLF_1716_00102654	Pseudogene, no cell wall sorting signal
spaE	Surface-anchored protein (minor pilus subunit)	CULC809_01950	PLF_1716_00007274	PLF_1716_00079271	Pseudogene, no signal peptide
spaF	Surface-anchored protein (pilus tip protein)	CULC809_01949	PLF_1716_00006760	PLF_1716_00006760	Pseudogene, no signal peptide
tox	Diphtheria toxin	CULC0102_0213	PLF_1716_00005191	PLF_1716_00005191	Present in 8 out of 38 genomes
tspA	Trypsin-like serine protease	CULC809_01848	PLF_1716_00007827	-	Absent
vsp1	Venom serine protease	CULC809_00509	PLF_1716_00104602	PLF_1716_00104343	64% identity with *C. ulcerans* 809
vsp2	Venom serine protease	CULC809_01964	PLF_1716_00015799	PLF_1716_00116381	74% identity with *C. ulcerans* 809
-	*C. diphtheriae* DIP0733 homolog	CULC22_00609	PLF_1716_00030114	PLF_1716_00030114	Present

https://doi.org/10.1371/journal.pone.0244210.t001

Table 2. Genomic islands in strain PO100/5 compared to *Corynebacterium pseudotuberculosis* ATCC19410^T and *C. ulcerans* NCTC7910^T.

n	Position compared to Cp	Size	Position compared to Cul	Size	Type	Prophage content
1	29362–38109	8.75 kb	-	-	-	-
2	55224–60448	5.22 kb	55224–60448	5.22 kb	PA	-
3	69256–74863	5.6 kb	-	-	PA, RE, SY	-
4	97842–103378	5.53 kb	-	-	PA, RE	-
5	167859–206438	38.58 kb	167859–206209	38.35 kb	-	Prophage I
6	311533–318567	7.03 kb	311533–318567	70.34 kb	-	-
7	422293–430839	85.46 kb	-	-	RE, SY	-
8	694806–746966	52.16 kb	694806–746966	52.16 kb	RE	Prophage II
9	925047–938225	13.18 kb	925047–938225	13.18 kb	-	-
10	1235769–1244625	8.86 kb	1235769–1244625	8.86 kb	-	-
11	1613625–1644953	31.33 kb	1614013–1639302	25.29 kb	-	Prophage III
12	1793740–1802246	8.5 kb	-	-	PA, ME	-
13	2029016–2035120	6.1 kb	-	-	-	-
14	2109299–2139505	30.2 kb	2110317–2139505	29.19 kb	-	Prophage IV
15	2255307–2261370	6.06 kb	-	-	-	-
16	2517632–2529863	12.23 kb	-	-	ME	-

Cp–*C. pseudotuberculosis*, Cul–*C. ulcerans*, PA–pathogenicity island, RE–resistance island, ME–metabolic island, SY–symbiotic island.

https://doi.org/10.1371/journal.pone.0244210.t002

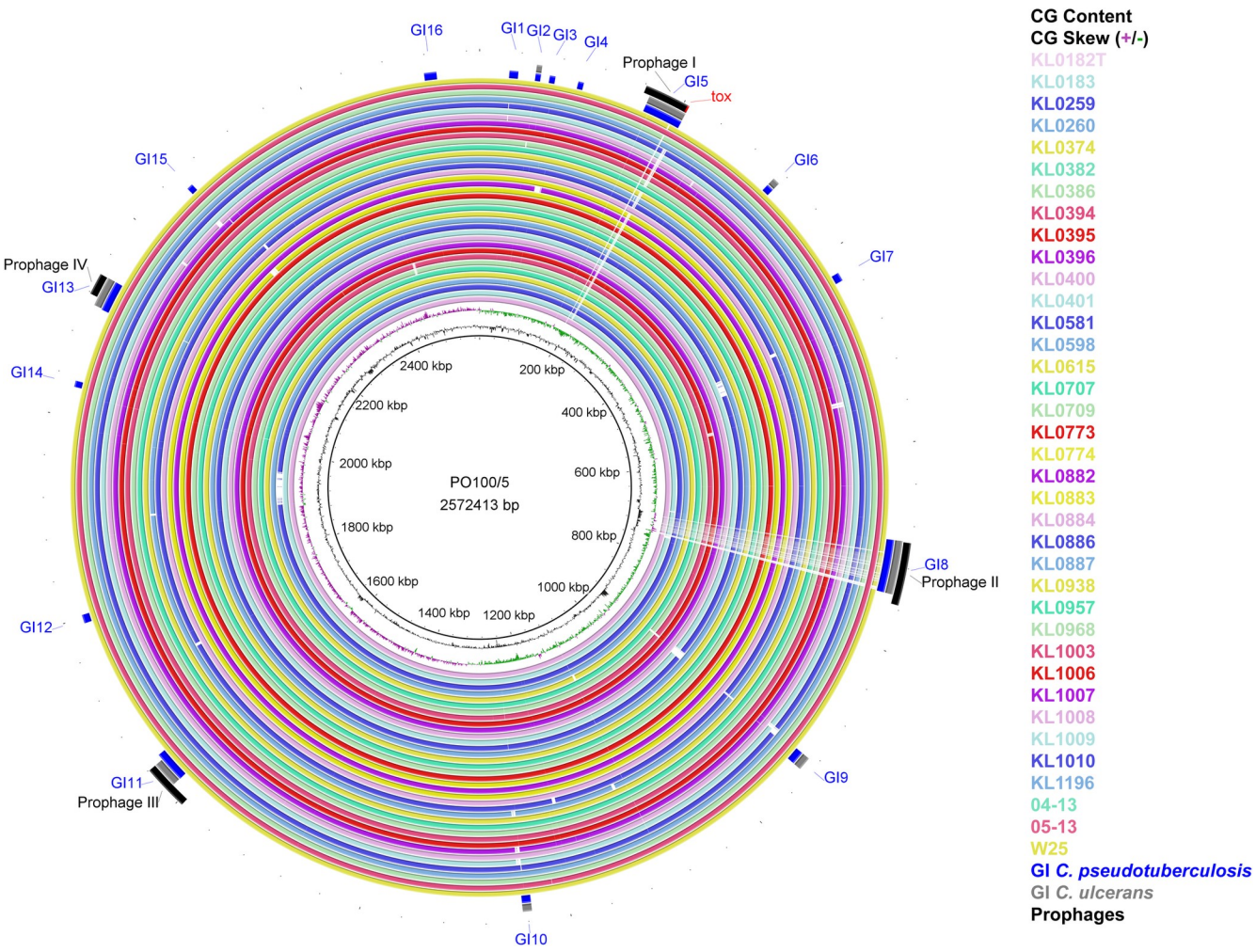


Fig 4. Circular map of *Corynebacterium silvaticum* genomes generated using BRIG v0.95. From inner to outer circle: strain PO100/5 (reference); CG content; CG Skew; strains KL0182^{T}, KL0183, KL0259, KL0260, KL0374, KL0382, KL0386, KL0394, KL0395, KL0396, KL0400, KL0401, KL0581, KL0598, KL0615, KL0707, KL0709, KL0773, KL0774, KL0882, KL0883, KL0884, KL0886, KL0887, KL0938, KL0957, KL0968, KL1003, KL1006, KL1007, KL1008, KL1009, KL1010, KL1196, 04–13, 05–13, W25; genomic islands compared to *C. pseudotuberculosis* ATCC19410^{T} (blue) and *C. ulcerans* strain NCTC7910^{T} (grey); and prophages (black). Genomic islands (GI) and prophage detection were performed using GIPSy and PHASTER, respectively.

https://doi.org/10.1371/journal.pone.0244210.g004

Discussion

Strain PO100/5 was originally classified as *C. pseudotuberculosis*. Its resistance profile was tested for 13 antimicrobial compounds (Amoxycillin/Clavulanic acid, Ampicillin, Chloramphenicol, Cephalexin, Gentamicin, Cefotaxime, Enrofloxacin, Nalidixic acid, Penicillin G, Streptomycin, Sulfamethoxazole/Trimethoprim, Tetracycline and Vancomycin) and it was found to be resistant to nalidixic acid and streptomycin [9]. It was suggested as *C. silvaticum* by a recent *rpoB* phylogeny [4]. We analyzed the genome diversity of this species, using publicly available genomes from the *C. diphtheriae* group (S1 File).

Taxonomic analysis showed that PO100/5, W25, 04–13 and 05–13 are strains of the recently described *C. silvaticum* [4]. This is supported by ANI values above 95% [20] (S2 File), genome and 16S rRNA GBDP clustering [24], dDDH > 70%, G+C content difference > 1% with *C. ulcerans* genomes [21, 24, 26] (S3–S5 Files), *rpoB* phylogenetic clustering (Fig 1) and the unique sequence type ST578 from *C. silvaticum* [4] (Fig 3). Strain PO100/5 has the new ST709

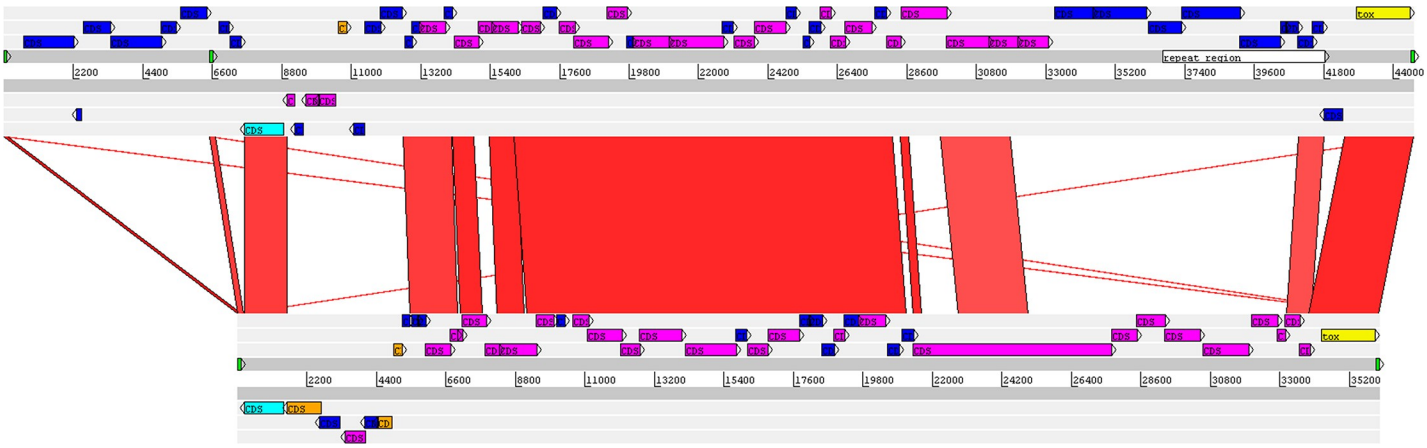


Fig 5. Alignment of *tox*$^{+}$ prophages form *Corynebacterium silvaticum* PO100/5 and *C. ulcerans* 0102. The red lines connect sequences with at least 75% identity. Light blue–Phage integrase; Orange–Phage-related transcriptional regulator; Pink–Phage-related coding sequence (CDS); Yellow–Diphtheria toxin CDS; Dark blue–other CDS; Green–tRNA.

https://doi.org/10.1371/journal.pone.0244210.g005

(S9 File, Fig 4). The misclassification of those strains is expected as, prior to the development of methods to identify *C. silvaticum* [4, 5], the use of biochemistry tests (API Coryne and VITEK2-compact) and the clinical picture would classify these strains as *C. pseudotuberculosis* [4, 48], while DNA sequence analysis and Fourier-transform Infrared Spectroscopy would classify it as *C. ulcerans* [6, 7, 48].

Analysis of genome plasticity identified unique characteristics *C. silvaticum*. The analysis of 16 known niche and virulence factors showed the absence of *rpb*, *rhuM*, *and tsA*. *spaB* was the only non-fragmented pili gene in *C. silvaticum* (Table 1). The Shiga-like ribosome-binding protein (*rpb*) has a ribosome inactivating protein domain that has only been reported in *C. ulcerans* 809 [35, 49]. The new species also has a RhuM-like protein (*rhuM*), which has only been seen previously in the *C. ulcerans* strain KL0387 [50]. A RhuM mutant of *Salmonella enterica* had a significant decrease in epithelial cell invasion [51]. We identified this protein in 15 other strains from humans, dogs and cats form Austria, France and Germany.

Serine proteases can promote the survival and dissemination of pathogens in the host [52], and we looked for these virulence factors in the genomes we analyzed (Table 1). Venom serine proteases (*vsp1* and *vsp2*) and Trypsin-like serine protease (*tspA*) are secreted proteases that could have multiple potential pathogenic functions [53]. There is homology between the two serine proteases found in *C. ulcerans* in this new species (Table 1), but *tspA* was not found in *C. silvaticum*. Its absence could be used as a marker to differentiate it from *C. ulcerans*.

Bacterial pili are adhesion structures required for colonization of host tissues. The *Corynebacterium* pili are SpaA-type, with a heterotrimeric structure composed by major (pilus shaft), minor and tip pilins, the last two required for adhesion. The pilus is assembled and anchored

Table 3. Number of genes (orthogroups) in subsets across *Corynebacterium silvaticum* and *C. ulcerans* genomes.

Species	Genomes	Pangenome	Core genome	Accessory genome	Shared genome	Singletons	α
C. silvaticum	38	3,002	2,209	703	603	190	0.9520
C. ulcerans	76	4,351	1,747	2,604	1,706	898	0.8142
C. silvaticum and *C. ulcerans*	114	4,916	1,618	3,298	2,349	949	_

The pangenome is the entire repertoire of orthogroups, the core genome is the subset conserved across all genomes (100%), the accessory genome is the subset not conserved across all genomes, singletons are exclusive from a genome, and the shared genome are orthogroups shared by two or more, but not all genomes.

https://doi.org/10.1371/journal.pone.0244210.t003

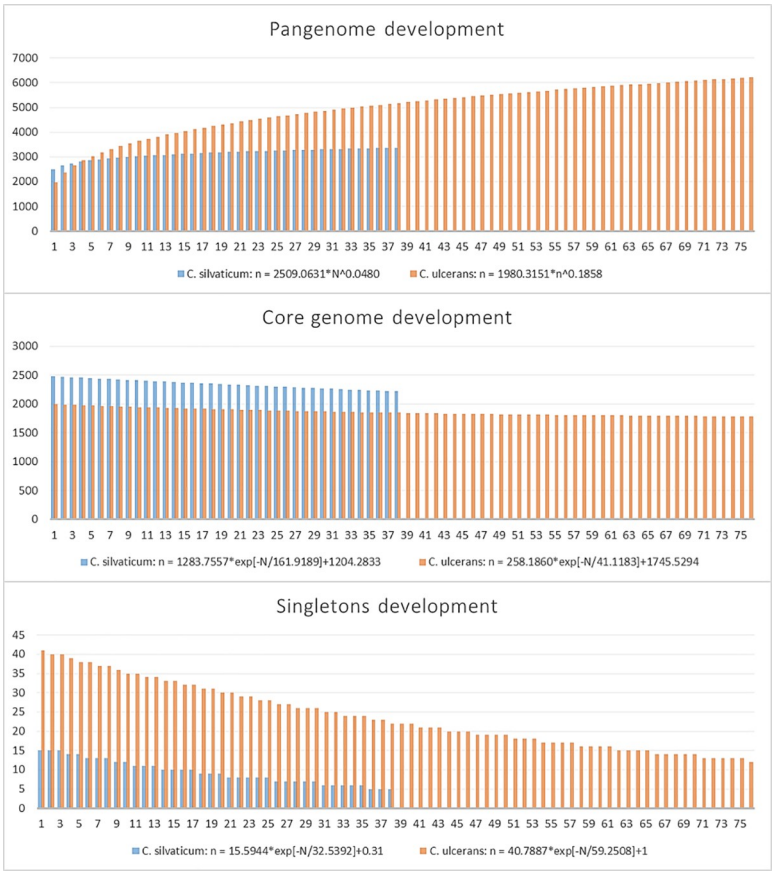


Fig 6. Pangenome, core genome and singletons development graphs and formulas for 38 genomes of *Corynebacterium silvaticum* and 76 *C. ulcerans* genomes.

https://doi.org/10.1371/journal.pone.0244210.g006

to the cell wall by the housekeeping sortase SrtA and pili sortases SrtB and SrtC [54]. As seen in *C. ulcerans* [35], *C. silvaticum* has the two pili gene clusters *spaB*C and *spaDEF*, although only *spaB* appears to be functional due to the presence of a signal peptide and a CWSS. The SpaB is a minor pilin that in *C. diphtheriae* has a role in adhesion on pharyngeal epithelial cells and could be functional when linked to the cell wall [55] as shown for the heterodimeric structure SpaB-SpaC in *C. diphtheriae* [56] and suggested for *C. ulcerans* [35].

The *tox* gene was found in only eight out of the 38 *C. silvaticum* strains (04–13, 05–13, KL0182, KL0884, KL0957, KL1196, PO100/5 and W25), although the strains lacking it were reported to be NTTB [6]. This can be seen in the circular map as a blank space in the *tox* gene region of the other 30 strains (Fig 4). The absence of the *tox* gene in the other strains could be the result of an assembly artifact, due to a repetitive region prior to this gene. Additionally, the *tox* sequences from PO100/5 and 05–13 lack the insertion of two guanines in position 44 (S1 and S2 Figs) that causes pseudogenization, characteristic of other *C. silvaticum* strains [10]. A recent publication showed that strains 04–13 and 05–13 from Austria produce the *tox* transcript by reverse transcriptase quantitative PCR (RT-qPCR) [8]. As 04–13 has the frameshift, 05–13 and PO10/5 could be the only known toxigenic *C. silvaticum* strains. The production of DT has yet to be tested.

In PO100/5, four incomplete prophages were found, one harboring the *tox* gene. When PO100/5 was compared to *C. pseudotuberculosis* ATCC 19410^{T}, sixteen genomic islands were

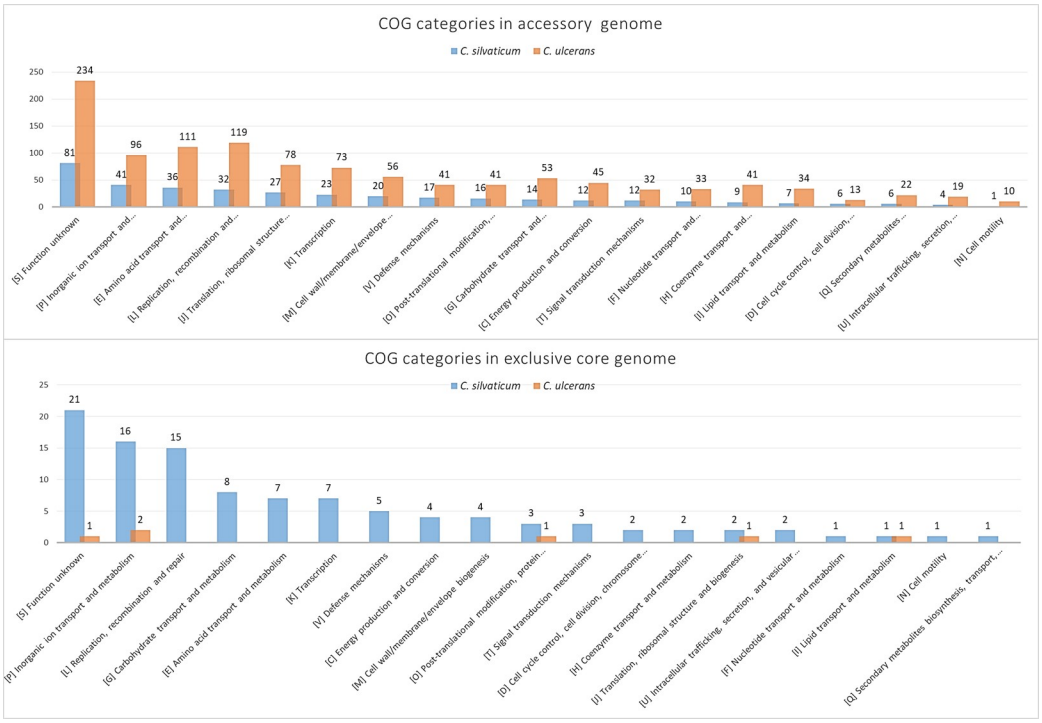


Fig 7. Clusters of orthologous groups (COGs) in the accessory and exclusive core genomes of *Corynebacterium silvaticum* and *C. ulcerans* annotated using eggNOG-mapper v2. COG categories are sorted from most abundant to less abundant in *C. silvaticum*. In *C. silvaticum*, 356 out of 793 proteins from the accessory genome and 93 out of 174 proteins of the exclusive core genome had a COG category. In *C. ulcerans*, 1,065 out of 2,064 proteins from the accessory genome and six out of nine proteins of the exclusive core genome had a COG category.

https://doi.org/10.1371/journal.pone.0244210.g007

identified. When it was compared to *C. ulcerans* NCTC7910^T, only eight islands were found. Four of the islands contained the prophages: GI5, GI8, GI11 and GI13 (S11 File, Figs 4 and 5). No island was found in comparison to *C. silvaticum* KL0182^T. Genomic islands are mobile genetic elements (MGEs) acquired by horizontal gene transfer that can provide adaptive traits [33]. In a previous study, MGEs containing *tox* in *C. diphtheriae* were identified as known prophages, while in *C. ulcerans* they can be different prophages or an alternative pathogenicity island. These mobile elements showed nearly species-specific clades, including the atypical *C. ulcerans* clade that now represents *C. silvaticum*. This implies independent events of acquisition of virulence factors in zoonotic species that could influence their pathogenic potential [6].

C. silvaticum was estimated to be more genetically homogeneous than *C. ulcerans* and to have a pangenome near to being closed, with bigger values of core genome development (Fig 6) and α closer to 1 (Table 3). This result could be influenced by the samples of *C. silvaticum* being from only two separate countries, Germany (n = 37) and Portugal (n = 1), and from two different species of host (*Sus scrofa* and *Capreolus capreolus*). This estimation could change once more genomes are sequenced. A total of 172 and 8 orthogroups were uniquely shared by all *C. silvaticum* and *C. ulcerans*, respectively, some in the described genomic islands (S12 File, Fig 7). For *C. silvaticum*, the most abundant functions are involved in nutrient acquisition such as transport and metabolism of inorganic ions, carbohydrates and amino acids (COG categories E, G and P), or are related to phages or immunity against them (COG category L). For example, two of them are a Type I restriction-modification system [57] in genomic island 11 and an "ABC-type dipeptide oligopeptide nickel transport system". The function of those genes in the phenotype and infection must be investigated, but they are candidates for genetic

markers for a rapid and cost-effective diagnostic using multiplex polymerase chain reaction (PCR) [58–60] and other established methods [4, 5].

In addition to being of veterinary importance, *C. silvaticum* could have medical relevance, as strains PO100/5 and KL1082 were predicted to be potential human pathogens (S13 File). The known host range of *C. silvaticum* is limited to wild boars, domestic pigs and roe deer [4, 7, 9]. Wild boars are reservoirs for viruses, bacteria and other parasites that can be transmitted to livestock and humans, during opportunities provided by deforestation and use of lands for agricultural purposes, hunting activities and consumption of wild boar meat [14]. Although they are transmitted additionally by other hosts, pigs and boars are a reservoir of *C. ulcerans*, which can cause zoonotic transmission to humans [11–13]. By the same route, *C. silvaticum* could be transmitted to humans and cause infection. In addition, it could be misidentified as *C. ulcerans* or *C. pseudotuberculosis* due to limitations in the standard methodology [4, 5].

Additionally, the TYGS results suggest that nine *C. ulcerans* corresponding to lineage 2 [49] is a potential new species, with dDDH of less than 70% with lineage 1 genomes. (S3–S6 Files). Further investigation is required to verify whether this lineage could be classified as a new species. Recently, *C. belfantii* and *C. diphtheriae lausannense* were suggested as synonyms [2]. Our analysis using TYGS corroborated that suggestion. In addition, besides strains FRC0043^{T}, CHUV2995^{T} and CMCNS703, the other nine genomes deposited in GenBank as *C. belfantii* (https://www.ncbi.nlm.nih.gov/genome/78252/) were classified as *C. diphtheriae* (S10 File). These results suggest the limitation of using only one cutoff as a parameter for taxonomic classification.

Conclusions

The taxonomic analysis shows PO100/5 and four other genomes deposited as *C. ulcerans* are from the recently described species *C. silvaticum*. The comparative genomic analysis showed this species is more genetically homogeneous than *C. ulcerans*, has SpaB as the only probably functional pilin subunit, and has conserved genomic islands and 172 genes that could be used as molecular markers for PCR identification. In contrast to the other strains from the same species, PO100/5 is the first one to be isolated from livestock and outside Germany and Austria, and to have the unique ST709. A non pseudogenized *tox* gene in PO100/5 and 05–13 suggest those strains could produce the diphtheria toxin.

Supporting information

S1 Fig. Alignment of tox gene from *Corynebacterium silvaticum*, *C. ulcerans*, *C. pseudotuberculosis* and *C. ulcerans*. The alignment was performed using MUSCLE algorithm implemented in MEGA v10.1.6. *C. silvaticum* strains PO100/5 and 05–13 do not have a two guanines insertion that lead to a frameshift in other strains from this species.
(TIF)

S2 Fig. Mapping of sequencing reads to the tox gene of *Corynebacterium silvaticum* strain PO100/5.
(TIF)

S1 File. Genomes of *Corynebacterium* species used for taxonomic analysis of the PO100/5 strain.
(XLSX)

S2 File. Average Nucleotide Identity and digital DNA-DNA hybridization among strains of *Corynebacterium*. Strains highlighted in blue are *Corynebacterium ulcerans* strains from

lineage 2. ANI values under 78.64% are showed as "NA".
(XLSX)

S3 File. Taxonomic classification of 80 *Corynebacterium ulcerans* genomes from GenBank by digital DNA-DNA hybridization and G+C content. The analysis was performed using Type Strain Genome Server.
(XLSX)

S4 File. Taxonomic classification of 20 *Corynebacterium ulcerans* genomes from GenBank. The analysis was performed using Type Strain Genome Server. Strains 04/13 and 05/13 were classified as *C. silvaticum*, while 03–8664, 04–7514, 131002 were classified as a potential new species.
(PDF)

S5 File. Taxonomic classification of 15 *Corynebacterium ulcerans* genomes from GenBank. The analysis was performed using Type Strain Genome Server. Strains PO100/5 and W25 were classified as *C. silvaticum*, while FRC11, LSPQ-04227, LSPQ-04228, NCTC8666 and NCTC12077 were classified as a potential new species.
(PDF)

S6 File. Taxonomic classification of 20 *Corynebacterium ulcerans* genomes from GenBank, available as sequencing data. The analysis was performed using Type Strain Genome Server. Strain KL0349 was classified as a potential new species.
(PDF)

S7 File. Taxonomic classification of 20 *Corynebacterium ulcerans* genomes from GenBank, available as sequencing data. The analysis was performed using Type Strain Genome Server. All strains were classified as *C. ulcerans*.
(PDF)

S8 File. Taxonomic classification of five *Corynebacterium ulcerans* genomes from GenBank, available as sequencing data. The analysis was performed using Type Strain Genome Server. All strains were classified as *C. ulcerans*.
(PDF)

S9 File. Taxonomic classification of 2 *Corynebacterium diphtheriae lausannense* and 10 *C. belfantii* genomes from GenBank. The analysis was performed using Type Strain Genome Server. *C. diphtheriae lausannense* strains CHUV2995^{T} and CMCNS703 were classified as *C. belfantii*, while *C. belfantii* strains except the type strain FRC0043^{T} were classified as *C. diphtheriae*.
(XLSX)

S10 File. Multilocus sequence typing data of *Corynebacterium silvaticum* and *C. ulcerans* genomes. The analysis was performed using MLSTchecker.
(PDF)

S11 File. Genomic islands content in strain PO100/5.
(XLSX)

S12 File. Pangenome analysis of 38 *Corynebacterium silvaticum* and 76 *C. ulcerans* samples. Gene homology groups were predicted using OrthoFinder v2.12.2 and functional annotation was performed using eggNOG-mapper v2.
(XLSX)

S13 File. Probability of pathogenicity for humans. PathogenFinder v. 1.1 was used to identify proteins associated to bacterial pathogens in the proteome of *C. silvaticum* PO100/5 (isolated from domestic pig) and KL0182^T (wild boar).
(XLSX)

Author Contributions

Conceptualization: Marcus Vinicius Canário Viana, Alice Rebecca Wattam.

Data curation: Marcus Vinicius Canário Viana, Janaína Canário Cerqueira.

Formal analysis: Marcus Vinicius Canário Viana.

Funding acquisition: Marcus Vinicius Canário Viana, Henrique Figueiredo, Preetam Ghosh, Artur Silva, Vasco Azevedo.

Investigation: Marcus Vinicius Canário Viana, Rodrigo Profeta, Raquel Hurtado, Alice Rebecca Wattam, Vasco Azevedo.

Methodology: Marcus Vinicius Canário Viana.

Resources: Manuela Oliveira, Luís Tavares, Henrique Figueiredo.

Software: Siomar de Castro Soares.

Supervision: Artur Silva, Vasco Azevedo.

Validation: Marcus Vinicius Canário Viana, Janaína Canário Cerqueira.

Writing – original draft: Marcus Vinicius Canário Viana, Alessandra Lima da Silva.

Writing – review & editing: Marcus Vinicius Canário Viana, Rodrigo Profeta, Alessandra Lima da Silva, Raquel Hurtado, Janaína Canário Cerqueira, Bruna Ferreira Sampaio Ribeiro, Marcelle Oliveira Almeida, Francielly Morais-Rodrigues, Siomar de Castro Soares, Manuela Oliveira, Luís Tavares, Alice Rebecca Wattam, Debmalya Barh, Preetam Ghosh, Vasco Azevedo.

References

1. Bernard AL, Funke G. *Corynebacterium*. Bergey's Manual of Systematic of Archaea and Bacteria (Online). John Wiley & Sons, Bergey's Manual Trust; 2015. pp. 1–70. https://doi.org/10.1002/9781118960608
2. Badell E, Hennart M, Rodrigues C, Passet V, Dazas M, Panunzi L, et al. *Corynebacterium rouxii* sp. nov., a novel member of the *diphtheriae* species complex. Res Microbiol. 2020; 171: 122–127. https://doi.org/10.1016/j.resmic.2020.02.003 PMID: 32119905
3. Dazas M, Badell E, Carmi-Leroy A, Criscuolo A, Brisse S. Taxonomic status of *Corynebacterium diphtheriae* biovar Belfanti and proposal of *Corynebacterium rouxii* sp. nov. Int J Syst Evol Microbiol. 2018; 68: 3826–3831. https://doi.org/10.1099/ijsem.0.003069 PMID: 30355399
4. Dangel A, Berger A, Rau J, Eisenberg T, Kämpfer P, Margos G, et al. *Corynebacterium silvaticum* sp. nov., a unique group of NTTB corynebacteria in wild boar and roe deer. Int J Syst Evol Microbiol. 2020. https://doi.org/10.1099/ijsem.0.004195 PMID: 32368999
5. Rau J, Eisenberg T, Peters M, Berger A, Kutzer P, Lassnig H, et al. Reliable differentiation of a non-toxigenic tox gene-bearing *Corynebacterium ulcerans* variant frequently isolated from game animals using MALDI-TOF MS. Vet Microbiol. 2019; 237: 108399. https://doi.org/10.1016/j.vetmic.2019.108399 PMID: 31585651
6. Dangel A, Berger A, Konrad R, Sing A. NGS-based phylogeny of diphtheria-related pathogenicity factors in different *Corynebacterium* spp. implies species-specific virulence transmission. BMC Microbiol. 2019; 19: 1–16. https://doi.org/10.1186/s12866-018-1372-8 PMID: 30616583

7. Eisenberg T, Kutzer P, Peters M, Sing A, Contzen M, Rau J. Nontoxigenic tox-bearing *Corynebacterium ulcerans* Infection among Game Animals, Germany. Emerg Infect Dis. 2014; 20: 448–452. https://doi.org/10.3201/eid2003.130423 PMID: 24572455

8. Schaeffer J, Huhulescu S, Stoeger A, Allerberger F, Ruppitsch W. Draft Genome Sequences of Six *Corynebacterium ulcerans* Strains Isolated from Humans and Animals in Austria, 2013 to 2019. Gill SR, editor. Microbiol Resour Announc. 2020; 9: 1–3. https://doi.org/10.1128/MRA.00946-20 PMID: 32883799

9. Oliveira M, Barroco C, Mottola C, Santos R, Lemsaddek A, Tavares L, et al. First report of *Corynebacterium pseudotuberculosis* from caseous lymphadenitis lesions in Black Alentejano pig (Sus scrofa domesticus). BMC Vet Res. 2014; 10: 218. https://doi.org/10.1186/s12917-014-0218-3 PMID: 25240511

10. Möller J, Musella L, Melnikov V, Geißdörfer W, Burkovski A, Sangal V. Phylogenomic characterisation of a novel corynebacterial species pathogenic to animals. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol. 2020; 5. https://doi.org/10.1007/s10482-020-01430-5 PMID: 32500295

11. Schuhegger R, Schoerner C, Dlugaiczyk J, Lichtenfeld I, Trouillier A, Zeller-Peronnet V, et al. Pigs as Source for Toxigenic *Corynebacterium ulcerans*. Emerg Infect Dis. 2009; 15: 1314–1315. https://doi.org/10.3201/eid1508.081568 PMID: 19751602

12. Berger A, Boschert V, Konrad R, Schmidt-Wieland T, Hörmansdorfer S, Eddicks M, et al. Two Cases of Cutaneous Diphtheria Associated with Occupational Pig Contact in Germany. Zoonoses Public Health. 2013; 60: 539–542. https://doi.org/10.1111/zph.12031 PMID: 23280311

13. Konig C, Meinel DM, Margos G, Konrad R, Sing A. Multilocus Sequence Typing of *Corynebacterium ulcerans* Provides Evidence for Zoonotic Transmission and for Increased Prevalence of Certain Sequence Types among Toxigenic Strains. J Clin Microbiol. 2014; 52: 4318–4324. https://doi.org/10.1128/JCM.02291-14 PMID: 25320226

14. Meng XJ, Lindsay DS, Sriranganathan N. Wild boars as sources for infectious diseases in livestock and humans. Philos Trans R Soc B Biol Sci. 2009; 364: 2697–2707. https://doi.org/10.1098/rstb.2009.0086 PMID: 19687039

15. Seth-Smith HMB, Egli A. Whole genome sequencing for surveillance of diphtheria in low incidence settings. Frontiers in Public Health. 2019. pp. 1–13. https://doi.org/10.3389/fpubh.2019.00001 PMID: 30809516

16. Berger A, Hogardt M, Konrad R, Sing A. Detection Methods for Laboratory Diagnosis of Diphtheria. *Corynebacterium diphtheriae* and Related Toxigenic Species. Dordrecht: Springer Netherlands; 2014. pp. 171–205. https://doi.org/10.1007/978-94-007-7624-1_9

17. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res. 2017; 45: D535–D542. https://doi.org/10.1093/nar/gkw1017 PMID: 27899627

18. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19: 455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

19. Brettin T, Davis JJ, Disz T, Edwards R A, Gerdes S, Olsen GJ, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep. 2015; 5: 1–6. https://doi.org/10.1038/srep08365 PMID: 25666585

20. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018; 9: 5114. https://doi.org/10.1038/s41467-018-07641-9 PMID: 30504855

21. Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun. 2019; 10: 2182. https://doi.org/10.1038/s41467-019-10210-3 PMID: 31097708

22. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016; 17: 132. https://doi.org/10.1186/s13059-016-0997-x PMID: 27323842

23. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

24. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013;14. https://doi.org/10.1186/1471-2105-14-14 PMID: 23323936

25. Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, Fiebig A, et al. Complete genome sequence of DSM 30083T, the type strain (U5/41T) of *Escherichia coli*, and a proposal for delineating

subspecies in microbial taxonomy. Stand Genomic Sci. 2014; 9: 2. https://doi.org/10.1186/1944-3277-9-2 PMID: 25780495

26. Meier-Kolthoff JP, Klenk HP, Göker M. Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. Int J Syst Evol Microbiol. 2014; 64: 352–356. https://doi.org/10.1099/ijs.0.056994-0 PMID: 24505073

27. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993; 10: 512–26. https://doi.org/10.1093/oxfordjournals.molbev.a040023 PMID: 8336541

28. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Battistuzzi FU, editor. Mol Biol Evol. 2018; 35: 1547–1549. https://doi.org/10.1093/molbev/msy096 PMID: 29722887

29. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019; 47: W256–W259. https://doi.org/10.1093/nar/gkz239 PMID: 30931475

30. Page AJ, Taylor B, Keane JA. Multilocus sequence typing by blast from de novo assemblies against PubMLST. J Open Source Softw. 2016; 1: 118. https://doi.org/10.21105/joss.00118

31. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. PHYLOViZ 2.0: Providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics. 2017; 33: 128–129. https://doi.org/10.1093/bioinformatics/btw582 PMID: 27605102

32. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016; 44: W16–W21. https://doi.org/10.1093/nar/gkw387 PMID: 27141966

33. Soares SC, Geyik H, Ramos RTJ, de Sá PHCG, Barbosa EGV, Baumbach J, et al. GIPSy: Genomic island prediction software. J Biotechnol. 2016; 232: 2–11. https://doi.org/10.1016/j.jbiotec.2015.09.008 PMID: 26376473

34. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011; 12: 402. https://doi.org/10.1186/1471-2164-12-402 PMID: 21824423

35. Trost E, Al-Dilaimi A, Papavasiliou P, Schneider J, Viehoever P, Burkovski A, et al. Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. BMC Genomics. 2011; 12: 383. https://doi.org/10.1186/1471-2164-12-383 PMID: 21801446

36. Tauch A, Burkovski A. Molecular armory or niche factors: virulence determinants of *Corynebacterium* species. FEMS Microbiol Lett. 2015; 67: fnv185. https://doi.org/10.1093/femsle/fnv185 PMID: 26449555

37. Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, et al. Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. Bioinformatics. 2008; 24: 2672–2676. https://doi.org/10.1093/bioinformatics/btn529 PMID: 18845581

38. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30: 1236–1240. https://doi.org/10.1093/bioinformatics/btu031 PMID: 24451626

39. Fimereli DK, Tsirigos KD, Litou ZI, Liakopoulos TD, Bagos PG, Hamodrakas SJ. CW-PRED: A HMM-Based Method for the Classification of Cell Wall-Anchored Proteins of Gram-Positive Bacteria. 1st ed. In: Maglogiannis I, Plagianakos V, Vlahavas I, editors. Artificial Intelligence: Theories and Applications. 1st ed. Springer, Berlin, Heidelberg; 2012. pp. 285–290. https://doi.org/10.1007/978-3-642-30448-4_36

40. CLC Genomics Workbench. Available: http://www.clcbio.com/products/clc-main-workbench/

41. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST plus: architecture and applications. BMC Bioinformatics. 2009; 10: 1. Artn 421\nDoi https://doi.org/10.1186/1471-2105-10-1 PMID: 19118496

42. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016; 44: D7–D19. https://doi.org/10.1093/nar/gkv1290 PMID: 26615191

43. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015; 16: 1–14. https://doi.org/10.1186/s13059-014-0572-2 PMID: 25583448

44. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008; 11: 472–477. https://doi.org/10.1016/j.mib.2008.09.006 PMID: 19086349

45. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol. 2017; 34: 2115–2122. https://doi.org/10.1093/molbev/msx148 PMID: 28460117

46. Cosentino S, Voldby Larsen M, Møller Aarestrup F, Lund O. PathogenFinder—Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. PLoS One. 2013; 8: e77302. https://doi.org/10.1371/journal.pone.0077302 PMID: 24204795

47. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22: 1658–1659. https://doi.org/10.1093/bioinformatics/btl158 PMID: 16731699

48. Contzen M, Sting R, Blazey B, Rau J. *Corynebacterium ulcerans* from Diseased Wild Boars. Zoonoses Public Health. 2011; 58: 479–488. https://doi.org/10.1111/j.1863-2378.2011.01396.x PMID: 21824349

49. Subedi R, Kolodkina V, Sutcliffe IC, Simpson-Louredo L, Hirata R, Titov L, et al. Genomic analyses reveal two distinct lineages of *Corynebacterium ulcerans* strains. New Microbes New Infect. 2018; 25: 7–13. https://doi.org/10.1016/j.nmni.2018.05.005 PMID: 29997890

50. Meinel DM, Margos G, Konrad R, Krebs S, Blum H, Sing A. Next generation sequencing analysis of nine *Corynebacterium ulcerans* isolates reveals zoonotic transmission and a novel putative diphtheria toxin-encoding pathogenicity island. Genome Med. 2014; 6: 113. https://doi.org/10.1186/s13073-014-0113-3 PMID: 25587356

51. Tenor JL, McCormick BA, Ausubel FM, Aballay A. *Caenorhabditis elegans*-Based Screen Identifies Salmonella Virulence Factors Required for Conserved Host-Pathogen Interactions. Curr Biol. 2004; 14: 1018–1024. https://doi.org/10.1016/j.cub.2004.05.050 PMID: 15182677

52. Liu H, Dang G, Zang X, Cai Z, Cui Z, Song N, et al. Characterization and pathogenicity of extracellular serine protease MAP3292c from *Mycobacterium avium* subsp. *paratuberculosis*. Microb Pathog. 2020; 142: 104055. https://doi.org/10.1016/j.micpath.2020.104055 PMID: 32058021

53. Ramirez NA, Das A, Ton-That H. New Paradigms of Pilus Assembly Mechanisms in Gram-Positive Actinobacteria. Trends Microbiol. 2020; 1–11. https://doi.org/10.1016/j.tim.2020.05.008 PMID: 32499101

54. Mandlik A, Swierczynski A, Das A, Ton-That H. Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. Trends Microbiol. 2008; 16: 33–40. https://doi.org/10.1016/j.tim.2007.10.010 PMID: 18083568

55. Mandlik A, Swierczynski A, Das A, Ton-That H. *Corynebacterium diphtheriae* employs specific minor pilins to target human pharyngeal epithelial cells. Mol Microbiol. 2007; 64: 111–124. https://doi.org/10.1111/j.1365-2958.2007.05630.x PMID: 17376076

56. Chang C, Mandlik A, Das A, Ton-That H. Cell surface display of minor pilin adhesins in the form of a simple heterodimeric assembly in *Corynebacterium diphtheriae*. Mol Microbiol. 2011; 79: 1236–1247. https://doi.org/10.1111/j.1365-2958.2010.07515.x PMID: 21205008

57. Loenen WAM, Dryden DTF, Raleigh EA, Wilson GG. Type I restriction enzymes and their relatives. Nucleic Acids Res. 2014; 42: 20–44. https://doi.org/10.1093/nar/gkt847 PMID: 24068554

58. Pacheco LGC, Pena RR, Castro TLP, Dorella F A., Bahia RC, Carminati R, et al. Multiplex PCR assay for identification of *Corynebacterium pseudotuberculosis* from pure cultures and for rapid detection of this pathogen in clinical samples. J Med Microbiol. 2007; 56: 480–486. https://doi.org/10.1099/jmm.0.46997-0 PMID: 17374887

59. Almeida S, Dorneles EMS, Diniz C, Abreu V, Sousa C, Alves J, et al. Quadruplex PCR assay for identification of *Corynebacterium pseudotuberculosis* differentiating biovar Ovis and Equi. BMC Vet Res. 2017; 13: 1–8. https://doi.org/10.1186/s12917-016-0931-1 PMID: 28049469

60. Badell E, Guillot S, Tulliez M, Pascal M, Panunzi LG, Rose S, et al. Improved quadruplex real-time PCR assay for the diagnosis of diphtheria. J Med Microbiol. 2019; 68: 1455–1465. https://doi.org/10.1099/jmm.0.001070 PMID: 31478826

4.2 Artigo 2 – Analysis of *Corynebacterium silvaticum* genomes from Portugal reveals a single cluster and a clade suggested to produce diphtheria toxin

Marcus Vinicius Canário Viana, José Henrique Souza Galdino Brandão, Rodrigo Profeta, Manuela Oliveira, Luís Tavares, Siomar de Castro Soares, Paulo Luiz Souza Carneiro Alice Rebecca Wattam, Vasco Azevedo. **Em revisão pela revista PeerJ.**

Neste capítulo investiguei oito genomas de linhagens isoladas de Portugal, o pangenoma da espécie, e a relação entre filogenia e a possível produção de DT.

Important declarations

Please remove this info from manuscript text if it is also present there.

Associated Data

Data supplied by the author:
The scripts used for analysis of OrthoFinder output are available as Supplemental Data S2.

Required Statements

Competing Interest statement:
Vasco Azevedo is an Academic Editor for PeerJ.

Funding statement:
This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and Universidade Federal de Minas Gerais (UFMG).

Analysis of *Corynebacterium silvaticum* genomes from Portugal reveals a single cluster and a clade suggested to produce diphtheria toxin

Marcus Vinicius Canario Viana [1], **José Henrique Galdino** [2], **Rodrigo Profeta** [1], **Manuela Oliveira** [3], **Luís Tavares** [3], **Siomar de Castro Soares** [4], **Paulo Carneiro** [2], **Alice Rebecca Wattam** [5], **Vasco Azevedo** [Corresp. 1]

[1] Department of Genetics, Ecology and Evolution, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[2] Department of Biological Sciences, State University of Southwest of Bahia, Jequié, Bahia, Brazil

[3] Centre for Interdisciplinary Research in Animal Health, Faculty of Veterinary Medicine, University of Lisbon, Lisbon, Portugal

[4] Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil

[5] Biocomplexity Institute, University of Virginia, Charlottesville, Virginia, United States

Corresponding Author: Vasco Azevedo
Email address: vasco@icb.ufmg.br

Background. *Corynebacterium silvaticum* is a pathogenic, gram-positive bacterial species that causes Caseous Lymphadenitis in wild boars, domestic pigs and roe deer in Western Europe. It can affect animal production and cause zoonosis. Genome analysis has suggested that one strain from Portugal and one from Austria could probably produce diphtheria toxin (DT), which inhibits protein synthesis and can cause death. **Methods.** To further investigate the species genetic diversity and probable production of DT by Portuguese strains, eight isolates from this country were sequenced and compared to 38 public ones. **Results.** Strains from Portugal are monophyletic, nearly identical, form a unique cluster and have 27 out of 36 known *Corynebacterium* virulence or niche factors. All of them lack a frameshift in the *tox* gene and were suggested to produce DT. A phylogenetic analysis shows that the species has diverged into two clades. Clade 1 is composed of strains that were suggested to have the ability to produce DT, represented by the monophyletic strains from Portugal and strain 05-13 from Austria. Clade 2 is composed of strains unable to produce DT due to a frameshifted *tox* gene. The second clade is represented by strains from Austria, Germany and Switzerland. Ten genome clusters were detected, in which strains from Germany are the most diverse. Strains from Portugal belong to an exclusive cluster. The pangenome has 2,961 proteins and is nearly closed ($\alpha = 0.968$). Exclusive genes shared by Cluster 1 and 2, and Portuguese strains are probably not related to disease manifestation as they share the same host but could play a role in their extra-host environmental adaptation . These results show the potential of the species to cause zoonosis, possibly diphtheria . The identified clusters, exclusively shaded genes,

and exclusive STs identified in Portugal could be applied in the identification and epidemiology of the species.

Analysis of *Corynebacterium silvaticum* genomes from Portugal reveals a single cluster and a clade suggested to produce diphtheria toxin

Marcus Vinicius Canário Viana[1], José Henrique Souza Galdino Brandão[2], Rodrigo Profeta[1], Manuela Oliveira[3], Luís Tavares[3], Siomar de Castro Soares[4], Paulo Luiz Souza Carneiro[2], Alice Rebecca Wattam[5], Vasco Azevedo[1]*

[1] Department of Genetics, Ecology and Evolution, Institute of Biological Sciences, Federal University of Minas Gerais, Minas Gerais, Brazil
[2] Department of Biological Sciences, State University of Southwest of Bahia, Bahia, Brazil
[3] Centre for Interdisciplinary Research in Animal Health, Faculty of Veterinary Medicine, University of Lisbon, Portugal
[4] Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil
[5] Biocomplexity Institute, University of Virginia, Virginia, USA

Corresponding Author:
Vasco Azevedo [1]
Av. Pres. Antônio Carlos, Belo Horizonte, Minas Gerais, 31270-901, Brazil
Email address: vasco@icb.ufmg.br

Abstract

Background. *Corynebacterium silvaticum* is a pathogenic, gram-positive bacterial species that causes Caseous Lymphadenitis in wild boars, domestic pigs and roe deer in Western Europe. It can affect animal production and cause zoonosis. Genome analysis has suggested that one strain from Portugal and one from Austria could probably produce diphtheria toxin (DT), which inhibits protein synthesis and can cause death.

Methods. To further investigate the species genetic diversity and probable production of DT by Portuguese strains, eight isolates from this country were sequenced and compared to 38 public ones.

Results. Strains from Portugal are monophyletic, nearly identical, form a unique cluster and have 27 out of 36 known *Corynebacterium* virulence or niche factors. All of them lack a frameshift in the *tox* gene and were suggested to produce DT. A phylogenetic analysis shows that the species has diverged into two clades. Clade 1 is composed of strains that were suggested to have the ability to produce DT, represented by the monophyletic strains from Portugal and strain 05-13 from Austria. Clade 2 is composed of strains unable to produce DT due to a frameshifted *tox*

gene. The second clade is represented by strains from Austria, Germany and Switzerland. Ten genome clusters were detected, in which strains from Germany are the most diverse. Strains from Portugal belong to an exclusive cluster. The pangenome has 2,961 proteins and is nearly closed ($\alpha = 0.968$). Exclusive genes shared by Cluster 1 and 2, and Portuguese strains are probably not related to disease manifestation as they share the same host but could play a role in their extra-host environmental adaptation. These results show the potential of the species to cause zoonosis, possibly diphtheria. The identified clusters, exclusively shaded genes, and exclusive STs identified in Portugal could be applied in the identification and epidemiology of the species.

Introduction

Corynebacterium silvaticum is a species recently described gram-positive pathogenic bacteria (Dangel et al., 2020), that has been isolated from wild board and roe deer in Germany (Dangel et al., 2020; Möller et al., 2020a), Austria (GenBank: JABGCO01), Switzerland (GenBank: JAEANX01), and from domestic pigs in Portugal (Oliveira et al., 2014; Viana et al., 2020). The infection manifests in a disease similar to Caseous Lymphadenitis (CL) (Dangel et al., 2020), which is caused by *C. pseudotuberculosis* in goats and sheep (Dorella et al., 2006). Prior to an analysis and designation of a new species, strains from this species were identified as *C. pseudotuberculosis* or *C. ulcerans* (Oliveira et al., 2014; Dangel et al., 2019; Rau et al., 2019).

C. silvaticum is part of a group of six phylogenetically related pathogenic species that include *C. diphtheriae*, *C. belfantii*, *C. rouxii*, *C. ulcerans*, and *C. pseudotuberculosis*, which can produce diphtheria toxin (DT) when lysogenized by *tox*$^+$ corynephages (Bernard & Funke, 2015; Dangel et al., 2020; Badell et al., 2020). This toxin causes cell death by inactivating protein synthesis (Murphy, 2011). One of the characteristics of *C. silvaticum* is that it is non-toxigenic, yet it has the *tox* gene (NTTB) (Dangel et al., 2020), caused by a frameshift in *tox* (Viana et al., 2020). Portuguese strain PO100/5 and Austrian strain 05-13 do not have the characteristic frameshift in the *tox* gene caused by the insertion of two guanines (Möller et al., 2020a; Viana et al., 2020), suggesting that those strains are producers of DT. The isolation from domestic pigs suggests a potential for zoonotic transmission (Viana et al., 2020). Besides production of DT, cytotoxicity in human epithelial cells has recently been demonstrated (Möller et al., 2021).

The possibility of zoonotic transmission and the impact it could have on animal production implicate *C. silvaticum* as a potential threat to human health. In this work, we investigated the genetic diversity of the species by sequencing eight genomes from Portugal, with the aim of identifying genomic features that could be used for its control.

Materials & Methods

Genome assembly and taxonomy

The eight *C. silvaticum* strains from domestic pigs in Portugal used in the analysis were isolated by Oliveira et al. (Oliveira et al., 2014) (*Data S1*). At the time the strains were classified as *C. pseudotuberculosis*. The genomes were sequenced using Illumina HiSeq 2500 (Illumina,

San Diego, CA, USA) with 2 × 150 bp paired-end libraries. The quality of the sequencing reads was assessed by FastQC v0.11.9 (Andrews, 2015). Each genome was assembled using both reference-based and *de novo* assemblies. For the reference-based assembly, we used as reference the first version of *C. silvaticum* PO100/5 genome (GenBank CP021417.1, BV-BRC 65058.108,). The tool used for read mapping and extraction of consensus sequence was UGENE 39 (Okonechnikov et al., 2012), with the plugins Bowtie v2.4.2 (Langmead & Salzberg, 2012) and SAMtools v0.1.19 (Li et al., 2009). For the *de novo* assemblies, we performed three assemblies using SPAdes v3.15.3 (Bankevich et al., 2012), Unicycler v0.4.8 (Wick et al., 2017) and Edena v3.131028 (Hernandez et al., 2008). Before finishing the assembly, the taxonomy of the sample was determined using Type Strain Genome Server (Meier-Kolthoff & Göker, 2019). Then, the best *de novo* assembly was determined by QUAST v5.1.0rc1 (Gurevich et al., 2013). This assembly was scaffolded using CONTIGuator v2 (Galardini et al., 2011) using CP021417.1 as reference. The beginning of the chromosome was moved to the *dnaA* gene using the script moveDNAA.py (https://github.com/dcbmariano/scripts/blob/master/moveDNAA.py). The gaps were automatically closed using the contigs of the other three assemblies using GFinisher v1.4 (Guizelini et al., 2016) with CP021417.1 as reference. The assembly completeness and contamination were evaluated using CheckM 2 (https://github.com/chklovski/CheckM2).

For comparison of *tox*$^+$ prophages, we reassembled the public genomes of strains 05-13 (GenBank SRA accession number SRR11485666) and KL0182^T (SRR7825394) (*Data S1*). The raw sequencing data was retrieved using fastq-dump from SRA Toolkit (https://github.com/ncbi/sratoolkit) and the assembly was performed with the method used in the strains from Portugal but replacing Edena's assembly for the respective public assembly.

Clustering, typing and annotation

Genome clusters were determined using PopPUNK v2.6.0 (Lees et al., 2019) and the network was visualized using Cytoscape v3.9.1 (Shannon et al., 2003). The sequence type (ST) of the strains was determined using MLST v 2.0.4 (Larsen et al., 2012). The genomes were annotated using the Rapid Annotation using Subsystems Technology (RASTtk) pipeline (Brettin et al., 2015), implemented in the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) (Olson et al., 2022), and submitted to GenBank (Benson et al., 2012).

Characterization of *C. silvaticum* genomes from Portugal

Plasmids, insertion sequences and prophages were predicted using PlasmidFinder v2.1 (Carattoli et al., 2014), ISEScan v1.7.2.3 (Xie & Tang, 2017) and PHASTER (Arndt et al., 2016), respectively. Genomics islands were predicted for strain PO100/5 using GIPSy v1.1.3 (Soares et al., 2016), with *C. glutamicum* ATCC13032 (CP025533.1) as a non-pathogenic reference. CRISPR-Cas systems were identified using CRISPRCasFinder (Couvin et al., 2018). Virulence factors were predicted with Abricate v1.0.1 https://github.com/tseemann/abricate), with minimum identity and coverage values of 60%. Virulence and niche factors were identified using BV-BRC's Proteome Comparison Tool, using a list of 37 genes described for *C.*

silvaticum, *C. ulcerans*, *C. pseudotuberculosis*, *C. diphtheriae*, *C. jeikeium* and *C. glutamicum* (Trost et al., 2010, 2011; Tauch & Burkovski, 2015; Reardon-Robinson et al., 2015; Weerasekera et al., 2019; Möller et al., 2020b). A circular map of PO100/5 was generated using BRIG v0.95 (Alikhan et al., 2011).

The *tox* gene sequence was compared across *Corynebacterium* species to look for the frameshift described in *C. silvaticum* (Dangel et al., 2020). The representatives for *C. silvaticum* included the KL0182^T and W25 strains from Germany, 5182 from Switzerland, 04-13 and 05-13 from Austria, and the eight strains from Portugal. Three other species were included in the comparison: *C. ulcerans* 0102 (AP012284.1), *C. pseudotuberculosis* 31 (CP003421.4) and *C. diphtheriae* NCTC13129 (NC_002935). The sequences were aligned using Jalview v2.11.1.4 (Waterhouse et al., 2009), with the MUSCLE algorithm (Edgar, 2004).

Comparative genomics with other strains

Thirty-eight public *C. silvaticum* genomes were obtained from BV-BRC and GenBank (*Data S1*) for a total of 46 when the Portuguese genomes were included. For samples available as sequencing reads (*Data S1*), we used the assemblies performed by Viana et al. (Viana et al., 2020). A phylogenomic tree of *C. silvaticum* was built using BV-BRC's Phylogenetic Tree Building tool, using the nucleotide and amino acid sequences from 1,000 shared genes. *C. ulcerans* NCTC 7910^T was used as an external group. Average Nucleotide Identity (ANI) was calculated using FastANI v1.0 (Jain et al., 2018).

The distribution of orthologous gene groups across all genomes from all three species was estimated using OrthoFinder v2.5.4 (Emms & Kelly, 2019) and in-house scripts (*Data S2*). Here, the core genome is defined by orthogroups shared by all genomes, accessory genome is defined by orthogroups shared by more than one but not all genomes, and singletons are exclusive genes of a single genome. A pangenome is the entire repertoire of orthogroups found across all genomes. We used an in-house script to identify subsets of gene groups exclusively shared by 1) *C. silvaticum* strains from Portugal, 2) *C. silvaticum* strains from Portugal and strain 05-13 from Austria, 3) the remaining *C. silvaticum* strains.

The prophages of strains PO100/5, 05-13 and KL0182 were compared using tBLASTx v2.9.0+ or BLASTn v2.9.0+ (Camacho et al., 2009) and visualized using Artemis Comparison Tool (ACT) v18.1.0 (Carver et al., 2008, 2012). Possible misassemblies were investigated by mapping sequencing reads to the assembled genome or a reference using UGENE. The reassembled version and KL0182^T was also used for genomic island prediction using GISPy to represent strains out of Portugal.

Results

Characterization of *C. silvaticum* genomes from Portugal

All strains from Portugal were identified as *C. silvaticum* (*Data S3*). The assemblies were estimated to be 99.9% complete with 0.19 or 0.2 % contamination. No plasmids were detected. The genome sizes were ~2.573 Mb, with 2,631 to 2,639 CDSs, 12 rRNA genes and 52 tRNA

genes (*Table 1*). All genomes have three insertion sequences families (*Table 1, Data S4*), two complete and one to two incomplete prophages (*Table 1, Data S5*), and a Type I-E CRISPR-Cas system (*Table 1, Data S6*). The ANI values ranged from 99.9948 to 99.9998 % (*Data S7*). The *tox*$^+$ prophage was ~38 Kb in all strains (*Data S5*). PO100/5 has 35 genomic islands in comparison to *C. glutamicum* ATCC13032 and one in comparison to KL0182^T (*Fig. 1, Data S8*). KL0182^T has 35 islands when compared to *C. glutamicum* (*Data S8*). Some islands are found in both PO100/5 and KL0182, and some are unique to each strain (Data S8). The new ST 795 was found in PO104/5, differing from the ST 578 by the new allele 65 of the gene *atpA* (*Table 1, Data S9*). Abricate identified virulence genes in four of the eight genomes (*tox*, *relA*, *ideR* and *ureB*) (*Table 2, Data S10*). The proteome comparison tool identified 27 out of 36 known *Corynebacterium* virulence or niche factors, although the pili genes *spaCDEF* were pseudogenized (*Table 2, S11 File*). All strains from Portugal and strain 05-13 from Austria have a *tox* gene that codes a DT with 560 amino acids with identical sequence (*Data S12*). The other strains have a frameshift that is caused by the insertion of two guanines in position 44 (*Data S13*). The frameshift results in a truncated protein of 17 amino acids *(Fig. 2*).

Comparative genomics with other strains

The strains were represented by 10 clusters, with strains from Germany in seven of them, being the most diverse. All strains from Portugal were in an exclusive cluster. Within strains from Austria, 05-13 had its own cluster, while 04-13 is part of a bigger cluster that includes strains from Germany and the single one from Switzerland (*Figs. 3 and 4*). The phylogenetic tree shows two clades. Clade 1 is monophyletic and has the strains from Portugal and strain 05-13 from Austria. Clade 2 has the remaining strains from Austria, Germany, and Switzerland (*Fig. 4*). The ANI values ranged from 99.7539 to 100 % for all strains, and 99.9632 to 100 % within the same clade. The difference between clades ranged from 0.2461 to 0.1729 % (*Data S7*). The alignment of the *tox*$^+$ prophages of PO100/5, 05-13 (Clade 1) and KL0182^T (Clade 2) shows nearly identical sequences for PO100/5 and 05-13, with a ~38 Kb insertion and an additional 5.8 Kb sequence upstream the *tox* gene in comparison to KL0182^T (Wild boar, Germany, ~32.8 Kb) (*Fig. 5*). The additional 5.8 Kb is a repeat and is also present in KL0182^T but in GI 18 rather than in *tox*$^+$ prophage (*Fig. 6*). Mapping the sequencing reads of KL0182^T to its assembled genome and to PO100/5 genome confirmed that part of the prophage sequence is in another region of the genome (*Data S14*). Besides the insertion, coding sequences can be fragmented or fused when the *tox*$^+$ prophage from PO100/5 and KL0182^T are compared (*Fig. 5*).

Across the 46 genomes, the pangenome, core genome, shared genome and singletons were 2,961, 2,227 (75.21 %), 623 (21.04 %) and 111 (3.75 %) orthogroups, respectively (*Data S15*). The value of α was 0.968 (*Data S16*). The subsets of exclusively shared orthogroups were: 19 in *C. silvaticum* strains from Portugal, 25 in *C. silvaticum* strains from Portugal and strain 05-13 from Austria, and 36 from the remaining strains (*Data S17 and S18*). Of the 27 *Corynebacterium* virulence/niche factors described in the literature and identified in strains from

Portugal (*Table 2, S11 File*), 18 are part of the core genome and 8 are part of the shared genome (*Data S15*).

Discussion

Strains from Portugal and production of diphtheria toxin

All the Portuguese strains are monophyletic (*Fig. 4*), from the same cluster (*Fig. 3*), and nearly identical (*Data S7*), suggesting that they were derived from a single clone. The fact that the Portuguese strains have a most recent common ancestor with Austrian strain 05-13 suggests that the initial infection originated in Austria.

Eight virulence and niche factors identified in strains from Portugal are in the accessory genome (*Data S14*). *spaDEF* and *srtC* are part of a pilus cluster and are fragmented in different genomes. The absence of the *tox* gene in some strains is certainly an assembly artifact, since the strains lacking it were shown to have it (Dangel et al., 2019). The absence of sialidase (*nanH*) and serine protease (cpfrc_00397) in some strains could also be a sequence or assembly artifact, since they were detected in 45 out of 46 strains. The venom serine protease (*vsp1*) was found in all strains from Portugal and KL1008 from Germany, but the advantages for host colonization in comparison to other strains that lack the protein must be elucidated. Among the others, the niche factors Trypsin-like serine protease (Uniprot A0A5C5F2T7), *cwlH* (A0A5C5F4U0) and *rfpI* (A0A5F0A739) are part of the core genome (*Data S14*) and shared by all of the strains examined. These three proteins have been shown to be the most abundant extracellular proteins produced *in vitro* by *C. silvaticum* W25, representing 88.1, 2.2 and 1.3 %, respectively (Möller et al., 2020b).

C. silvaticum was described as non-toxigenic *tox* gene-bearing (Dangel et al., 2020). An insertion of two guanines in *tox* caused a frameshift and the pseudogenization of this gene (Möller et al., 2020a). However, the strains PO100/5 from Portugal and 05-13 lack the insertion and therefore could produce the toxin (Viana et al., 2020). We confirmed that all eight strains from Portugal lack that insertion that knocks out this gene (*Fig. 2, Data S13*) and are probable DT producers. The production of DT in strains 4-13 and 05-13 from Austria was inferred by RT-qPCR (Schaeffer et al., 2020), targeting subunits A and B of the DT (Mothershed et al., 2002). However, strain 04-13 has the typical frameshift in *tox*. The primers binding sites start from position 312, while the frameshift occurs in position 44 (*Fig. 2, Data S13*), so the frameshift cannot not be detected. It was suggested that DT production must be confirmed on *tox*-positive isolates by an Elek test, due to the description of *tox*-positive and Elek-negative strains (Schuhegger et al., 2008; Berger et al., 2014). This shows a limitation of the current RT-qPCR for detection of DT, and that 04-13 probably does not produce the toxin.

DT is a virulence factor because it only works inside of host cells and damages them (Tauch & Burkovski, 2015), but that may not impact *C. silvaticum*'s host range in the wild. The species causes CLA in its known hosts (Dangel et al., 2020), a disease manifestation like the non-DT-producing *C. pseudotuberculosis* biovar ovis causes in goats and sheep (Dorella et al.,

2006). CLA is related to the toxin Phospholipase D (*pld*) (Dorella et al., 2006) that is also produced by *C. silvaticum* (Dangel et al., 2020). Another possibility is that the DT is required only for the infection of as yet unidentified hosts. It has been hypothesized that the production of DT is required for *C. pseudotuberculosis* biovar equi to infect buffalo (Viana et al., 2017). If DT production is not required, this could relax the selective pressure to keep a functional *tox* gene and could lead to accumulation mutations that would result in the loss of function (*Figs. 2 and 5*) seen in some populations of Germany, Austria and Switzerland.

Genome diversity of *C. silvaticum*

The phylogenetic tree showed that strains from Portugal and 05-13 from Austria formed Clade 1 and the remaining strains formed Clade 2 (*Fig. 4*). The two strains from Austria (04-13 and 05-13) are in different clades, suggesting different geographical origins. The strains from Portugal are in the same clade and nearly identical to 05-13, which suggests an Austrian origin.

Although the genomes had high ANI values, of at least 99.7 % (*Data S7*), we could identify 10 clusters (*Figs. 3 and 4*). Strains from Germany had higher diversity, with strains in seven clusters, probably due to the higher number of samples and isolation from wild animals. All strains from Portugal had an exclusive cluster. As they were isolated from domestic pigs in two farms (Oliveira et al., 2014) this represents the spread of only one clone. More strains must be analyzed to assess the diversity of the species.

Across the 46 genomes, we identified a pangenome of 2,961 orthogroups, core genome of 2,227, 623 shared orthogroups, 111 singletons and an α of 0.968 (*Data S5* and *S15*). The pangenome is nearly closed, which agrees with the high ANI values between the genomes. We identified exclusively shared orthogroups in strains from Portugal (19 orthogroups), Clade 1 (27 orthogroups), and Clade 2 (36 orthogroups) (*Data S17 and S18*). Those genes are probably not involved in the manifestation of the disease as it is the same manifestation (CL) in the known hosts but could be required for survival outside of the host.

With the presence of *C. silvaticum* in wild and domestic animals, cytotoxicity to human epithelial cells (Möller et al., 2021) and the probable production of DT, this species has the potential to cause zoonosis and diphtheria. Human transmission could occur via occupational exposure, as seen in *C. pseudotuberculosis* (Dorella et al., 2006). In this context, we identified 10 clusters (*Figs. 3*), exclusively shared genes of clades and Portuguese strains (*Data S17 and S18*), and the exclusive STs from Portugal (*Data S9*). This information can be applied to the identification and epidemiology of *C. silvaticum*.

Conclusions

In *C. silvaticum*, Clade 1 includes strains that have the potential to produce DT, which is missing in Clade 2 (non-DT-producing). Both clades can be identified by genes that, while probably not important for their interactions within the host environment, could play a role in survival in the environment. Portuguese strains are monophyletic, nearly identical, form a unique cluster and probably produce DT. We showed that the species has the potential to cause zoonosis

and diphtheria, and genome clusters, STs and exclusive genes that can be applied to its epidemiology.

References

Alikhan N-F, Petty NK, ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. DOI: 10.1186/1471-2164-12-402.

Andrews S. 2015.FASTQC A Quality Control tool for High Throughput Sequence Data. *Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/*

Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research* 44:W16–W21. DOI: 10.1093/nar/gkw387.

Badell E, Hennart M, Rodrigues C, Passet V, Dazas M, Panunzi L, Bouchez V, Carmi–Leroy A, Toubiana J, Brisse S. 2020. *Corynebacterium rouxii* sp. nov., a novel member of the diphtheriae species complex. *Research in Microbiology* 171:122–127. DOI: 10.1016/j.resmic.2020.02.003.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A v., Sirotkin A v., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19:455–477. DOI: 10.1089/cmb.2012.0021.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. *Nucleic Acids Research* 41. DOI: 10.1093/nar/gks1195.

Berger A, Hogardt M, Konrad R, Sing A. 2014. Detection Methods for Laboratory Diagnosis of Diphtheria. In: *Corynebacterium diphtheriae and Related Toxigenic Species*. Dordrecht: Springer Netherlands,. DOI: 10.1007/978-94-007-7624-1_9.

Bernard AL, Funke G. 2015. *Corynebacterium*. In: *Bergey's Manual of Systematic of Archaea and Bacteria (Online)*. London: John Wiley & Sons, Bergey's Manual Trust, 1–70. DOI: 10.1002/9781118960608.

Brettin T, Davis JJ, Disz T, Edwards R a, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason J a, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports* 5:1–6. DOI: 10.1038/srep08365.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. DOI: 10.1186/1471-2105-10-421.

Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. In Silico Detection and Typing of Plasmids using

PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrobial Agents and Chemotherapy* 58:3895–3903. DOI: 10.1128/AAC.02412-14.

Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream MA. 2008. Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* 24:2672–2676. DOI: 10.1093/bioinformatics/btn529.

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics (Oxford, England)* 28:464–469. DOI: 10.1093/bioinformatics/btr703.

Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* 46:W246–W251. DOI: 10.1093/NAR/GKY425.

Dangel A, Berger A, Konrad R, Sing A. 2019. NGS-based phylogeny of diphtheria-related pathogenicity factors in different *Corynebacterium* spp. implies species-specific virulence transmission. *BMC Microbiology* 19:1–16. DOI: 10.1186/s12866-019-1402-1.

Dangel A, Berger A, Rau J, Eisenberg T, Kämpfer P, Margos G, Contzen M, Busse H-J, Konrad R, Peters M, Sting R, Sing A. 2020. *Corynebacterium silvaticum* sp. nov., a unique group of NTTB corynebacteria in wild boar and roe deer. *International Journal of Systematic and Evolutionary Microbiology*. DOI: 10.1099/ijsem.0.004195.

Dorella FA, Carvalho Pacheco L, Oliveira SC, Miyoshi A, Azevedo V. 2006. *Corynebacterium pseudotuberculosis* : microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Veterinary Research* 37:201–218. DOI: 10.1051/vetres:2005056.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792–1797. DOI: 10.1093/nar/gkh340.

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology* 20:1–14. DOI: 10.1186/s13059-019-1832-y.

Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source code for biology and medicine* 6:11. DOI: 10.1186/1751-0473-6-11.

Guizelini D, Raittz RT, Cruz LM, Souza EM, Steffens MBR, Pedrosa FO. 2016. GFinisher: a new strategy to refine and finish bacterial genome assemblies. *Scientific Reports* 6:34963. DOI: 10.1038/srep34963.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. DOI: 10.1093/bioinformatics/btt086.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. 2008. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Research* 18:802–809. DOI: 10.1101/gr.072033.107.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* 9:5114. DOI: 10.1038/s41467-018-07641-9.

359 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*
360 9:357–359. DOI: 10.1038/nmeth.1923.
361 Larsen M v., Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-
362 Pontén T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus Sequence Typing of Total-
363 Genome-Sequenced Bacteria. *Journal of Clinical Microbiology* 50. DOI:
364 10.1128/JCM.06094-11.
365 Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD,
366 Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK.
367 *Genome Research* 29:304–316. DOI: 10.1101/gr.241455.118.
368 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.
369 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
370 DOI: 10.1093/bioinformatics/btp352.
371 Meier-Kolthoff JP, Göker M. 2019. TYGS is an automated high-throughput platform for state-
372 of-the-art genome-based taxonomy. *Nature Communications* 10:2182. DOI:
373 10.1038/s41467-019-10210-3.
374 Möller J, Busch A, Berens C, Hotzel H, Burkovski A. 2021. Newly Isolated Animal Pathogen
375 *Corynebacterium silvaticum* Is Cytotoxic to Human Epithelial Cells. *International Journal*
376 *of Molecular Sciences* 22. DOI: 10.3390/ijms22073549.
377 Möller J, Musella L, Melnikov V, Geißdörfer W, Burkovski A, Sangal V. 2020a. Phylogenomic
378 characterisation of a novel corynebacterial species pathogenic to animals. *Antonie van*
379 *Leeuwenhoek* 113:1225–1239. DOI: 10.1007/s10482-020-01430-5.
380 Möller J, Schorlemmer S, Hofmann J, Burkovski A. 2020b. Cellular and Extracellular Proteome
381 of the Animal Pathogen *Corynebacterium silvaticum*, a Close Relative of Zoonotic
382 *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis*. *Proteomes* 8. DOI:
383 10.3390/proteomes8030019.
384 Mothershed EA, Cassiday PK, Pierson K, Mayer LW, Popovic T. 2002. Development of a Real-
385 Time Fluorescence PCR Assay for Rapid Detection of the Diphtheria Toxin Gene. *Journal*
386 *of Clinical Microbiology* 40. DOI: 10.1128/JCM.40.12.4713-4719.2002.
387 Murphy JR. 2011. Mechanism of Diphtheria Toxin Catalytic Domain Delivery to the Eukaryotic
388 Cell Cytosol and the Cellular Factors that Directly Participate in the Process. *Toxins* 3:294–
389 308. DOI: 10.3390/toxins3030294.
390 Okonechnikov K, Golosova O, Fursov M, team the U. 2012. Unipro UGENE: a unified
391 bioinformatics toolkit. *Bioinformatics* 28:1166–1167. DOI:
392 10.1093/BIOINFORMATICS/BTS091.
393 Oliveira M, Barroco C, Mottola C, Santos R, Lemsaddek A, Tavares L, Semedo-Lemsaddek T.
394 2014. First report of *Corynebacterium pseudotuberculosis* from caseous lymphadenitis
395 lesions in Black Alentejano pig (*Sus scrofa domesticus*). *BMC Veterinary Research* 10:218.
396 DOI: 10.1186/s12917-014-0218-3.
397 Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, Dempsey DM, Dickerman A,
398 Dietrich EM, Kenyon RW, Kuscuoglu M, Lefkowitz EJ, Lu J, Machi D, Macken C, Mao C,

Niewiadomska A, Nguyen M, Olsen GJ, Overbeek JC, Parrello B, Parrello V, Porter JS, Pusch GD, Shukla M, Singh I, Stewart L, Tan G, Thomas C, VanOeffelen M, Vonstein V, Wallace ZS, Warren AS, Wattam AR, Xia F, Yoo H, Zhang Y, Zmasek CM, Scheuermann RH, Stevens RL. 2022. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Research*. DOI: 10.1093/nar/gkac1003.

Rau J, Eisenberg T, Peters M, Berger A, Kutzer P, Lassnig H, Hotzel H, Sing A, Sting R, Contzen M. 2019. Reliable differentiation of a non-toxigenic tox gene-bearing *Corynebacterium ulcerans* variant frequently isolated from game animals using MALDI-TOF MS. *Veterinary Microbiology* 237:108399. DOI: 10.1016/j.vetmic.2019.108399.

Reardon-Robinson ME, Osipiuk J, Jooya N, Chang C, Joachimiak A, Das A, Ton-That H. 2015. A thiol-disulfide oxidoreductase of the Gram-positive pathogen *Corynebacterium diphtheriae* is essential for viability, pilus assembly, toxin production and virulence. *Molecular Microbiology* 98. DOI: 10.1111/mmi.13172.

Schaeffer J, Huhulescu S, Stoeger A, Allerberger F, Ruppitsch W. 2020. Draft Genome Sequences of Six *Corynebacterium ulcerans* Strains Isolated from Humans and Animals in Austria, 2013 to 2019. *Microbiology Resource Announcements* 9. DOI: 10.1128/mra.00946-20.

Schuhegger R, Lindermayer M, Kugler R, Heesemann J, Busch U, Sing A. 2008. Detection of Toxigenic *Corynebacterium diphtheriae* and *Corynebacterium ulcerans* Strains by a Novel Real-Time PCR. *Journal of Clinical Microbiology* 46. DOI: 10.1128/JCM.01010-08.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13:2498–504. DOI: 10.1101/gr.1239303.

Soares SC, Geyik H, Ramos RTJ, de Sá PHCG, Barbosa EGV, Baumbach J, Figueiredo HCP, Miyoshi A, Tauch A, Silva A, Azevedo V. 2016. GIPSy: Genomic island prediction software. *Journal of Biotechnology* 232:2–11. DOI: 10.1016/j.jbiotec.2015.09.008.

Tauch A, Burkovski A. 2015. Molecular armory or niche factors: virulence determinants of *Corynebacterium* species. *FEMS Microbiology Letters* 67:fnv185. DOI: 10.1093/femsle/fnv185.

Trost E, Al-Dilaimi A, Papavasiliou P, Schneider J, Viehoever P, Burkovski A, Soares SC, Almeida SS, Dorella F a, Miyoshi A, Azevedo V, Schneider MP, Silva A, Santos CS, Santos LS, Sabbadini P, Dias A a, Hirata R, Mattos-Guaraldi AL, Tauch A. 2011. Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. *BMC genomics* 12:383. DOI: 10.1186/1471-2164-12-383.

Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, Husemann P, Stoye J, Dorella FA, Rocha FS, Soares SDC, D'Afonseca V, Miyoshi A, Ruiz J, Silva A, Azevedo V, Burkovski A, Guiso N, Join-Lambert OF, Kayal S, Tauch A. 2010. The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl

with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC genomics* 11:728. DOI: 10.1186/1471-2164-11-728.

Viana MVC, Figueiredo H, Ramos R, Guimarães LC, Pereira FL, Dorella FA, Selim SAK, Salaheldean M, Silva A, Wattam AR, Azevedo V. 2017. Comparative genomic analysis between *Corynebacterium pseudotuberculosis* strains isolated from buffalo. *PLOS ONE* 12:e0176347. DOI: 10.1371/journal.pone.0176347.

Viana MVC, Profeta R, da Silva AL, Hurtado R, Cerqueira JC, Ribeiro BFS, Almeida MO, Morais-Rodrigues F, Soares S de C, Oliveira M, Tavares L, Figueiredo H, Wattam AR, Barh D, Ghosh P, Silva A, Azevedo V, de Castro Soares S, Oliveira M, Tavares L, Figueiredo H, Wattam AR, Barh D, Ghosh P, Silva A, Azevedo V. 2020. Taxonomic classification of strain PO100/5 shows a broader geographic distribution and genetic markers of the recently described *Corynebacterium silvaticum*. *PLOS ONE* 15:e0244210. DOI: 10.1371/journal.pone.0244210.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189–1191. DOI: 10.1093/bioinformatics/btp033.

Weerasekera D, Möller J, Kraner ME, Azevedo Antunes C, Mattos-Guaraldi AL, Burkovski A. 2019. Beyond diphtheria toxin: cytotoxic proteins of *Corynebacterium ulcerans* and *Corynebacterium diphtheriae*. *Microbiology* 165. DOI: 10.1099/mic.0.000820.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* 13:e1005595. DOI: 10.1371/journal.pcbi.1005595.

Xie Z, Tang H. 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* 33:3340–3347. DOI: 10.1093/bioinformatics/btx433.

Table 1(on next page)

Genome features of *Corynebacterium silvaticum* strains from Portugal

Strain	**PO25/4**	**PO38/4**	**PO39/4**	**PO100/5**	**PO101/5**	**PO102/5**	**PO104/5**	**PO105/5**
Genbank accession	CP080461	CP081182	CP081179	CP021417.2	CP081180	CP080459	CP080460	CP081181
Completeness (%)	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
Contamination (%)	0.2	0.19	0.2	0.19	0.19	0.19	0.2	0.2
Size (bp)	2,572,825	2,572,864	2,572,860	2,572,864	2,572,991	2,572,936	2,572,895	2,572,843
Plasmid	-	-	-	-	-	-	-	-
CG content (%)	54.40	54.40	54.40	54.40	54.40	54.40	54.40	54.40
CDS	2,631	2,633	2,636	2,633	2,636	2,635	2,634	2,639
tRNA	52	52	52	52	52	52	52	52
rRNA	12	12	12	12	12	12	12	12
Repeat region	24	24	24	24	24	24	24	24
Sequence type	578	578	578	709	578	709	795	578
Insertion sequence	IS21, IS110 and IS256	IS21, IS110 and IS257	IS21, IS110 and IS258	IS21, IS110 and IS259	IS21, IS110 and IS260	IS21, IS110 and IS261	IS21, IS110 and IS262	IS21, IS110 and IS263
Prophages	2 questionable, 2 incomplete	2 questionable, 1 incomplete	2 questionable, 2 incomplete	2 questionable, 2 incomplete	2 questionable, 1 incomplete	2 questionable, 2 incomplete	2 questionable, 2 incomplete	2 questionable, 2 incomplete
CRISPR-Cas system	Type I-E	Type I-E	Type I-E	Type I-E	Type I-E	Type I-E	Type I-E	Type I-E

1

Table 2(on next page)

Presence of niche and virulence factors of *Corynebacterium* in *C. silvaticum* strains from Portugal.

Type	Gene	Product	Presence	Reference locus tag	Reference species	Reference
Niche	-	*C. diphtheriae* DIP0733 homolog	Yes	CULC22_00609	Cul	Tauch and Burkovski 2015
Niche	-	Secreted subtilisin-like serine protease	Yes	cpfrc_00397	Cp	Trost et al. 2010
Niche	-	Secreted subtilisin-like serine protease	Yes, except in PO25/4	cpfrc_01634	Cp	Trost et al. 2010
Niche	-	Secreted trypsin-like serine protease	No	cpfrc_00562	Cp	Trost et al. 2010
Niche	-	Secreted SGNH-hydrolase	Yes	cpfrc_00536	Cp	Trost et al. 2010
Niche	*accD3*	Acyl-CoA carboxylase b-subunit involved in mycolic acid synthesis	Yes	cpfrc_01953	Cp	Trost et al. 2010
Niche	*asa*	Ceramidase	No	jk1103	Cj	Tauch et al. 2005, Tauch and Burkovski 2015
Niche	*che*	Cholesterol esterase	No	jk2054	Cj	Tauch et al. 2005, Tauch and Burkovski 2015
Niche	*choE*	Cholesterol oxidase	No	jk0629	Cj	Tauch et al. 2005, Tauch and Burkovski 2015
Niche	*cwlH*	Cell wall-associated hydrolase	Yes	CULC809_01521	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*dtsR1*	Acetyl-CoA carboxylase b-subunit involved in fatty acid synthesis	Yes	cpfrc_00492	Cp	Trost et al. 2010
Niche	*dtsR2*	Acyl-CoA carboxylase b-subunit involved in mycolic acid synthesis	Yes	cpfrc_00491	Cp	Trost et al. 2010
Niche	*endoE*	Endoglycosidase E (former corynebacterial protease CP40)	Yes	CULC809_01974	Cul	Trost et al. 2011
Niche	*mdbA*	Thiol-disulfide oxidoreductase	Yes	DIP1880	Cd	Reardon-Robinson, Osipiuk and Jooya 2015, Tauch and Burkovski 2015
Niche	*nanH*	Sialidase (neuraminidase H)	Yes	CULC809_00434	Cul	Trost et al. 2011
Niche	*nor*	Nitric oxide reductase	No	cpfrc_00128	Cp	Trost et al. 2010

Niche	*nrpS1*	Nonribosomal peptide synthetase 1	Yes	cpfrc_00565	Cp	Trost et al. 2010
Niche	*nrpS2*	Nonribosomal peptide synthetase 2	No	cpfrc_00180	Cp	Trost et al. 2010
Niche	*rhuM*	RhuM-like protein	No	CulFRC58_0285	Cul	Trost et al. 2011, Weerasekera et al. 2019
Niche	*rpfA*	Resuscitation-promoting factor A (muralytic enzyme)	Yes	cpfrc_00594	Cp	Trost et al. 2010
Niche	*rpfB*	Resuscitation-promoting factor B (muralytic enzyme)	Yes	cpfrc_00679	Cp	Trost et al. 2010
Niche	*rpfI*	Resuscitation-promoting factor-interacting protein	Yes	CULC809_01133	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*spaB*	Surface-anchored protein (minor pilus subunit)	Yes	CULC809_01980	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*spaC*	Surface-anchored protein (tip pilus protein)	Pseudogene, no CWSS	CULC809_01979	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*spaD*	Surface-anchored protein (major pilus subunit)	Pseudogene, no CWSS	CULC809_01952	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*spaE*	Surface-anchored protein (minor pilus subunit)	Pseudogene, no SP	CULC809_01950	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*spaF*	Surface-anchored protein (tip pilus protein)	Pseudogene, no SP	CULC809_01949	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*srtA*	Sortase A	Yes	CULC809_01981	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*srtB*	Sortase B	Yes	CULC809_01953	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*srtC*	Sortase C	Yes	CULC809_01951	Cul	Trost et al. 2011, Tauch and Burkovski 2015
Niche	*tspA*	Trypsin-like serine protease	No	CULC809_01848	Cul	Trost et al. 2011
Niche	*vsp1*	Venom serine protease	Yes	CULC809_00509	Cul	Trost et al. 2011
Niche	*vsp2*	Venom serine protease	Yes	CULC809_01964	Cul	Trost et al. 2011
Virulence	*pld*	Phospholipase D	Yes	ET810_03855	Cs	Trost et al. 2011, Tauch and Burkovski 2015
Virulence	*rbp*	Shiga-like ribosome-binding protein	No	CULC809_00177	Cul	Trost et al. 2011, Tauch and Burkovski 2015

Virulence	*tox*	Diphtheria toxin	Yes	DIP0222	Cd	Tauch and Burkovski 2015
Virulence	*relA*	Guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase / GTP pyrophosphokinase, (p)ppGpp synthetase II	Yes	-	-	Abricate
Virulence	*ideR*	Iron-dependent repressor IdeR/DtxR	Yes	-	-	Abricate
Virulence	*ureB*	Urease alpha subunit	Yes	-	-	Abricate
Resistance	*rpoB2*	Rifampin-resistant beta-subunit of RNA polymerase	Yes	-	-	Abricate
Resistance	*mtrA*	Two component system response regulator MtrA	Yes	-	-	Abricate
Resistance	*rbpA*	RNA-polymerase binding protein which confers resistance to rifampin	Yes	-	-	Abricate

1 CWSS - cell wall sorting signal, SP - signal peptide

Figure 1

Circular map of *Corynebacterium silvaticum* PO100/5 genome generated using BRIG v0.95.

Genomic islands and prophage detection were performed using GIPSy and PHASTER, respectively. CuI – *C. ulcerans*, GIs – genomic islands, Cg – *C. glutamicum*, Cs – *C. silvaticum*.

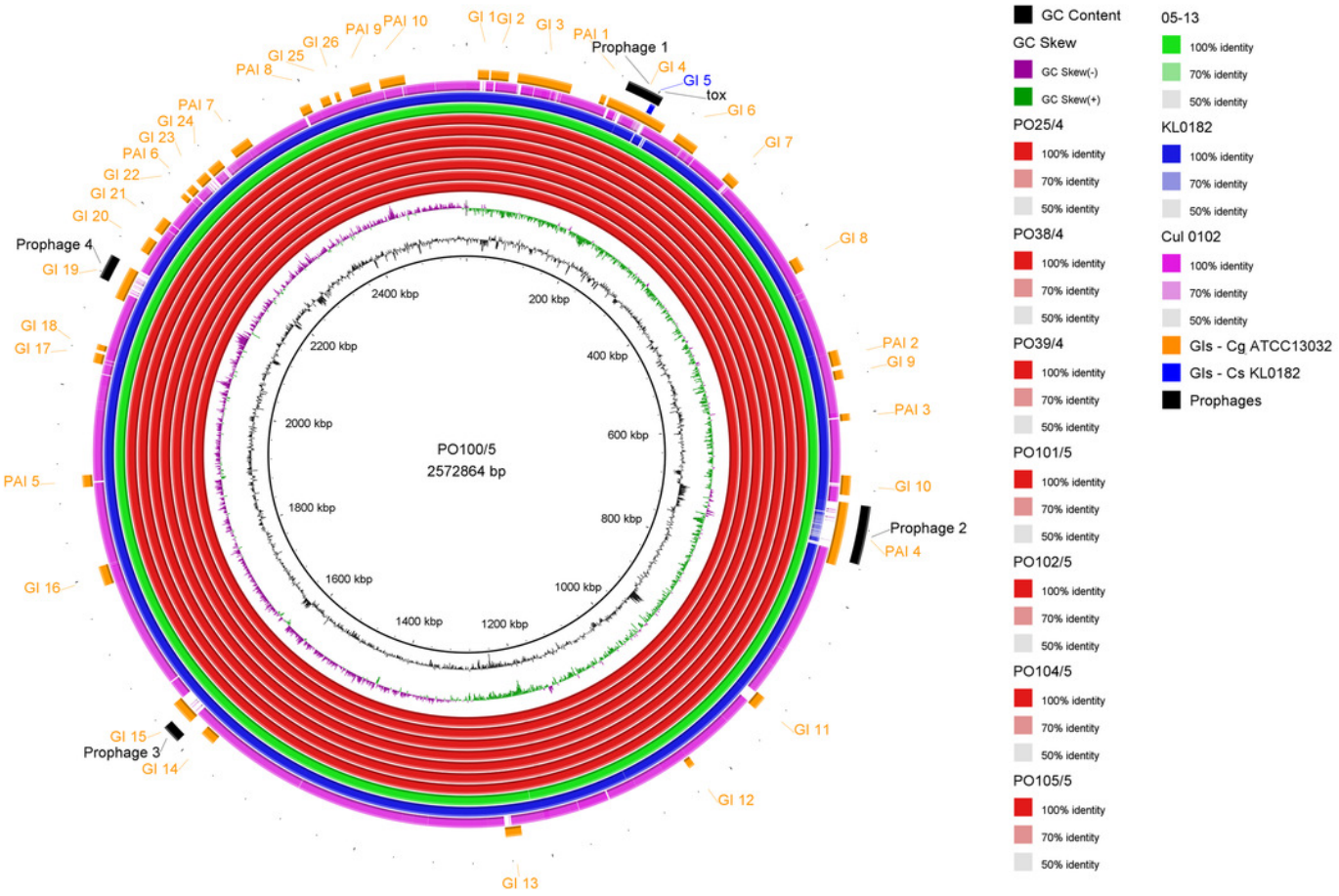


Figure 2

Alignment of the *tox* gene's first 60 nucleotides and respective amino acids of strains from *Corynebacterium silvaticum*, *C. ulcerans*, *C. pseudotuberculosis* and *C. diphtheriae*.

C. silvaticum strains from Portugal PO25/4, PO38/4, PO39/4, PO100/5, PO101/5, PO102/5, PO104/5 and PO105/5, and the strain 05-13 from Switzerland do not have the insertion of two guanines that led to a frameshift in other strains from this species. The frameshift results in a truncated protein with 17 amino acids. The alignment was performed using MUSCLE algorithm implemented in Jalview v2.11.1.4.

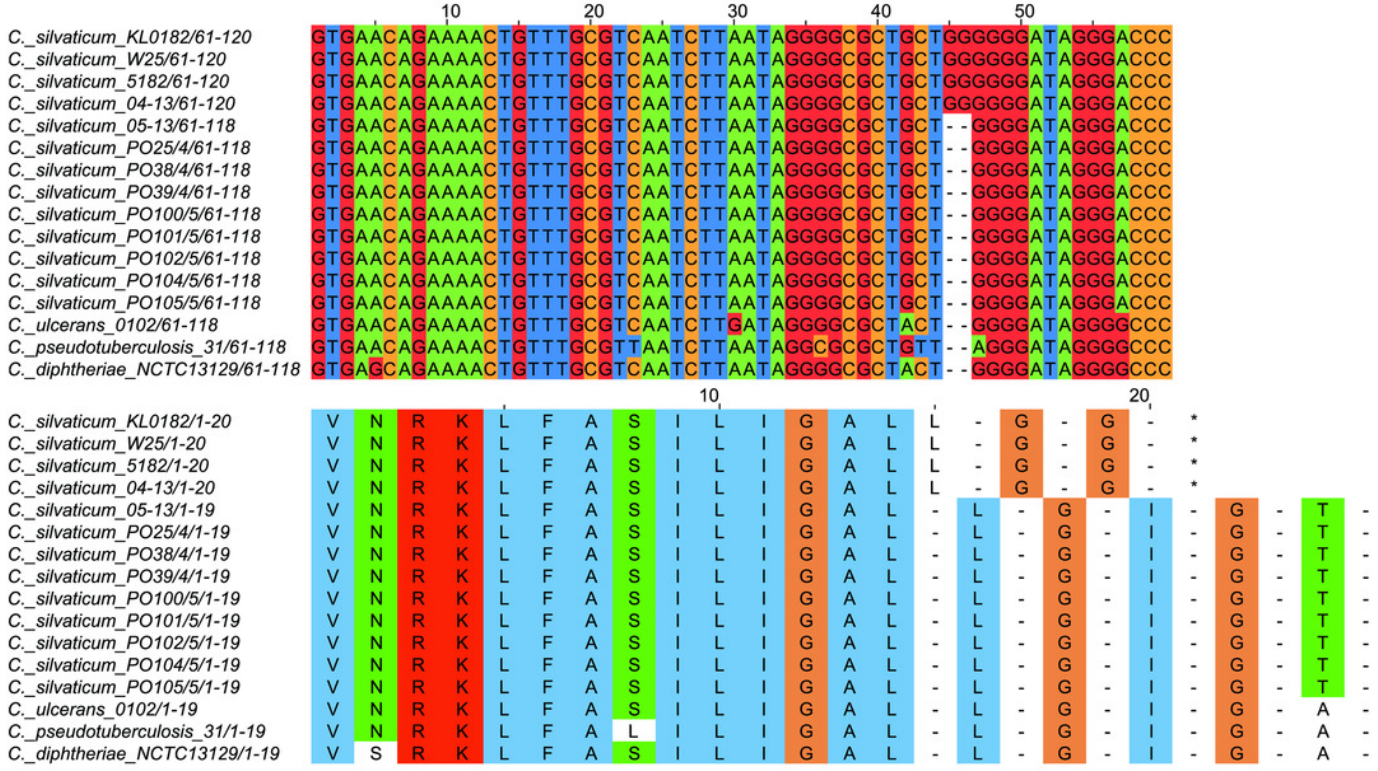


Figure 3

Clusters of *Corynebacterium silvaticum* genomes.

Countries of sample isolation: Austria (green), Germany (black), Portugal (red), Switzerland (blue).

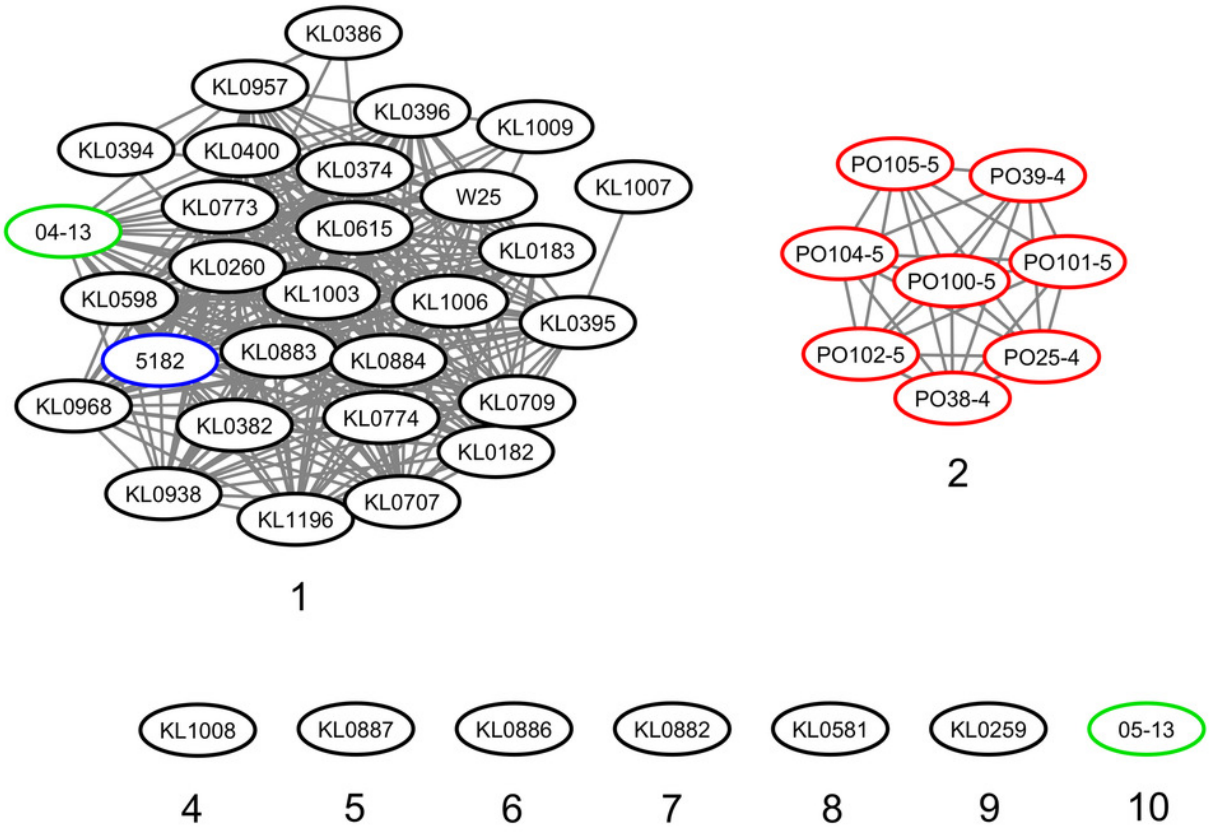


Figure 4

Phylogenomic tree of *Corynebacterium silvaticum* strains.

The phylogeny was inferred using the Phylogenetic Tree service in PATRIC, which uses RAxML with 100 rounds of “rapid bootstrapping”, and the codon sequences from 1,000 single-copy shared genes. Bootstrap values are represented by a color scale from red (70 %) to green (100 %). *Corynebacterium ulcerans* NCTC7910 was used as an outgroup.

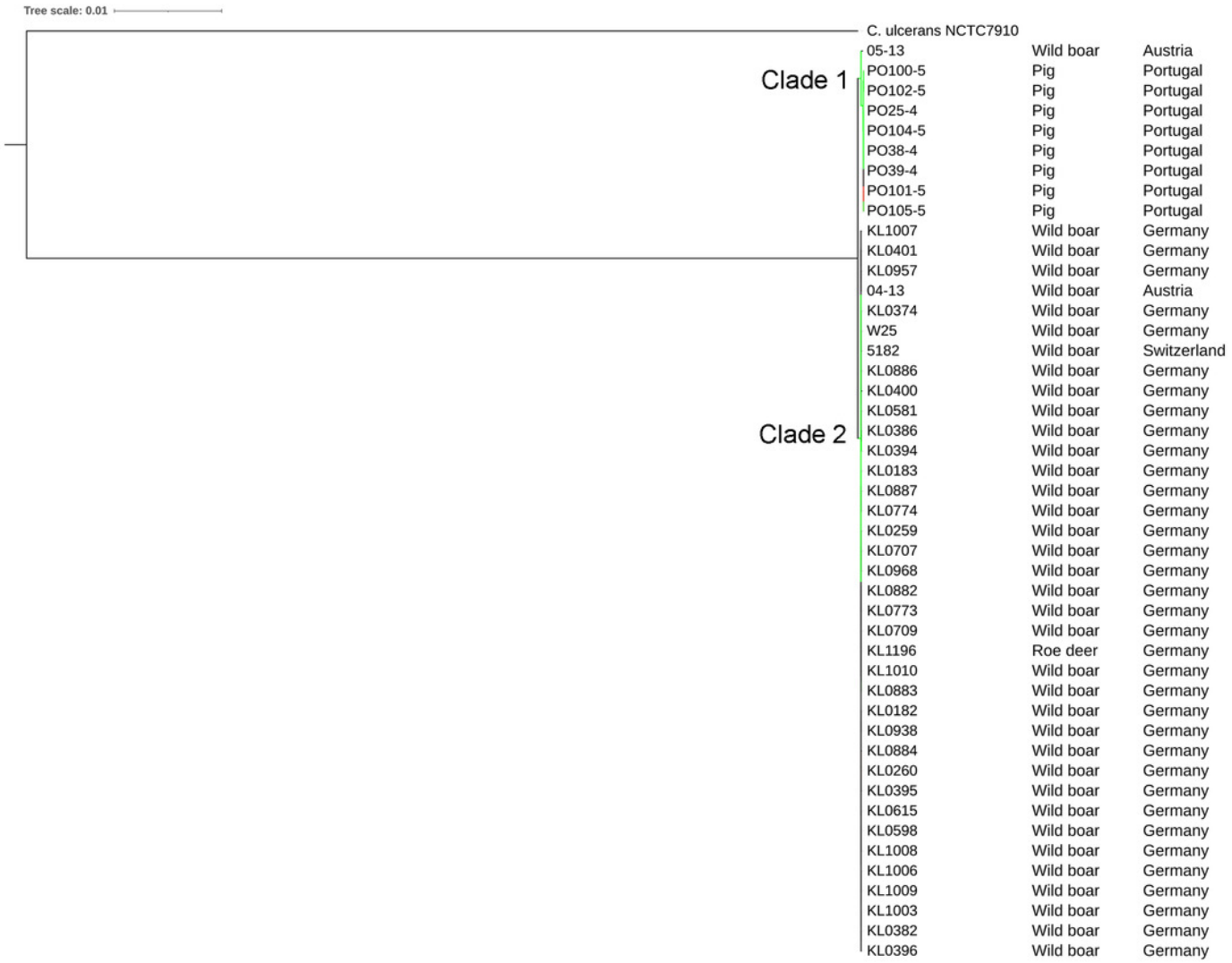


Figure 5

Alignment of *tox*$^{+}$ prophages from *Corynebacterium silvaticum* strains.

Coding sequences in dark blue are fragmented or fused in at least one genome. Coding sequences in pink are not present in strain KL0182^{T} prophage.

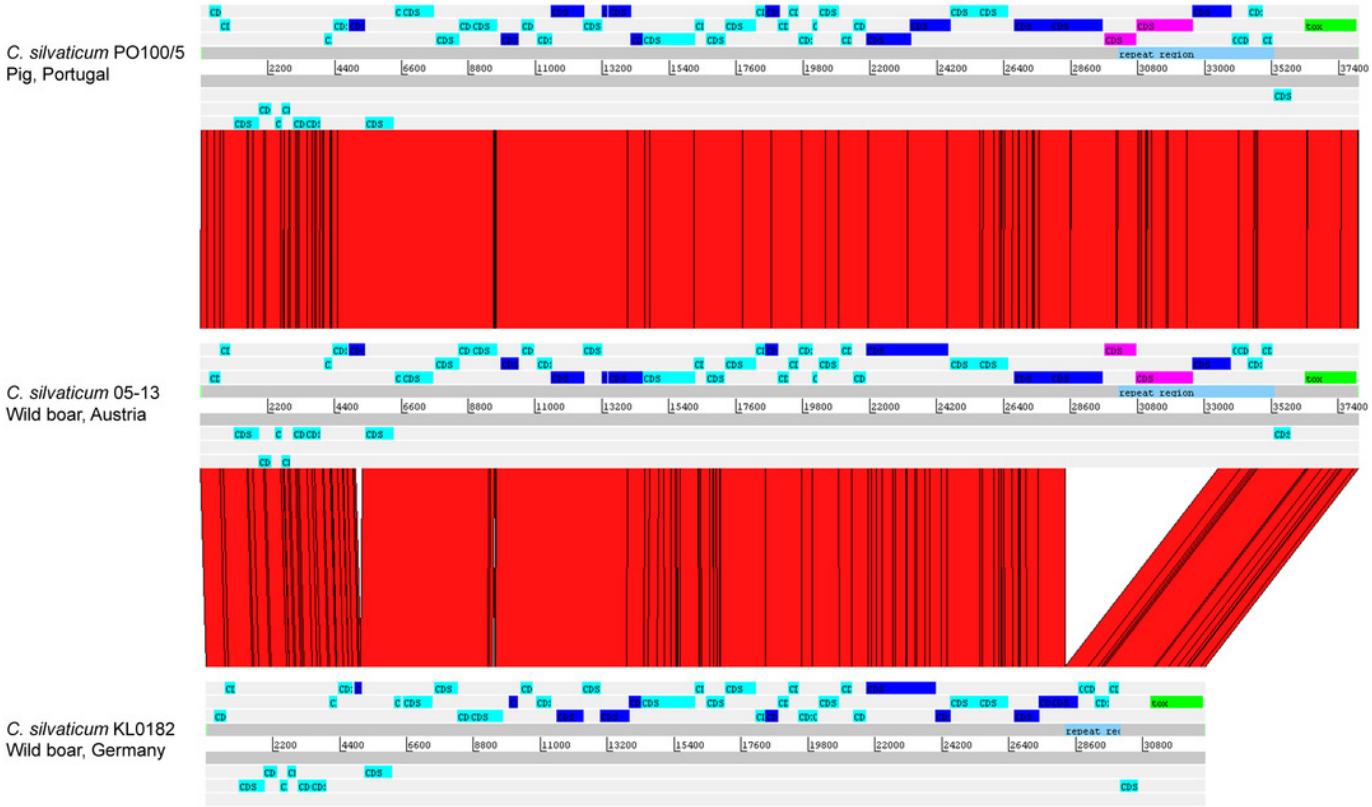
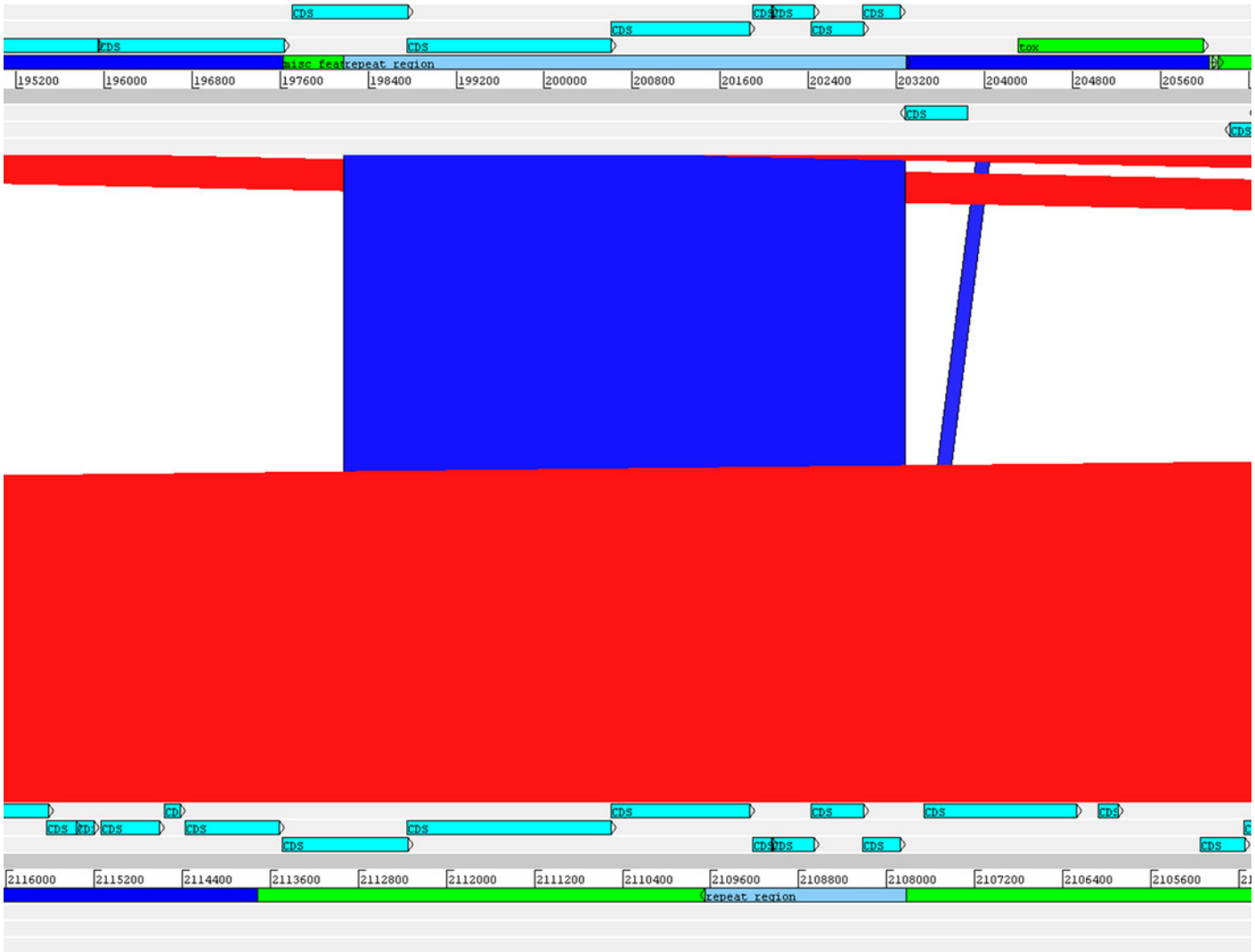


Figure 6

Alignment of part of the *tox*$^{+}$ prophage from *Corynebacterium silvaticum* PO100/5 (top) and KL0182^{T} (bottom).

The absent sequence in strain KL0182^{T} *tox*$^{+}$ prophage is part of a repeat region and is in another part of the genome , inside genomic island GI 18 (bottom, green). Dark blue highlight in chromosome – prophage, green highlight in chromosome – genomic island.



4.3 Artigo 3 – Evidence of episodic positive selection in *Corynebacterium diphtheriae* complex of species and its implementations in identification of drug and vaccine targets

Marcus Vinicius Canário Viana, Rodrigo Profeta, Janaína Canário Cerqueira, Alice Rebecca Wattam, Debmalya Barh, Artur Silva, Vasco Azevedo. **Publicado na revista PeerJ.** https://doi.org/10.7717/peerj.12662.

Neste capítulo realizei uma análise de seleção positiva para identificar adaptações exclusivas em genes codificadores de proteínas em cada uma das seis espécies do complexo da *Corynebacterium diphtheriae*: *C. belfantii*, *C. diphtheriae*, *C. pseudotuberculosis*, *C. rouxii*, *C. silvaticum* e *C. ulcerans*.

Evidence of episodic positive selection in *Corynebacterium diphtheriae* complex of species and its implementations in identification of drug and vaccine targets

Marcus Vinicius Canário Viana[1,2], Rodrigo Profeta[1], Janaína Canário Cerqueira[1], Alice Rebecca Wattam[3], Debmalya Barh[1,4], Artur Silva[2] and Vasco Azevedo[1]

[1] Departamento de Genética, Ecologia e Evolução, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2] Departamento de Genética, Universidade Federal do Pará, Belém, Pará, Brazil
[3] Biocomplexity Institute, University of Virginia, Charlottesville, Virginia, United States
[4] Institute of Integrative Omics and Applied Biotechnology, Nonakuri, West Bengal, India

ABSTRACT

Background: Within the pathogenic bacterial species *Corynebacterium* genus, six species that can produce diphtheria toxin (*C. belfantii*, *C. diphtheriae*, *C. pseudotuberculosis*, *C. rouxii*, *C. silvaticum* and *C. ulcerans*) form a clade referred to as the *C. diphtheria* complex. These species have been found in humans and other animals, causing diphtheria or other diseases. Here we show the results of a genome scale analysis to identify positive selection in protein-coding genes that may have resulted in the adaptations of these species to their ecological niches and suggest drug and vaccine targets.

Methods: Forty genomes were sampled to represent species, subspecies or biovars of *Corynebacterium*. Ten phylogenetic groups were tested for positive selection using the PosiGene pipeline, including species and biovars from the *C. diphtheria* complex. The detected genes were tested for recombination and had their sequences alignments and homology manually examined. The final genes were investigated for their function and a probable role as vaccine or drug targets.

Results: Nineteen genes were detected in the species *C. diphtheriae* (two), *C. pseudotuberculosis* (10), *C. rouxii* (one), and *C. ulcerans* (six). Those were found to be involved in defense, translation, energy production, and transport and in the metabolism of carbohydrates, amino acids, nucleotides, and coenzymes. Fourteen were identified as essential genes, and six as virulence factors. Thirteen from the 19 genes were identified as potential drug targets and four as potential vaccine candidates. These genes could be important in the prevention and treatment of the diseases caused by these bacteria.

Subjects Bioinformatics, Genomics, Microbiology, Molecular Biology
Keywords Corynebacterium, Positive selection, Drug target, Vaccine target

Submitted 22 July 2021
Accepted 30 November 2021
Published 16 February 2022

Corresponding author
Vasco Azevedo, vasco@icb.ufmg.br

Academic editor
Joseph Gillespie

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.12662**

How to cite this article Canário Viana MV, Profeta R, Cerqueira JC, Wattam AR, Barh D, Silva A, Azevedo V. 2022. Evidence of episodic positive selection in *Corynebacterium diphtheriae* complex of species and its implementations in identification of drug and vaccine targets. **PeerJ 10:e12662 DOI 10.7717/peerj.12662**

INTRODUCTION

The genus *Corynebacterium* are gram-positive bacteria of biotechnological, medical, and veterinary importance (*Bernard & Funke, 2015*). Within the pathogenic species, some can produce diphtheria toxin (DT) after lysogenization by *tox*+ corynephages (*Bernard & Funke, 2015*). Three species that compose a clade were initially described as potential diphtheria toxin (DT) producers: *C. diphtheriae*, *C. ulcerans* and *C. pseudotuberculosis* (*Bernard & Funke, 2015*). The number of species in the clade of potential DT producers increased to six with the inclusion of the recently described *C. belfantii* (*Dazas et al., 2018*), *C. rouxii* (*Badell et al., 2020*) and *C. silvaticum* (*Dangel et al., 2020*). Those six species are described here as the "*C. diphtheria* complex" (*Badell et al., 2020*).

C. diphtheriae, *C. belfantii* and *C. rouxii* infect mainly humans (*Bernard & Funke, 2015*; *Dazas et al., 2018*; *Badell et al., 2020*). *C. ulcerans*, *C. pseudotuberculosis* and *C. silvaticum* infect mainly wild and domesticated mammals and/or can cause zoonosis (*Bernard & Funke, 2015*; *Dangel et al., 2020*). *C. belfantii* and *C. rouxii* have recently been reclassified species from some of the *C. diphtheriae* biovar Belfanti strains (*Dazas et al., 2018*; *Badell et al., 2020*). *C. ulcerans*, *C. pseudotuberculosis* and *C. silvaticum* infect mainly wild and domesticated mammals but can also be zoonotic (*Bernard & Funke, 2015*; *Dangel et al., 2020*).

The *C. diphtheria* complex have an impact on public health, and also on the production of animal-based foods. Some of the species contain both DT and strains that lack the toxin. DT-producing *C. diphtheriae* strains cause cutaneous and respiratory diphtheria (*Zasada, 2013*; *Grosse-Kock et al., 2017*). The report of multidrug-resistant strains from Brazil is a new concern (*Zasada, 2014*; *Hennart et al., 2020*). DT-producing *C. pseudotuberculosis* from biovar equi causes Oedematous Skin Disease in buffalos (*Selim et al., 2015*). *C. ulcerans* infects a broad range of mammal species and DT-producing strains have caused diphtheria (*Hacker et al., 2016*). Some non-DT producing strains of *C. diphtheriae* cause endocarditis, septic arthritis, osteomyelitis and sepsis in humans (*Zasada, 2013*; *Grosse-Kock et al., 2017*). Non-DT producing strains of *C. ulcerans* are associated with ulcers in humans (*Hacker et al., 2016*). *C. pseudotuberculosis* also contains non-DT strains, with those in biovar equi causing ulcerative lymphangitis in horses, and those in the biovar ovis causing caseous lymphadenitis in goat and sheep, and lymphadenitis and abscesses in humans (*Selim et al., 2015*).

There are also *C. diphtheria* complex species that never produce DT but do cause disease. *C. belfantii* causes laryngitis and bronchopathy (*Dazas et al., 2018*). *C. rouxii* causes chronic arteritis leading to ulcerations on feet and legs, and peritonitis (*Badell et al., 2020*). *C. silvaticum* has only been isolated from pigs and roe deer to date, causing caseous lymphadenitis (*Dangel et al., 2020*), and is cytotoxic for human epithelial cells (*Möller et al., 2021*).

The host ranges and virulence mechanisms of these species are not entirely known, and better understanding of their biology could be helpful in controlling this group of pathogens. Diphtheria outbreaks were reported globally between 1921 and 2018. The disease is still endemic in some countries, with thousands of annual cases reported in

Asia and Africa. The disease can emerge when the recommended vaccination programs are not applied or sustained (*Sharma et al., 2019*). The current vaccine is based on the DT toxoid (*Rappuoli & Malito, 2014*) but does not prevent the colonization, transmission, and disease manifestation. In addition, the acquired immunity has been found to decrease with time (*Truelove et al., 2020*). Isolation of symptomatic individuals, antitoxin and antibiotics are still essential in the control of these diseases (*Truelove et al., 2020*). Furthermore, the diversity of DT toxin sequences across strains could reduce the effectiveness of diphtheria toxoid–based vaccines and diphtheria antitoxins (*Otsuji et al., 2019*). Another factor to consider in the control of these pathogens is that non-DT producing strains can cause other diseases, such as ulcers and caseous lymphadenitis, the latter associated with the Phospholipase D toxin produced by *C. ulcerans*, *C. pseudotuberculosis* and *C. silvaticum* (*Bernard & Funke, 2015*; *Dangel et al., 2020*).

Adaptive mutations for a specific ecological niche can be identified using genomic analyses, including genome-scale positive selection analysis (*Kopac et al., 2014*). At an ecological level, routine selection favors the maintenance of a stable population structure over time, while episodic selection is the effect of a sudden environmental disturbance, such as host change (*Brasier, 1995*). At the molecular level, positive selection can help fix adaptive mutations (*Anisimova & Liberles, 2012*). Episodes of positive selection can act on specific codons at specific times (phylogenetic branches), for which branch-site statistical models were developed (*Zhang, 2005*). Information on the amino acids under selection could be used for drug design (*Farhat et al., 2013*), or even reverse vaccinology if the amino acids are surface exposed (*Goodswen, Kennedy & Ellis, 2018*).

In this work, we used a genome scale positive selection analysis to identify the genes that could be involved in ecological adaptation and identified genes that can be used to develop preventive or therapeutic strategies against this group of important pathogens.

MATERIALS AND METHODS

Samples and taxonomy

Episodic positive selection across species or other phylogenetic groups of the diphtheriae group was investigated using a branch-site test (*Zhang, 2005*). This test is more appropriate for inter-specific samples, because it assumes that the observed mutations have already been fixed by selection (*Kryazhimskiy & Plotkin, 2008*; *Anisimova & Liberles, 2012*; *Kosiol & Anisimova, 2012*), and one genome could represent a species. For this reason, we limited the samples to one per species, subspecies or biovar. For the foregrounds (target groups), we selected the six type strains from the *C. diphtheria* complex and other strains to represent biovars and lineages. *C. diphtheriae* biovars could not be used as foregrounds as they are united in a single clade (*Sangal et al., 2014*). As background, we selected 40 total genomes that included 30 representative (*O'Leary et al., 2016*), eight complete and three WGS genomes of *Corynebacterium* species, all of which were available in the Pathosystems Resource Integration Center (PATRIC) (*Davis et al., 2020*). These were annotated by RASTtk (*Brettin et al., 2015*) and downloaded from PATRIC (Table S1).

The taxonomy of the samples was verified using TYGS (*Meier-Kolthoff & Göker, 2019*). TYGS determines the closest related type strains using the MASH algorithm (*Ondov et al., 2016*) for entire genomes and BLASTn for 16S sequences. It calculates the pairwise distances of 16S and genome sequences using GBDP (*Meier-Kolthoff et al., 2013*), followed by inference of 16S and genome phylogenies based on the pairwise GBDP distances using FastME (*Lefort, Desper & Gascuel, 2015*), digital DNA-DNA hybridization (dDDH) using GGDC (*Meier-Kolthoff et al., 2013*), and deviation of G+C content. Genomes with >70% of dDDH and <1% of G+C content deviation are considered to be in the same species (*Meier-Kolthoff & Göker, 2019*).

Positive selection analysis

A genome-scale positive selection analysis was performed using the PosiGene pipeline (*Sahm et al., 2017*) on the *Corynebacterium* genomes. Ten foreground genomes were tested (Table S2) representing 10 target clades or subclades. Six clades represent the species from the *C. diphtheria* complex (*C. belfantii*, *C. diphtheriae*, *C. pseudotuberculosis*, *C. rouxii*, *C. silvaticum* and *C. ulcerans*), two subclades represent *C. pseudotuberculosis* (biovars equi and ovis), and two subclades representing *C. ulcerans* lineages (lineage 1 and 2).

In the module "create_catalog", homologous genes were identified using BLASTp (*Camacho et al., 2009*) with the best-bidirectional hit criterion (*Altenhoff & Dessimoz, 2009*). In the module "alignments", orthologous genes are identified, gene trees are built, and a species tree is built from the gene trees. In this module, the anchor species is the genome that the orthologous gene sequences are aligned to, and the reference species is the genome from which the gene names are extracted. *C. diphtheriae* NCTC 11397^{T} was selected as both the reference and anchor genome. In the first step, orthologous gene sequences were aligned to the anchor genome sequences using CLUSTALW (*Larkin et al., 2007*), with the parameters' minimum identity and minimum pairs identity set to 40%. The aligned sequences were filtered by GBLOCKS for gaps and unreliable alignment columns (*Jordan & Goldman, 2012*). In the next step, a phylogeny was built for each gene using the parsimony method and jackknifing implemented in DNAPARS from the PHYLIP package (*Felsenstein, 2005*). In the third step, a species tree was built based on the gene trees consensus, using CONSENSE from the PHYLIP package. The species tree is required to test for positive selection along specific lineages (*Yang & Nielsen, 2002*). The consensus tree was visualized using FigTree v1.4.4 and rooted using *C. kroppenstedtii* DSM 44385 as an outgroup. This strain was identified as an outgroup based on another tree generated by the same pipeline including *M. tuberculosis* H37RV to find the correct *Corynebacterium* species to use as an outgroup (Table S1) in that tree. The *M. tuberculosis* tree was not included in the downstream analysis.

In the module "positive_selection", a likelihood ratio test compares the non-synonymous to synonymous substitution rate $\omega = d_N/d_S$ in the foreground and the background. Here, d_N is the number of non-synonymous substitutions per non-synonymous site and d_S is the number of synonymous substitutions per synonymous site. The episodic positive selection model assumes $\omega > 1$ in the foreground and $\omega = 0$ or $\omega < 1$ in the background, while the null model assumes $\omega = 0$ or $\omega < 1$ for foreground and

background (*Yang & Dos Reis, 2011*). We considered positive selection if $p < 0.05$ for False Discovery Rate, as this correction is more suitable for genome wide experiments (*Storey & Tibshirani, 2003*).

Due to an assumption of no recombination by the branch-site models we used (*Yang, 2005*), recombination could cause false positive results (*Anisimova, Nielsen & Yang, 2003*). To avoid that artifact, the genes identified as positively selected by PosiGene were then tested for intragenic recombination using PhiPack (*Bruen, Philippe & Bryant, 2006*) that calculates Pairwise Homoplasy Index method (PHI) (*Bruen, Philippe & Bryant, 2006*), Neighbor Similarity Score (NSS) (*Jakobsen & Easteal, 1996*), and Maximum Chi-Square (*Smith, 1992*). In our analysis, we considered that recombination had occurred when $q < 0.05$ for PHI and at least NSS or Maximum Chi-Square (*Hongo et al., 2015*). Genes identified as recombinant were discarded for downstream analysis.

To minimize the false positive results caused by misalignments, frameshifts and ortholog prediction (*Schneider et al., 2009*; *Markova-Raina & Petrov, 2011*), we visually checked the alignments and checked the homology prediction by comparing the orthologs protein domains using PATRIC's annotation of Local and Global Families (*Davis et al., 2016*) and the Conserved Domain Database (CDD) (*Lu et al., 2020*).

Gene annotation

Gene function was predicted using annotations from PATRIC and eggNOG-mapper (*Huerta-Cepas et al., 2017*), and InterProScan (*Jones et al., 2014*) was used to examine protein domains. Subcellular localization of the proteins was assessed using SufG+ v1.2.1 (*Barinov et al., 2009*). Protter (*Omasits et al., 2014*) was used to identify the position of the positively selected amino acids in relation to the cytoplasmic, transmembrane, and surface exposed portions of the proteins.

GIPSy v1.1.2 (*Soares et al., 2016*) was used to identify genomic islands (GIs) in *C. diphtheriae* NCTC 11397^{T} using *C. glutamicum* ATCC 1302 (NC_006958.1) as the non-pathogenic reference. The islands were predicted by genes with C+G content and codon usage deviation, presence of transposases, presence of specific genes, non-conservation in comparison to reference genome, and flanking tRNA genes. The islands were classified according to the proportion of specific genes as pathogenicity islands, metabolic islands, resistance islands and symbiotic islands. Finally, prophages were predicted using PHASTER (*Arndt et al., 2016*).

Prediction of drug and vaccine targets

Virulence genes and drug targets were predicted using the Pipeline Builder for Identification of Targets (PBIT) (*Shende et al., 2016*). PBIT predicted drug targets among the cytoplasmic proteins by a subtractive approach, identifying sequences of interest using BLASTp and specific databases in the following way. First, homologs to the human proteome, anti-targets and gut microbiota proteomes were filtering out to avoid cross-reactivity of drugs. Then the essential genes were identified using the Database of Essential Genes (*Zhang, 2004*) and virulence genes using the Virulence Factor Database (*Chen et al., 2005*). The druggability of the remaining candidates was predicted by

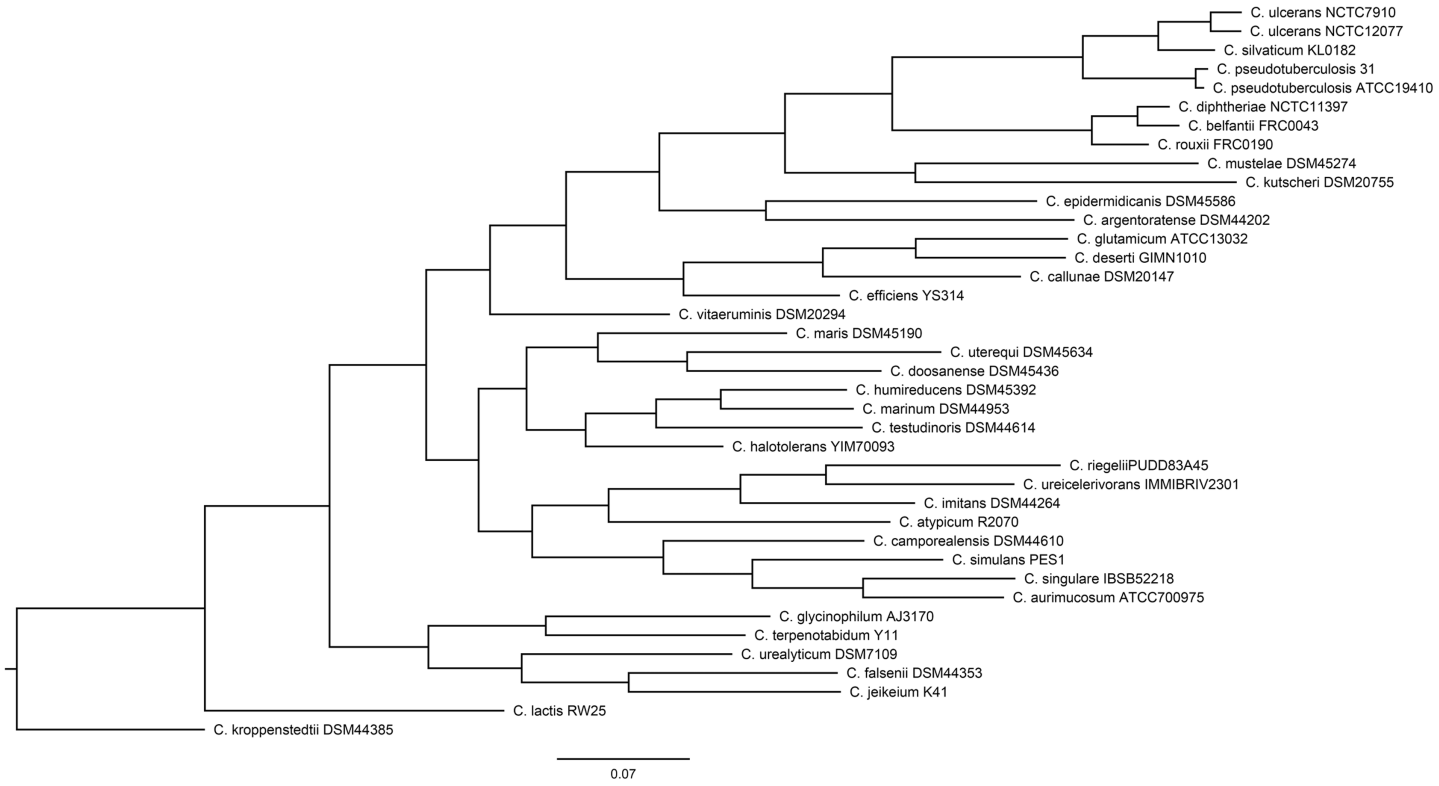


Figure 1 **The *Corynebacterium* species tree that was generated by the PosiGene pipeline, using CONSENSE from the PHYLIP package.**
Full-size DOI: 10.7717/peerj.12662/fig-1

similarity to experimentally validated druggable targets from the Therapeutic Target Database (*Li et al., 2018*).

Vaccine targets were predicted from the transmembrane, putative surface exposed and secreted proteins predicted using SurfG+ and Vaxign (*He, Xiang & Mobley, 2010*) with prediction of human MHC Class I and II epitopes.

RESULTS

Of the 40 genomes examined, the genome of *C. casei* LMG S-19264 was discarded due to its being identified as *Brevibacterium linens* based on the TYGS pipeline (Data S1). For this reason, only 39 *Corynebacterium* genomes were used for the downstream analysis. Figure 1 shows the phylogeny of the *Corynebacterium* genomes built by the PosiGene pipeline.

The PosiGene analysis showed zero to nine genes under positive selection ($p < 0.05$ for FDR) depending upon the 10 foregrounds that were used (Cb, Cd, Cp, Cpequi, Cpovis, Cr, Cs, Cul, Cul1, Cul2 (Tables S3–S12)), with 22 total genes shown to be under selection. Additional analysis of the 22 genes using PhiPack ($q < 0.05$ for PHI and at least NSS or Maximum Chi-Square tests) showed recombination in two out of four genes when Cd was the foreground, and one out of nine genes when Cp was used (Table S13). Manual curation of the 19 remaining proteins did not reveal any false positives caused by misalignments and ortholog prediction by comparison of protein domains (Table S14).

SurfG+ predicted 15 cytoplasmic, two membrane, one putative surface exposed and one secreted protein (Table S15). Thirty-five genomic islands and two prophages were predicted in *C. diphtheriae* NCTC 11397^{T} (Tables S16 and S17).

From the 19 genes that were identified as positively selected across the species, two were identified in *C. diphtheriae*, 10 in *C. pseudotuberculosis*, one in *C. rouxii*, and six in *C. ulcerans* (Table 1 and Table S15). The COG categories of those genes were shown to be involved in defense, translation, energy production, and transport and metabolism of carbohydrates, amino acids, nucleotides, and coenzymes (Fig. 2). Based on our *in silico* prediction, 14 genes were found to be essential, six were virulence factors, and three were found to be in genomic islands (Table 1 and Table S15).

Thirteen of the genes were identified as potential drug targets by different analyses. Three genes were predicted based on our pipeline (Tables 1 and 2, and Table S15). The other 10 genes were not tagged as potential targets by the homology or druggability filters of the pipeline, but were included due to the possibility of targeting them with other methods (see Discussion section). For vaccine targets, four genes were predicted based on our pipeline (Tables 1 and 2, and Table S15).

DISCUSSION

We identified 19 genes under positive selection, 13 potential drug targets and four potential vaccine targets. From the 13 potential drug targets, 10 were includes despite not passing the homology or druggability filters of the PBIT pipeline (Tables 1 and 2, and Table S15). The problem of having homology with the human proteome or human gut microbiota proteome can be solved by screening compounds that selectively inhibit the pathogen protein (*Arya et al., 2015*). The druggability prediction of the PBIT pipeline is based on sequence similarity to experimentally validated targets (*Shende et al., 2016*). So, the lack of predicted druggability by that pipeline can be solved by prediction of druggable pockets of a protein based on its own structure (*Volkamer et al., 2012*). Additionally, some of those proteins are known drug targets in other species.

C. diphtheriae

In *C. diphtheriae* (foreground Cd), we identified two genes encoding proteins predicted to be essential and involved in translation, amino acid transport and metabolism (Table 1 and Table S15). The gene *ansB* is in GI10 and encodes secreted L-asparaginase type II which is a high-affinity enzyme that catalyzes the conversion of L-asparagine to L-aspartate and ammonia. The *E. coli*, *Dickeya dadantii* and human homologs to this gene are used for leukemia treatment, where the consequent low L-asparagine levels in plasma leads to apoptosis of the leukemia cells (*Lubkowski & Wlodawer, 2021*). This gene was suggested as a candidate vaccine target due to its classification as a secreted protein and predicted epitopes (Table 2). The second protein, SSU ribosomal protein S3p (*rpsC*), is a 30S ribosomal subunit that binds to the initiator Met-tRNA (*Burd & Dreyfuss, 1994*). Possible reasons for the selective pressure on these genes could be the effects on L-aspartate uptake (*ansB*) and translation efficiency (*rpsC*). *rpsC* was suggested as a drug target but

Table 1 **Characterization and possible application of 19 genes under positive selection in different species of the *Corynebacterium diphtheria* complex.**

n	Product (Gene)	PS sites	PS positions	Local	COG	Essential	VF	Target	GenBank ID
	C. diphtheriae								
1	L-asparaginase, type II (EC 3.5.1.1) (*ansB*)	3	69, 182, 339	S	EJ	Yes	No	Va[2]	ERS451417_00414
2	SSU ribosomal protein S3p (S3e) (*rpsC*)	1	89	C	J	Yes	No	Dr[4]	ERS451417_00402
	C. pseudotuberculosis								
3	ABC transporter, permease protein (*mntC*)	1	111	M	P	Yes	Yes	Va[2]	ERS451417_00548
4	Adenosine deaminase (*add*)	1	25	C	F	–	No	–	ERS451417_00570
5	Adhesin SpaE (*spaE*)	20	23, 33, 35, 108, 119, 122, 125, 223, 232, 238, 239, 243, 244, 247, 251, 252, 253, 255, 257, 259	SE	–	No	Yes	Va[2]	ERS451417_00159
6	Dihydropteroate synthase 2 (nonfunctional) (*folP*)	2	135, 158	C	H	Yes	No	Dr[1,3]	ERS451417_00887
7	HNH endonuclease	4	48, 111, 284, 352	C	V	Yes	No	Dr[4]	ERS451417_00880
8	Peptide chain release factor 1 (*prfA*)	1	60	C	J	Yes	No	Dr[3]	ERS451417_00951
9	Putative oxidoreductase	8	33, 42, 48, 52, 109, 209, 226, 285	C	CH	No	Yes	–	ERS451417_02135
10	Putative phosphoglycerate mutase (*pgmB*)	2	69, 143	C	G	Yes	No	Dr[1,3]	ERS451417_02267
	***C. pseudotuberculosis* equi**								
11	Methionine aminopeptidase (EC 3.4.11.18) (*mapB*)	2	96, 97	C	E	Yes	No	Dr[3]	ERS451417_01521
12	Tyrosyl-tRNA synthetase (EC 6.1.1.1) (*tyrS*)	4	23, 58, 59, 403	V	J	Yes	Yes	Dr[3]	ERS451417_01169
	C. rouxii								
13	Hypothetical protein	4	5, 74, 95, 154	M	S	–	No	Va[2]	ERS451417_00470
	C. ulcerans								
14	Serine hydroxymethyltransferase (EC 2.1.2.1) (*glyA*)	1	385	C	E	Yes	No	Dr[4]	ERS451417_00836
	***C. ulcerans* lineage 1**								
15	Hypothetical protein	1	32	C	–	–	–	Dr[4]	ERS451417_00635
16	Phosphoenolpyruvate-dihydroxyacetone phosphotransferase, dihydroxyacetone binding subunit DhaK (*dnaK*)	4	248, 251, 255, 256	C	G	Yes	Yes	Dr[4]	ERS451417_02360
17	Similar to citrate lyase beta chain, 3 (*citE*)	1	235	C	G	Yes	Yes	Dr[1]	ERS451417_00750
	C. ulcerans* lineage *2								
18	DNA polymerase III epsilon subunit (EC 2.7.7.7) (*dnaQ*)	1	142	C	L	Yes	No	Dr[4]	ERS451417_00985
19	Precorrin-6A reductase (EC 1.3.1.54) (*cobK*)	1	206	C	H	Yes	No	Dr[4]	ERS451417_01234

Notes:
Columns: VF, virulence factor; COG, Clusters of Orthologous Groups; PS sites, positively selected sites; SE, surface exposed (sites).
Column Local: C, cytoplasm; M, membrane; SE, surface exposed; S, secreted.
Column COG: C, energy production and conversion; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; J, translation, ribosomal structure and biogenesis; S, function unknown; V, defense mechanisms.
Column Target: Dr, drug target; Va, vaccine target; [1]–predicted by PBIT pipeline, [2]–predicted by essentiality, local and Vaxign, [3]–described in literature for other species, [4]–suggested despite not attending one or more pipeline filters.

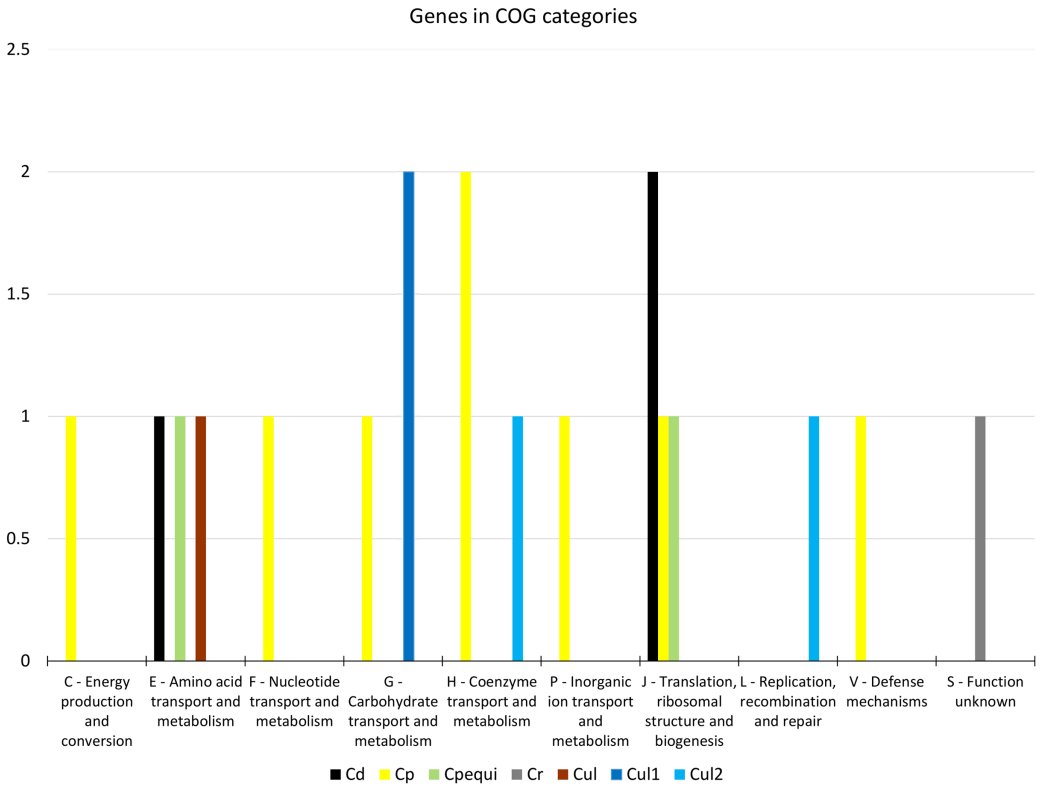


Figure 2 Distribution of 19 genes under positive selection in COG categories. Target groups: Cd, *C. diphtheriae*; Cp, *C. pseudotuberculosis*; Cpequi, *C. pseudotuberculosis* biovar equi; Cr, *C. rouxii*; Cul, *C. ulcerans*; Cul1, *C. ulcerans* lineage 1; Cul2, *C. ulcerans* lineage 2. COG categories: C-Q, metabolism; J-L, information storage and processing; M-V, cellular processes and signaling; S, poorly characterized.
Full-size DOI: 10.7717/peerj.12662/fig-2

has homology to a protein in the human gut microbiota proteome, requiring compounds that can selectively inhibit it.

C. pseudotuberculosis

When *C. pseudotuberculosis* was the foreground (Cp), eight positively selected genes were identified. Five of the genes were tagged as essential, and three as virulence factors. They are involved in translation, coenzyme transport and metabolism, inorganic ion transport, defense from foreign DNA, nucleotide transport and metabolism and adhesion (Table 1 and Table S15). Among the essential genes, an ABC transporter permease protein (*mntC*) plays a role in the transport of Mn^{2+} and Zn^{2+} (*Claverys, 2001*) and was classified as a virulence factor. The Dihydropteroate synthase 2 (*folP2*) is nonfunctional according to PATRIC annotation but functional genes with this annotation are essential for the *de novo* synthesis of folate (*Bertacine Dias et al., 2018*). The HNH endonucleases degrade foreign DNA, and can also be involved in DNA repair, replication and recombination (*Wu, Lin & Yuan, 2020*). Peptide chain release factor 1 (*prfA*) recognizes the stop codons UAA and UAG, promoting the end of translation (*Scolnick et al., 1968*). The Putative phosphoglycerate mutase (*pgmB*) is capable of interconverting

Table 2 **Final drug and vaccine target candidates for *Corynebacterium* species based on a positive selection analysis by foreground, application, and priority.**

Foreground	Application	Priority	Product (Gene)	PS sites (exposed sites)	Local	GenBank ID
Cd	Drug target	1 (gut microbiota homolog)	SSU ribosomal protein S3p (S3e) (*rpsC*)	89	C	ERS451417_00402
Cd	Vaccine	1	L-asparaginase (*ansB*)	69, 182, 339 (69, 182, 339)	S	ERS451417_00414
Cp	Drug target	1	Putative phosphoglycerate mutase (*pgmB*)	69, 143	C	ERS451417_02267
Cp	Drug target	2 (predicted as nonfunctional)	Dihydropteroate synthase 2 (nonfunctional) (*folP2*)	135, 158	C	ERS451417_00887
Cp	Drug target	3 (human and gut microbiota homolog, no predicted druggability)	Peptide chain release factor 1 (*prfA*)	60	C	ERS451417_00951
Cp	Drug target	4 (no predicted druggability)	HNH endonuclease	48, 111, 284, 352	C	ERS451417_00880
Cp	Vaccine	1	Adhesin SpaE (*spaE*)	23, 33, 35, 108, 119, 122, 125, 223, 232, 238, 239, 243, 244, 247, 251, 252, 253, 255, 257, 259 (23, 33, 35, 108, 119, 122, 125, 223)	SE	ERS451417_00159
Cp	Vaccine	2 (no exposed PS site)	ABC transporter, permease protein (*mntC*)	111	M	ERS451417_00548
Cpequi	Drug target	1 (virulence factor, more PS sites, gut microbiota homolog, target in literature)	Tyrosyl-tRNA synthetase (*tyrS*)	23, 58, 59, 403	C	ERS451417_01169
Cpequi	Drug target	2 (less PS sites, gut microbiota homolog, target in literature)	Methionine aminopeptidase (*mapB*)	96, 97	C	ERS451417_01521
Cr	Vaccine	1	Hypothetical protein	5, 74, 95, 154 (154)	M	ERS451417_00470
Cul	Drug target	1 (human and gut microbiota homolog)	Serine hydroxymethyltransferase (*glyA*)	385	C	ERS451417_00836
Cul1	Drug target	1 (virulence factor)	Similar to citrate lyase beta chain (*citE*)	235	C	ERS451417_00750
Cul1	Drug target	2 (virulence factor, gut microbiota homolog)	Phosphoenolpyruvate-dihydroxyacetone phosphotransferase, dihydroxyacetone binding subunit DhaK (*dhaK*)	248, 251, 255, 256	–	ERS451417_02360
Cul1	Drug target	3 (hypothetical protein, no predicted druggability)	Hypothetical protein	32	C	ERS451417_00635
Cul2	Drug target	Equal. No predicted druggability	DNA polymerase III epsilon subunit (*dnaQ*)	142	C	ERS451417_00985
Cul2	Drug target	Equal. No predicted druggability	Precorrin-6A reductase (*cobK*)	206	C	ERS451417_01234

Note:
PS sites, positively selected sites.

2- and 3-phosphoglycerate in glycolysis (*Rigden, 2008*), although this particular gene is annotated as putative.

The other three identified genes (*add, spaE* and the putative oxidoreductase) are not characterized as essential, so may not be suitable drug targets. Adenosine deaminase (*add*) catalyzes the hydrolytic deamination of adenosine into inosine (*Chang et al., 1991*). *spaE*, from the operon *spaDEF*, encodes the minor pilin SpaE in *C. diphtheriae* (*Mandlik et al., 2008*) and is in GI5. It's ortholog in *C. pseudotuberculosis* also encodes the minor pilin and was first described as *spaB* from the operon *spaABC* (*Trost et al., 2010*). The putative oxidoreductase is a flavoenzyme with a "FAD-binding domain, ferredoxin reductase-type" (IPR017927), but its specific reaction is unknown.

Why would these particular genes be under positive selection? One could hypothesize that there would be more efficient manganese uptake (*mntC*), tissue adhesion on a new host range (*spaE*), improved efficiency for defense against foreign DNA (HNH endonuclease), translation (*prfA*), and metabolism of nucleotides (*add*) and carbohydrates (*pgmB*).

The vaccine targets (*mntC* and *spaE*) were indicated due to either their membrane location, their predicted epitopes, and that they might have surface exposed sites that are under positive selection. There were four drug targets (Table 2). *pgmB* was predicted as a target by PBIT and is this same gene is a drug target in helminth parasites (*Timson, 2016*). *folP2* was also predicted and is a well know target of sulfa and imidazole derivatives in human pathogens such as *Staphylococcus aureus*, *M. tuberculosis*, *Bacillus anthracis*, *Streptococcus pneumoniae*, *Burkholderia cenocepacia* and *Yersinia pestis* (*Bertacine Dias et al., 2018*). Although it was annotated as non-functional, the evidence of positive selection in this protein suggests an unknown adaptation due to specific amino acids that could be targeted. *prfA* has homology to human and human gut microbiota proteome and has no predicted druggability, but it is a known target of the drug Apidaecin in gram negative bacteria (*Matsumoto et al., 2017*). The HNH endonuclease had no predicted druggability.

C. pseudotuberculosis biovar equi

When *C. pseudotuberculosis* biovar equi was used as the foreground (Cpequi), two genes were identified as being under positive selection, and were also characterized as essential. These two genes (*mapB* and *tyrS*) are both involved in translation (Table 1 and Table S15). Methionine aminopeptidase (*mapB*) cleaves the initiator methionine from newly synthesized polypeptides (*Helgren et al., 2016*; *Pillalamarri et al., 2021*). Tyrosyl-tRNA synthetase (*tyrS*) attaches the amino acid tyrosine to the appropriate tRNA (*Hughes et al., 2020*; *Othman et al., 2021*). These two genes could be under positive selection as it could affect translation efficiency. Both genes were predicted as homologs to human gut microbiota and have been identified as drug targets in other studies (Table 2). *tyrS* was predicted as a virulence factor and an ortholog from *Pseudomonas aeruginosa* was found to be targeted by four drug-like compounds (*Hughes et al., 2020*), and a *S. aureus* ortholog to this gene was targeted by new pyrazolone and dipyrazolotriazine derivatives (*Othman et al., 2021*). *mapB* from *M. tuberculosis* and

S. pneumoniae were shown to be selectively targeted despite homology to human protein (*Krátký et al., 2012*; *Arya et al., 2015*).

C. rouxii

A single hypothetical protein (ERS451417_00470) was identified when *C. rouxii* was used as the foreground (Cr). It had no predicted domains, but it could be a vaccine target candidate (Table 2) due to its transmembrane location and the surface exposed sites under positive selection.

C. ulcerans

A single essential gene (*glyA*) was identified when *C. ulcerans* was used as the foreground (Cul). *glyA* encodes a serine hydroxymethyltransferase enzyme (Table 1 and Table S15). This gene is known to participate in the one-carbon metabolism of serine/glycine interconversion and also in the folate/methionine cycle (*Batool et al., 2020*), which could explain its being under selective pressure. This gene was also identified as a potential drug target (Table 2), but it does have homology to human and human gut microbiota proteome. It has been shown to play a key role in lysostaphin resistance in *Staphylococcus aureus* (*Batool et al., 2020*).

C. ulcerans lineage 1

Three genes were identified when the *C. ulcerans* lineage 1 genome was used as the foreground (Cul1). Two of them were essential, involved in carbohydrate transport and metabolism (Table 1 and Table S15). The Phosphoenolpyruvate-dihydroxyacetone phosphotransferase, dihydroxyacetone binding subunit DhaK (*dhaK*) gene is in GI34. This enzyme phosphorylates ketones and short chain aldoses using adenosine triphosphate (ATP) (*Peiro et al., 2019*). The second gene is annotated in PATRIC as "Similar to citrate lyase beta chain, 3" (*citE*) and is probably one of the catalytic subunits of citrate lyase, the enzyme that catalyzes the cleavage of citrate to acetate and oxaloacetate during citrate fermentation (*Schneider, Dimroth & Bott, 2000*). Both proteins were predicted as virulence factors by PBIT. The third gene encodes a hypothetical protein (ERS451417_00635) with no predicted domain or cellular localization.

The two genes with predicted function (*dhaK* and *citE*) appear to be related to metabolism inside the host. In *Listeria monocytogenes*, Phosphoenolpyruvate-dihydroxyacetone phosphotransferase (DhaK and other subunits) is required to utilize carbon sources for amino acid synthesis inside murine macrophages (*Eylert et al., 2008*). In *Enterococcus faecalis*, mutants of citrate fermentation genes (*citE* and others) were less pathogenic for the model *Galleria mellonella* (*Martino et al., 2018*). These same two genes are possible drug targets (Table 2). *citE* was predicted as a virulence factor and drug target candidate, while *dhaK* was predicted as a virulence factor and suggested despite the homology to a protein in the gut microbiota homology.

C. ulcerans lineage 2

Two essential genes (Table 1 and Table S15) were identified when the *C. ulcerans* lineage 2 was used as the foreground (Cul2). The DNA polymerase III epsilon subunit (*dnaQ*) has a

domain with 3′–5′ exonuclease proofreading activity (*Raia, Delarue & Sauguet, 2019*). The other gene, *cobK*, encodes Precorrin-6A reductase which is involved in part I of the cobalamin cofactor (vitamin B12) biosynthesis pathway (*Kipkorir et al., 2021*). The mutations seen in these genes could provide the organism with a more efficient means of DNA replication (*dnaQ*) and biosynthesis of the essential cofactor cobalamin (*cobK*). Neither of these genes had any predictable druggability.

Probable adaptations across groups

It is reported that most of the genes identified as being under positive selection are exposed on the surface and are involved in host colonization, and resistance to phage and antibiotics (*Petersen et al., 2007*; *Anisimova & Liberles, 2012*). Those under positive selection that are not surface exposed have been shown to be involved in metabolism (*Petersen et al., 2007*; *Rao, Sivakumar & Jayakumar, 2019*) or gene regulation (*Zhang et al., 2011*). Considering the function of the identified genes, most of the probable adaptations appear to be related to metabolism. A notable exception is in *C. pseudotuberculosis*, where the pilin SpaE could have enhanced adhesion to different host species tissues. Although the specific adaptations are not clear, an amino acid fixed by positive selection is an attractive target for a therapeutic molecule, as a non-synonymous mutation that could avoid interaction would decrease fitness. These genes could be used for reverse vaccinology and *in silico* drug targeting methods.

CONCLUSION

In this analysis, we predicted 19 genes with non-synonymous mutations that are probably involved in adaptations found in the pathogens *C. diphtheriae*, *C. pseudotuberculosis*, *C. rouxii* and *C. ulcerans*. Based on our pipeline and literature data, 13 genes are candidate drug targets and four are potential vaccine targets, but their effectiveness would require experimental validation.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPEMIG (Fundação de Amparo à Pesquisa de Minas Gerais). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil.
Conselho Nacional de Desenvolvimento Científico e Tecnológico.
Fundação de Amparo à Pesquisa de Minas Gerais.

Competing Interests

Debmalya Barh and Vasco Azevedo are Academic Editors for PeerJ.

Author Contributions

- Marcus Vinicius Canário Viana conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Rodrigo Profeta conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Janaína Canário Cerqueira conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Alice Rebecca Wattam conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Debmalya Barh conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Artur Silva conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Vasco Azevedo conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data are the gene nucleotide sequences listed in Table S1 and Tables 1 and 2.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12662#supplemental-information.

REFERENCES

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLOS Computational Biology* **5**:e1000262 DOI 10.1371/journal.pcbi.1000262.

Anisimova M, Liberles DA. 2012. Detecting and understanding natural selection. In: Cannarozzi GM, Schneider A, eds. *Codon Evolution: Mechanisms and Models*. Oxford: Oxford University Press, 73–96.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**:1229–1236 DOI 10.1093/bioinformatics/btn086.

Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research* **44**:W16–W21 DOI 10.1093/nar/gkw387.

Arya T, Reddi R, Kishor C, Ganji RJ, Bhukya S, Gumpena R, McGowan S, Drag M, Addlagatta A. 2015. Identification of the molecular basis of inhibitor selectivity between the

human and streptococcal type I methionine aminopeptidases. *Journal of Medicinal Chemistry* **58**:2350–2357 DOI 10.1021/jm501790e.

Badell E, Hennart M, Rodrigues C, Passet V, Dazas M, Panunzi L, Bouchez V, Carmi-Leroy A, Toubiana J, Brisse S. 2020. Corynebacterium rouxii sp. nov., a novel member of the diphtheriae species complex. *Research in Microbiology* **171**:122–127 DOI 10.1016/j.resmic.2020.02.003.

Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlichh D, Maguin E, van Guchte MD. 2009. Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria. *Proteomics* **9**:61–73 DOI 10.1002/pmic.200800195.

Batool N, Ko KS, Chaurasia AK, Kim KK. 2020. Functional identification of serine hydroxymethyltransferase as a key gene involved in lysostaphin resistance and virulence potential of staphylococcus aureus strains. *International Journal of Molecular Sciences* **21**:9135 DOI 10.3390/ijms21239135.

Bernard AL, Funke G. 2015. Corynebacterium. In: *Bergey's Manual of Systematic of Archaea and Bacteria (Online)*. London: John Wiley & Sons, Bergey's Manual Trust, 1–70.

Bertacine Dias MV, Santos JC, Libreros-Zúñiga GA, Ribeiro JA, Chavez-Pacheco SM. 2018. Folate biosynthesis pathway: mechanisms and insights into drug design for infectious diseases. *Future Medicinal Chemistry* **10**:935–959 DOI 10.4155/fmc-2017-0168.

Brasier CM. 1995. Episodic selection as a force in fungal microevolution, with special reference to clonal speciation and hybrid introgression. *Canadian Journal of Botany* **73**:1213–1221 DOI 10.1139/b95-381.

Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports* **5**:1–6 DOI 10.1038/srep08365.

Bruen TC, Philippe HH, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**:2665–2681 DOI 10.1534/genetics.105.048975.

Burd CG, Dreyfuss G. 1994. Conserved structures and diversity of functions of RNA-binding proteins. *Science* **265**:615–621 DOI 10.1126/science.8036511.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Chang ZY, Nygaard P, Chinault AC, Kellems RE. 1991. Deduced amino acid sequence of Escherichia coli adenosine deaminase reveals evolutionarily conserved amino acid residues: implications for catalytic function. *Biochemistry* **30**:2273–2280 DOI 10.1021/bi00222a033.

Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research* **33**:D325–D328 DOI 10.1093/nar/gki008.

Claverys JP. 2001. A new family of high-affinity ABC manganese and zinc permeases. *Research in Microbiology* **152**:231–243 DOI 10.1016/s0923-2508(01)01195-0.

Dangel A, Berger A, Rau J, Eisenberg T, Kämpfer P, Margos G, Contzen M, Busse H-J, Konrad R, Peters M, Sting R, Sing A. 2020. Corynebacterium silvaticum sp. nov., a unique group of NTTB corynebacteria in wild boar and roe deer. *International Journal of Systematic and Evolutionary Microbiology* **70(6)**:3614–3624 DOI 10.1099/ijsem.0.004195.

Davis JJ, Gerdes S, Olsen GJ, Olson R, Pusch GD, Shukla M, Vonstein V, Wattam AR, Yoo H. 2016. PATtyFams: protein families for the microbial genomes in the PATRIC database. *Frontiers in Microbiology* **7**:1–12 DOI 10.3389/fmicb.2016.00118.

Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler RRM, Butler RRM, Chlenski P, Conrad N, Dickerman A, Dietrich EM, Gabbard JL, Gerdes S, Guard A, Kenyon RW, Machi D, Mao C, Murphy-Olson D, Nguyen M, Nordberg EK, Olsen GJ, Olson RD, Overbeek JC, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomas C, VanOeffelen M, Vonstein V, Warren AS, Xia F, Xie D, Yoo H, Stevens R. 2020. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Research* **48**:D606–D612 DOI 10.1093/nar/gkz943.

Dazas M, Badell E, Carmi-Leroy A, Criscuolo A, Brisse S. 2018. Taxonomic status of Corynebacterium diphtheriae biovar Belfanti and proposal of Corynebacterium belfantii sp. nov. *International Journal of Systematic and Evolutionary Microbiology* **68**:3826–3831 DOI 10.1099/ijsem.0.003069.

Eylert E, Schär J, Mertins S, Stoll R, Bacher A, Goebel W, Eisenreich W. 2008. Carbon metabolism of Listeria monocytogenes growing inside macrophages. *Molecular Microbiology* **69**:1008–1017 DOI 10.1111/j.1365-2958.2008.06337.x.

Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M. 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nature Genetics* **45**:1183–1189 DOI 10.1038/ng.2747.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by Author. Seattle: Department of Genome Sciences, University of Washington. *Available at http://evolution.genetics.washington.edu/phylip.html*.

Goodswen SJ, Kennedy PJ, Ellis JT. 2018. A gene-based positive selection detection approach to identify vaccine candidates using toxoplasma gondii as a test case protozoan pathogen. *Frontiers in Genetics* **9**:1–16 DOI 10.3389/fgene.2018.00332.

Grosse-Kock S, Kolodkina V, Schwalbe EC, Blom J, Burkovski A, Hoskisson PA, Brisse S, Smith D, Sutcliffe IC, Titov L, Sangal V. 2017. Genomic analysis of endemic clones of toxigenic and non-toxigenic Corynebacterium diphtheriae in Belarus during and after the major epidemic in 1990s. *BMC Genomics* **18**:1–10 DOI 10.1186/s12864-017-4276-3.

Hacker E, Antunes CA, Mattos-Guaraldi AL, Burkovski A, Tauch A. 2016. Corynebacterium ulcerans, an emerging human pathogen. *Future Microbiology* **11**:1191–1208 DOI 10.2217/fmb-2016-0085.

He Y, Xiang Z, Mobley HLT. 2010. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *Journal of Biomedicine and Biotechnology* **2010**:297505 DOI 10.1155/2010/297505.

Helgren TR, Wangtrakuldee P, Staker BL, Hagen TJ. 2016. Advances in bacterial methionine aminopeptidase inhibition. *Current Topics in Medicinal Chemistry* **16**:397–414 DOI 10.2174/1568026615666150813145410.

Hennart M, Panunzi LG, Rodrigues C, Gaday Q, Baines SL, Barros-Pinkelnig M, Carmi-Leroy A, Dazas M, Wehenkel AM, Didelot X, Toubiana J, Badell E, Brisse S. 2020. Population genomics and antimicrobial resistance in Corynebacterium diphtheriae. *Genome Medicine* **12**:107 DOI 10.1186/s13073-020-00805-7.

Hongo JA, de Castro GM, Cintra LC, Zerlotini A, Lobo FP. 2015. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC Genomics* **16**:567 DOI 10.1186/s12864-015-1765-0.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**:2115–2122 DOI 10.1093/molbev/msx148.

Hughes CA, Gorabi V, Escamilla Y, Dean FB, Bullard JM. 2020. Two forms of tyrosyl-tRNA synthetase from pseudomonas aeruginosa : characterization and discovery of inhibitory compounds. *SLAS DISCOVERY: Advancing the Science of Drug Discovery* **25**:1072–1086 DOI 10.1177/2472555220934793.

Jakobsen IB, Easteal S. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Computer Applications in the Biosciences* **12**:291–295 DOI 10.1093/bioinformatics/12.4.291.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240 DOI 10.1093/bioinformatics/btu031.

Jordan G, Goldman N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular Biology and Evolution* **29**:1125–1139 DOI 10.1093/molbev/msr272.

Kipkorir T, Mashabela GT, de Wet TJ, Koch A, Wiesner L, Mizrahi V, Warner DF. 2021. De novo cobalamin biosynthesis, transport, and assimilation and cobalamin-mediated regulation of methionine biosynthesis in mycobacterium smegmatis. *Journal of Bacteriology* **203**:e00620 DOI 10.1128/JB.00620-20.

Kopac S, Wang Z, Wiedenbeck J, Sherry J, Wu M, Cohan FM. 2014. Genomic heterogeneity and ecological speciation within one subspecies of bacillus subtilis. *Applied and Environmental Microbiology* **80**:4842–4853 DOI 10.1128/AEM.00576-14.

Kosiol C, Anisimova M. 2012. Selection on the protein-coding genome. In: Anisimova M, ed. *Evolutionary Genomics*. New York, Dordrecht, Heidelberg, London: Humana Press, 113–140.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLOS Genetics* **4**:e1000304 DOI 10.1371/journal.pgen.1000304.

Krátký M, Vinšová J, Novotná E, Mandíková J, Wsól V, Trejtnar F, Ulmann V, Stolaříková J, Fernandes S, Bhat S, Liu JO. 2012. Salicylanilide derivatives block Mycobacterium tuberculosis through inhibition of isocitrate lyase and methionine aminopeptidase. *Tuberculosis* **92**:434–439 DOI 10.1016/j.tube.2012.06.001.

Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**:2947–2948 DOI 10.1093/bioinformatics/btm404.

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution* **32**:2798–2800 DOI 10.1093/molbev/msv150.

Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, Fu T, Zhang X, Cui X, Tu G, Zhang Y, Li S, Yang F, Sun Q, Qin C, Zeng X, Chen Z, Chen YZ, Zhu F. 2018. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Research* **46**:D1121–D1127 DOI 10.1093/nar/gkx1076.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research* **48**:D265–D268 DOI 10.1093/nar/gkz991.

Lubkowski J, Wlodawer A. 2021. Structural and biochemical properties of L-asparaginase. *The FEBS Journal* **288**:4183–4209 DOI 10.1111/febs.16042.

Mandlik A, Swierczynski A, Das A, Ton-That H. 2008. Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. *Trends in Microbiology* **16**:33–40 DOI 10.1016/j.tim.2007.10.010.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Research* **21**:863–874 DOI 10.1101/gr.115949.110.

Martino GP, Perez CE, Magni C, Blancato VS. 2018. Implications of the expression of Enterococcus faecalis citrate fermentation genes during infection. *PLOS ONE* **13**:e0205787 DOI 10.1371/journal.pone.0205787.

Matsumoto K, Yamazaki K, Kawakami S, Miyoshi D, Ooi T, Hashimoto S, Taguchi S. 2017. In vivo target exploration of apidaecin based on acquired resistance induced by gene overexpression (ARGO assay). *Scientific Reports* **7**:12136 DOI 10.1038/s41598-017-12039-6.

Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**:60 DOI 10.1186/1471-2105-14-60.

Meier-Kolthoff JP, Göker M. 2019. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature Communications* **10**:2182 DOI 10.1038/s41467-019-10210-3.

Möller J, Busch A, Berens C, Hotzel H, Burkovski A. 2021. Newly isolated animal pathogen corynebacterium silvaticum is cytotoxic to human epithelial cells. *International Journal of Molecular Sciences* **22**:3549 DOI 10.3390/ijms22073549.

Omasits U, Ahrens CH, Müller S, Wollscheid B. 2014. Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* **30**:884–886 DOI 10.1093/bioinformatics/btt607.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**:132 DOI 10.1186/s13059-016-0997-x.

Othman IMM, Gad-Elkareem MAM, Hassane Anouar E, Aouadi K, Snoussi M, Kadri A. 2021. New substituted pyrazolones and dipyrazolotriazines as promising tyrosyl-tRNA synthetase and peroxiredoxin-5 inhibitors: design, synthesis, molecular docking and structure-activity relationship (SAR) analysis. *Bioorganic Chemistry* **109**:104704 DOI 10.1016/j.bioorg.2021.104704.

Otsuji K, Fukuda K, Ogawa M, Saito M. 2019. Mutation and diversity of diphtheria toxin in Corynebacterium ulcerans. *Emerging Infectious Diseases* **25**:2122–2123 DOI 10.3201/eid2511.181455.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**:D733–D745 DOI 10.1093/nar/gkv1189.

Peiro C, Millard P, de Simone A, Cahoreau E, Peyriga L, Enjalbert B, Heux S. 2019. Chemical and metabolic controls on dihydroxyacetone metabolism lead to suboptimal growth of Escherichia coli. *Applied and Environmental Microbiology* **85**:1–17 DOI 10.1128/AEM.00768-19.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in Escherichia coli. *Genome Research* **17**:1336–1343 DOI 10.1101/gr.6254707.

Pillalamarri V, Reddy CG, Bala SC, Jangam A, Kutty VV, Addlagatta A. 2021. Methionine aminopeptidases with short sequence inserts within the catalytic domain are differentially inhibited: Structural and biochemical studies of three proteins from Vibrio spp. *European Journal of Medicinal Chemistry* **209**:112883 DOI 10.1016/j.ejmech.2020.112883.

Raia P, Delarue M, Sauguet L. 2019. An updated structural classification of replicative DNA polymerases. *Biochemical Society Transactions* **47**:239–249 DOI 10.1042/BST20180579.

Rao RT, Sivakumar N, Jayakumar K. 2019. Analyses of livestock-associated staphylococcus aureus pan-genomes suggest virulence is not primary interest in evolution of its genome. *OMICS: A Journal of Integrative Biology* **23**:224–236 DOI 10.1089/omi.2019.0005.

Rappuoli R, Malito E. 2014. History of diphtheria vaccine development. In: Burkovski A, ed. *Corynebacterium Diphtheriae and Related Toxigenic Species*. Dordrecht: Springer Netherlands, 225–238.

Rigden DJ. 2008. The histidine phosphatase superfamily: structure and function. *The Biochemical journal* **409**:333–348 DOI 10.1042/BJ20071097.

Sahm A, Bens M, Platzer M, Szafranski K. 2017. PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. *Nucleic Acids Research* **45**:1–11 DOI 10.1093/nar/gkx179.

Sangal V, Burkovski A, Hunt AC, Edwards B, Blom J, Hoskisson PA. 2014. A lack of genetic basis for biovar differentiation in clinically important Corynebacterium diphtheriae from whole genome sequencing. *Infection, Genetics and Evolution* **21**:54–57 DOI 10.1016/j.meegid.2013.10.019.

Schneider K, Dimroth P, Bott M. 2000. Biosynthesis of the prosthetic group of citrate lyase. *Biochemistry* **39**:9438–9450 DOI 10.1021/bi000401r.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biology and Evolution* **1**:114–118 DOI 10.1093/gbe/evp012.

Scolnick E, Tompkins R, Caskey T, Nirenberg M. 1968. Release factors differing in specificity for terminator codons. *Proceedings of the National Academy of Sciences of the United States of America* **61**:768–774 DOI 10.1073/pnas.61.2.768.

Selim SA, Mohamed FH, Hessain AM, Moussa IM. 2015. Immunological characterization of diphtheria toxin recovered from Corynebacterium pseudotuberculosis. *Saudi Journal of Biological Sciences* **23**:282–287 DOI 10.1016/j.sjbs.2015.11.004.

Sharma NC, Efstratiou A, Mokrousov I, Mutreja A, Das B, Ramamurthy T. 2019. Diphtheria. *Nature Reviews Disease Primers* **5**:81 DOI 10.1038/s41572-019-0131-y.

Shende G, Haldankar H, Barai RS, Bharmal MH, Shetty V, Idicula-Thomas S. 2016. PBIT: pipeline builder for identification of drug targets for infectious diseases. *Bioinformatics* **33(6)**:929–931 DOI 10.1093/bioinformatics/btw760.

Smith JM. 1992. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**:126–129 DOI 10.1007/BF00182389.

Soares SC, Geyik H, Ramos RTJ, de Sá PHCG, Barbosa EGV, Baumbach J, Figueiredo HCP, Miyoshi A, Tauch A, Silva A, Azevedo V. 2016. GIPSy: genomic island prediction software. *Journal of Biotechnology* **232**:2–11 DOI 10.1016/j.jbiotec.2015.09.008.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**:9440–9445 DOI 10.1073/pnas.1530509100.

Timson DJ. 2016. Metabolic enzymes of helminth parasites: potential as drug targets. *Current Protein & Peptide Science* **17**:280–295 DOI 10.2174/1389203717999160226180733.

Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, Husemann P, Stoye J, Dorella FA, Rocha FS, Soares SDC, D'Afonseca V, Miyoshi A, Ruiz J, Silva A, Azevedo V, Burkovski A, Guiso N, Join-Lambert OF, Kayal S, Tauch A. 2010. The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC Genomics* **11**:728 DOI 10.1186/1471-2164-11-728.

Truelove SA, Keegan LT, Moss WJ, Chaisson LH, Macher E, Azman AS, Lessler J. 2020. Clinical and epidemiological aspects of diphtheria: a systematic review and pooled analysis. *Clinical Infectious Diseases* **71**:89–97 DOI 10.1093/cid/ciz808.

Volkamer A, Kuhn D, Rippmann F, Rarey M. 2012. DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics* **28**:2074–2075 DOI 10.1093/bioinformatics/bts310.

Wu C-C, Lin JLJ, Yuan HS. 2020. Structures, mechanisms, and functions of his-me finger nucleases. *Trends in Biochemical Sciences* **45**:935–946 DOI 10.1016/j.tibs.2020.07.002.

Yang Z. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**:1107–1118 DOI 10.1093/molbev/msi097.

Yang Z, Dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution* **28**:1217–1228 DOI 10.1093/molbev/msq303.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**:908–917 DOI 10.1093/oxfordjournals.molbev.a004148.

Zasada AA. 2013. Nontoxigenic highly pathogenic clone of Corynebacterium diphtheriae, Poland, 2004–2012. *Emerging Infectious Diseases* **19**:1870–1872 DOI 10.3201/eid1911.130297.

Zasada AA. 2014. Antimicrobial susceptibility and treatment. In: *Corynebacterium Diphtheriae and Related Toxigenic Species*. Dordrecht: Springer Netherlands, 239–246.

Zhang R. 2004. DEG: a database of essential genes. *Nucleic Acids Research* **32**:271D–272D DOI 10.1093/nar/gkh024.

Zhang J. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* **22**:2472–2479 DOI 10.1093/molbev/msi237.

Zhang Y, Zhang H, Zhou T, Zhong Y, Jin Q. 2011. Genes under positive selection in Mycobacterium tuberculosis. *Computational Biology and Chemistry* **35**:319–322 DOI 10.1016/j.compbiolchem.2011.08.001.

5. DISCUSSÃO GERAL

Na Revisão Bibliográfica da tese apresentei o conhecimento atual sobre a biologia de *C. silvaticum*, incluindo sua classificação taxonômica, patogenicidade e epidemiologia, além dos conceitos de genômica aplicados nas análises dos capítulos.

No capítulo I investigamos a taxonomia da linhagem *C. silvaticum* PO100/5, exploramos a diversidade genômica e buscamos por genes que poderiam ser utilizados como marcadores para a espécie. O genoma da PO100/5 foi primeiro sequenciado desta espécie oriundo de um isolado de animal doméstico e fora da Alemanha ou Áustria. Reclassificamos como *C. silvaticum* a PO100/5 e as linhagens 04-13 e 05-13 (porco selvagem, Áustria) e W25 (veado, Alemanha) que haviam sido depositadas no GenBank como *C. ulcerans*. PO100/5 possui um novo ST (ST709), e juntamente com 05-13, possuem um gene *tox* sem *frameshift*, potencialmente sendo capazes de produzir DT. Outros nove fatores de virulência estão presentes e possivelmente funcionais (*endoE*, *cwlH*, *nanH*, *pld*, *rpfI*, *spaB*, *vsp1*, *vsp2* e DIP0733). Os 38 genomas de *C. silvaticum* analisados possuem um pangenoma de 3002 genes, 73.6% dos genes conservados, um pangenoma perto de ser fechado ($\alpha > 0.952$) e quatro profagos. Em comparação com *C. ulcerans*, existem oito ilhas genômicas e 172 genes que poderiam ser utilizados como marcadores específicos para a espécie. Além disso, *C. silvaticum* foi predita como potencial patógeno para humanos, provavelmente podendo causar zoonose.

No capítulo II adicionamos sete novas linhagens isoladas de porcos domésticos em Portugal à análise genômica, aumentando para oito a amostragem deste país, e incluímos 38 outos genomas, totalizando 46. As linhagens de Portugal são idênticas (ANI > 99.99%), parte de um *cluster* único e exclusivo, monofiléticas e filogeneticamente mais próximas à linhagem 05-13 da Áustria, o que sugere origem neste país. Em relação aos fatores de virulência e nicho, 27 de 36 genes conhecidos em *Corynebacterium* estão presentes nas oito linhagens. Todas as linhagens de Portugal e a 05-13 não possuem o *frameshift* no gene *tox* presente nas demais linhagens, o que sugere que podem produzir a DT. Os *primers* utilizados na técnica de RT-qPCR para a expressão de *tox* não abrangem a região do *frameshift*, portanto pode causar falsos positivos. Os 46 genomas de *C. silvaticum* analisados foram classificados em dois clados. O Clado 1 contém as linhagens de Portugal e a 05-13, que são prováveis produtoras de DT. O Clado 2 contém as demais linhagens. O novo pangenoma está mais próximo de ser fechado ($\alpha = 0.968$). 18 dos 36 fatores de virulência de *Corynebacterium* são parte do genoma central. Encontramos genes exclusivos que podem ser marcadores para o Clado 1 (25), Clado 2 (36), e

para as linhagens de Portugal (19). Estes genes provavelmente não interferem na manifestação da doença em porcos, mas podem estar envolvidos adaptação a ambientes fora do hospedeiro. Devido à predição da espécie como patógeno humano, e demonstração de citotoxicidade em células humanos *in vitro*, e provável produção de DT, *C. silvaticum* tem o potencial de causar zoonose e difteria. A transmissão pode ser ocupacional, como em *C. pseudotuberculosis*.

No capítulo III, foi realizada uma análise de seleção positiva em escala genômica nas seis espécies do complexo da *C. diphtheriae*, para identificar adaptações e sugerir alvos de vacinas e quimioterápicos. Nenhum gene foi identificado em *C. silvaticum*. A adaptação ao nicho provavelmente ocorre pela presença de genes, ao invés de mudanças nas sequências destes. 19 genes foram detectados em *C. diphtheriae* (2), *C. pseudotuberculosis* (10), *C. rouxii* (1), e *C. ulcerans* (6). Destes, 14 são essenciais, seis são fatores de virulência, 13 alvos de quimioterápicos e quatro alvos vacinais.

6. CONCLUSÃO E PERSPECTIVAS

As análises genômicas realizadas contribuíram para o conhecimento sobre a biologia de *C. silvaticum*. Demonstramos que as amostras isoladas de porco doméstico de Portugal são idênticas e monofiléticas, e juntamente com a linhagem 05-13 da Áustria formam um clado que provavelmente produz DT. *C. silvaticum* possui genes que podem ser utilizados como marcadores moleculares para os dois clados detectados e linhagens de Portugal. Nenhum gene codificador de proteína foi identificado como sob seleção positiva. Em relação a espécie mais próxima *C. ulcerans*, *C. silvaticum* é mais homogênea, com um pangenoma menor e mais próximo de ser fechado, e possui oito ilhas genômicas. *C. silvaticum* foi predita como patógeno em humanos, podendo causar zoonose, tornando o porco um possível reservatório para a difteria.

Como perspectivas, esperamos montar genomas completos da espécie para analisar com mais precisão a sintenia e plasticidade genômica da espécie, e testar experimentalmente a produção de DT em linhagens de Portugal e na linhagem 05-13.

REFERÊNCIAS

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

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

ANISIMOVA, M. Parametric models of codon evolution. Em: CANNAROZZI, G. M.; SCHNEIDER, A. (Eds.). **Codon Evolution: Mechanisms and Models**. 1. ed. Oxford: Oxford University Press, 2012. p. 12–33.

ANISIMOVA, M.; LIBERLES, D. A. Detecting and understanding natural selection. Em: CANNAROZZI, G. M.; SCHNEIDER, A. (Eds.). **Codon Evolution: Mechanisms and Models**. 1. ed. Oxford: Oxford University Press, 2012. p. 73–96.

AUCH, A. F.; KLENK, H. P.; GÖKER, M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. **Standards in Genomic Sciences**, v. 2, n. 1, p. 142–148, 2010.

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

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

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

BAUER, K. A. et al. Review of rapid diagnostic tests used by antimicrobial stewardship programs. **Clinical infectious diseases : an official publication of the Infectious Diseases Society of America**, v. 59, n. Suppl 3, p. S134–S145, 2014.

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

BENSON, D. A. GenBank. **Nucleic Acids Research**, v. 33, n. Database issue, p. D34–D38, 17 dez. 2004.

BERGER, A. et al. Two Cases of Cutaneous Diphtheria Associated with Occupational Pig Contact in Germany. **Zoonoses and Public Health**, v. 60, n. 8, p. 539–542, 2013.

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

BUSCH, A. et al. Genome sequence of a pathogenic *Corynebacterium ulcerans* strain isolated from a wild boar with necrotizing lymphadenitis. **BMC Research Notes**, v. 12, n. 1, p. 10–12, 2019.

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

CHAN, C. X.; RAGAN, M. A. Next-generation phylogenomics. p. 1–6, 2013.

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

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

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

CLARK, A. E. et al. Matrix-Assisted laser desorption ionization-time of flight mass spectrometry: A fundamental shift in the routine practice of clinical microbiology. **Clinical Microbiology Reviews**, v. 26, n. 3, p. 547–603, 2013.

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

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

DANGEL, A. et al. NGS-based phylogeny of diphtheria-related pathogenicity factors in different *Corynebacterium* spp. implies species-specific virulence transmission. **BMC Microbiology**, v. 19, n. 1, p. 1–16, 2019.

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

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

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

EISENBERG, T. et al. Nontoxigenic tox-bearing *Corynebacterium ulcerans* Infection among Game Animals, Germany. **Emerging Infectious Diseases**, v. 20, n. 3, p. 448–452, mar. 2014.

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

HACKER, E. et al. *Corynebacterium ulcerans*, an emerging human pathogen. **Future Microbiology**, v. 11, n. 9, p. 1191–1208, set. 2016.

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

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

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

KONIG, C. et al. Multilocus Sequence Typing of *Corynebacterium ulcerans* Provides Evidence for Zoonotic Transmission and for Increased Prevalence of Certain Sequence Types among Toxigenic Strains. **Journal of Clinical Microbiology**, v. 52, n. 12, p. 4318–4324, 1 dez. 2014.

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

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

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

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

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

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

MEIER-KOLTHOFF, J. P.; GÖKER, M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. **Nature Communications**, v. 10, n. 1, p. 2182, 16 dez. 2019.

MENG, X. J.; LINDSAY, D. S.; SRIRANGANATHAN, N. Wild boars as sources for infectious diseases in livestock and humans. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 364, n. 1530, p. 2697–2707, 27 set. 2009.

MÖLLER, J. et al. Phylogenomic characterisation of a novel corynebacterial species pathogenic to animals. **Antonie van Leeuwenhoek**, v. 113, n. 8, p. 1225–1239, 4 ago. 2020.

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

NOVAIS, Â. et al. Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing. **European Journal of Clinical Microbiology & Infectious Diseases**, v. 38, n. 3, p. 427–448, 27 mar. 2019.

OGATA, H. et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research**, v. 27, n. 1, p. 29–34, 1 jan. 1999.

OLIVEIRA, M. et al. First report of *Corynebacterium pseudotuberculosis* from caseous lymphadenitis lesions in Black Alentejano pig (Sus scrofa domesticus). **BMC Veterinary Research**, v. 10, n. 1, p. 218, dez. 2014.

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

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

RAMOS, C. P. et al. Molecular characterization of *Corynebacterium pseudotuberculosis*, *C. silvaticum*, and *C. auriscanis* by ERIC-PCR. **Ciência Rural**, v. 52, n. 11, 2022.

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

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

RAU, J. et al. Reliable differentiation of a non-toxigenic tox gene-bearing *Corynebacterium ulcerans* variant frequently isolated from game animals using MALDI-TOF MS. **Veterinary Microbiology**, v. 237, n. August, p. 108399, 2019.

SCHUHEGGER, R. et al. Pigs as Source for Toxigenic *Corynebacterium ulcerans*. **Emerging Infectious Diseases**, v. 15, n. 8, p. 1314–1315, ago. 2009.

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

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

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

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

TOFT, C.; ANDERSSON, S. G. E. Evolutionary microbial genomics: Insights into bacterial host adaptation. **Nature Reviews Genetics**, v. 11, n. 7, p. 465–475, 2010.

WASMUTH, E. V et al. UniProt: The universal protein knowledgebase. **Nucleic Acids Research**, v. 45, n. D1, p. D158–D169, 4 jan. 2017.

WINDSOR, P. A. Control of Caseous Lymphadenitis. **Veterinary Clinics of North America: Food Animal Practice**, v. 27, n. 1, p. 193–202, mar. 2011.

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

APÊNDICE

APÊNDICE A – Linhas de comando

1. FastANI

../fastANI -t 8 --ql all_vs_all2.txt --rl all_vs_all2.txt -o all_vs_all_matrix2.out --matrix

2. MLSTcheck

2.1. Baixando a base de dados de MLST

sudo docker run --rm -it -v /home/marcus/softwares/mlstcheck/Cs/:/data sangerpathogens/mlst_check download_mlst_databases

2.2. Linha de comando do MLSTcheck

sudo docker run --rm -it -v /home/marcus/softwares/mlstcheck/Cs/:/data sangerpathogens/mlst_check get_sequence_type -d 8 -c -s 'Corynebacterium diphtheriae' *.fasta

2.3. Mudando permissões dos arquivos de saída

sudo chmod 777 mlst_results.allele.csv mlst_results.genomic.csv concatenated_alleles.fa

3. OrhoFinder e scripts *in house*

3.1.OrthoFinder

Aumentando o número de arquivos que podem ser abertos pelo sistema

sudo su
ulimit -n 500000

Rodando o OrthoFinder como *root*

./orthofinder -t 8 -f Cs_38

./orthofinder -t 8 -f Cul_76

./orthofinder -t 8 -f CsCul_114

Mudando permissões dos arquivos de saída

sudo chmod -R 777 *

3.2.Scripts *in house*

Obtendo a lista de genes do pangenoma, genoma core, acessório, compartilhado e singletons

perl orthofinder_pangenome_splitter.pl Orthogroups.tsv Orthogroups_UnassignedGenes.tsv

cat shared.txt singletons.txt > accessory.txt

perl first_non_empty_column_per_line.pl accessory.txt

perl -pi -e "s/\,.+\n/\n/g" first_sample_of_each_orthogroup.txt

cat core.txt shared.txt singletons.txt > pangenome.txt

perl pandev.pl pangenome.txt > pandev_results.txt

Obtendo uma proteína para representar cada ortogrupo

perl cat *.faa > allproteins.faa

perl list2ffnfaa.pl first_non_empty_column_per_line.txt allproteins.faa > accessory.faa

4. Vacinologia reversa

4.1.Preparação do arquivo contendo as proteínas *core* representadas no genoma referência KL0182^{T}

perl -pi -e "s/\,.+\n/\n/g" core_KL0182.txt

perl -pi -e "s/\"//g" core_KL0182.txt

perl list2ffnfaa.pl core_KL0182.txt allproteins.faa > core_KL0182.faa

4.2.BLASTp contra o genoma do hospedeiro

formatdb -p T -i Sscrofa11.1.faa

blastall -p blastp -i core_KL0182.faa -d Sscrofa11.1.faa -o core-vs-pig.out.tab -F F -m 8

cp core-vs-pig.out.tab core-vs-pig.list

perl -pi -e "s/\t.+\n/\n/g" core-vs-pig.list

perl get_non_host_homolog.pl core-vs-pig.list core_KL0182.faa > core-non_pig_host.faa

4.3.Obtendo um arquivo multifasta de proteínas a partir da lista das proteínas essenciais

perl list2ffnfaa.pl core_non_pig_host_essential.txt core-non_pig_host.faa > core-non_pig_host_essential.faa

4.4.SurfG+

surfg.sh -g pos -i core-non_pig_host.faa

perl surfg2files.pl out.txt

perl list2ffnfaa.pl vaccine_targets.list core-non_host.faa > vaccine_targets.faa

perl list2ffnfaa.pl drug_targets.list core-non_host.faa > drug_targets.faa

APÊNDICE B – Artigos em primeira ou última autoria

Artigo de pesquisa 1

Referência bibliográfica: ALMEIDA, M. O. ; CARVALHO, R. D. O. ; ABURJAILE, F. F. ; MIRANDA, F. ; CERQUEIRA, J. C. ; BRENIG, B. ; GHOSH, P. ; RAMOS, ROMMEL ; KATO, R. B. ; SOARES, S. C. ; SILVA, A. ; AZEVEDO, V. ; **VIANA, M. V. C.** .

Characterization of the first vaginal *Lactobacillus crispatus* genomes isolated in Brazil. PeerJ, v. 9, p. e11079, 2021. https://doi.org/10.7717/peerj.11079.

Contribuição: Coorientação, análises genômicas, escrita e revisão do manuscrito.

Abstract: *Lactobacillus crispatus* is the dominant species in the vaginal microbiota associated with health and considered a homeostasis biomarker. Interestingly, some strains are even used as probiotics. However, the genetic mechanisms of *L. crispatus* involved in the control of the vaginal microbiome and protection against bacterial vaginosis (BV) are not entirely known. To further investigate these mechanisms, we sequenced and characterized the first four *L. crispatus* genomes from vaginal samples from Brazilian women and used genome-wide association study (GWAS) and comparative analyses to identify genetic mechanisms involved in healthy or BV conditions and selective pressures acting in the vaginal microbiome. The four genomes were sequenced, assembled using ten different strategies and automatically annotated. The functional characterization was performed by bioinformatics tools comparing with known probiotic strains. Moreover, it was selected one representative strain (*L. crispatus* CRI4) for in vitro detection of phages by electron microscopy. Evolutionary analysis, including phylogeny, GWAS and positive selection were performed using 46 public genomes strains representing health and BV conditions. Genes involved in probiotic effects such as lactic acid production, hydrogen peroxide, bacteriocins, and adhesin were identified. Three hemolysins and putrescine production were predicted, although these features are also present in other probiotic strains. The four genomes presented no plasmids, but 14 known families insertion sequences and several prophages were detected. However, none of the mobile genetic elements contained antimicrobial resistance genes. The genomes harbor a CRISPR-Cas subtype II-A system that is probably inactivated due to fragmentation of the genes csn2 and cas9. No genomic feature was

associated with a health condition, perhaps due to its multifactorial characteristic. Five genes were identified as under positive selection, but the selective pressure remains to be discovered. In conclusion, the Brazilian strains investigated in this study present potential protective properties, although in vitro and in vivo studies are required to confirm their efficacy and safety to be considered for human use.

Artigo de pesquisa 2

Referência bibliográfica: VIANA, M. V. C.; SAHM, A. ; GOES NETO, A. ; FIGUEIREDO, HENRIQUE CÉSAR PEREIRA ; WATTAM, A. R. ; AZEVEDO, V.. **Rapidly evolving changes and gene loss associated with host switching in *Corynebacterium pseudotuberculosis*.** Plos One, v. 13, p. e0207304, 2018. https://doi.org/10.1371/journal.pone.0207304.

Contribuição: Análises genômica, escrita e revisão do manuscrito.

Abstract: Phylogenomics and genome scale positive selection analyses were performed on 29 *Corynebacterium pseudotuberculosis* genomes that were isolated from different hosts, including representatives of the Ovis and Equi biovars. A total of 27 genes were identified as undergoing adaptive changes. An analysis of the clades within this species and these biovars, the genes specific to each branch, and the genes responding to selective pressure show clear differences, indicating that adaptation and specialization is occurring in different clades. These changes are often correlated with the isolation host but could indicate responses to some undetermined factor in the respective niches. The fact that some of these more-rapidly evolving genes have homology to known virulence factors, antimicrobial resistance genes and drug targets shows that this type of analysis could be used to identify novel targets, and that these could be used as a way to control this pathogen.

Artigo de revisão 1

Referência bibliográfica: ALMEIDA, M.O. ; CARMO, F.L.R. DO ; GALA-GARCÍA, A. ; KATO, R. ; GOMIDE, A.C. ; DRUMMOND, R.M.N. ; DRUMOND, M.M. ; AGRESTI, P.M.

; BARH, D. ; BRENING, B. ; GHOSH, P. ; SILVA, A. ; AZEVEDO, V. ; **VIANA, M.V.C.**. ***Lactobacillus crispatus*** **protects against bacterial vaginosis.** Genetics and Molecular Research, v. 18, p. gmr18475, 2019. https://doi.org/10.4238/gmr18475.

Contribuição: Escrita e revisão do manuscrito.

Abstract: In medicine, the 20th century was marked by one of the most important revolutions in infectious-disease management, the discovery and increasing use of antibiotics. However, their indiscriminate use has led to the emergence of multidrug-resistant (MDR) bacteria. Drug resistance and other factors, such as the production of bacterial biofilms, have resulted in high recurrence rates of bacterial diseases. Bacterial vaginosis (BV) syndrome is the most prevalent vaginal condition in women of reproductive age, leading to considerable discomfort. BV can be a consequence of gynecological and obstetric complications, as well as sexually transmitted diseases. Given the decrease in efficiency of antibiotic therapy and high rates of recurrence, probiotics have become promising alternatives for both prevention and treatment of BV, or as an adjuvant to conventional therapy. Currently, *Lactobacillus* species are the most extensively studied for use as probiotics. Probiotics act through stimulation of the host immune system, competitive exclusion and antimicrobial activity; the latter involves production of substances such as lactic acid, hydrogen peroxide and bacteriocins. *Lactobacillus crispatus* is considered to be a biomarker of a healthy vaginal tract and is indicated for a probiotic approach to maintaining and restoring of a healthy vaginal ecosystem. Some *L. crispatus* probiotic strains are already commercially available with encouraging results; however, control of BV syndrome still presents many challenges.

APÊNDICE C – Artigos em colaboração

Artigo de pesquisa 3

Referência bibliográfica: VALDEZ-BAEZ, J. ; DA COSTA, F. M. R. ; PINTO GOMIDE, A. CYBELLE ; PROFETA, R. ; DA SILVA, A. L. ; SOUSA, T. J. ; **VIANA, M. V. C.** ; BENTES KATO, R. ; AMERICO, M. . ; DOS SANTOS FREITAS, A. ; CARVALHO, R. D. O. ; BRENIG, B. ; MARTINS, F. S. ; ABURJAILE, F. ; AZEVEDO, V. . **Comparative Genomics and In Silico Evaluation of Genes Related to the Probiotic Potential of *Bifidobacterium breve* 1101A.** Bacteria, v. 1, p. 161-182, 2022. https://doi.org/10.1016/j.gene.2021.145781.

Contribuição: Supervisão, escrita e revisão do manuscrito.

Abstract: The *Bifidobacterium longum* 51A strain of isolated from feces of a healthy child, has demonstrated probiotic properties by in vivo and in vitro studies, which may be assigned to its production of metabolites such as acetate. Thus, through the study of comparative genomics, the present work sought to identify unique genes that might be related to the production of acetate. To perform the study, the DNA strain was sequenced using Illumina HiSeq technology, followed by assembly and manual curation of coding sequences. Comparative analysis was performed including 19 complete *B. longum* genomes available in Genbank/NCBI. In the phylogenetic analysis, the CECT 7210 and 157F strains of *B. longum* subsp. *infantis* aggregated within the subsp. *longum* cluster, suggesting that their taxonomic classification should be reviewed. The strain 51A of *B. longum* has 26 unique genes, six of which are possibly related to carbohydrate metabolism and acetate production. The phosphoketolase pathway from *B. longum* 51A showed a difference in acetyl-phosphate production. This result seems to corroborate the analysis of their unique genes, whose presence suggests the strain may use different sources of carbohydrates that allow a greater production of acetate and consequently offer benefits to the host health.

Artigo de pesquisa 4

Referência bibliográfica: HURTADO, R. ; BARH, D. ; WEIMER, B. C. ; **VIANA, M. V. C.** ; PROFETA, R. ; SOUSA, T. J. ; ABURJAILE, F. F. ; QUINO, W. ; SOUZA, R. P. ; MESTANZA, O. ; GAVILÁN, R. G. ; AZEVEDO, V.. **WGS-Based Lineage and Antimicrobial Resistance Pattern of *Salmonella* Typhimurium Isolated during 2000-2017 in Peru.** ANTIBIOTICS-BASEL, v. 11, p. 1170, 2022. https://doi.org/10.3390/antibiotics11091170.

Contribuição: Coorientação, escrita e revisão do manuscrito.

Abstract: *Salmonella* Typhimurium is associated with foodborne diseases worldwide, including in Peru, and its emerging antibiotic resistance (AMR) is now a global public health problem. Therefore, country-specific monitoring of the AMR emergence is vital to control this pathogen, and in these aspects, whole genome sequence (WGS)-based approaches are better than gene-based analyses. Here, we performed the antimicrobial susceptibility test for ten widely used antibiotics and WGS-based various analyses of 90 *S.* Typhimurium isolates (human, animal, and environment) from 14 cities of Peru isolated from 2000 to 2017 to understand the lineage and antimicrobial resistance pattern of this pathogen in Peru. Our results suggest that the Peruvian isolates are of Typhimurium serovar and predominantly belong to sequence type ST19. Genomic diversity analyses indicate an open pan-genome, and at least ten lineages are circulating in Peru. A total of 48.8% and 31.0% of isolates are phenotypically and genotypically resistant to at least one antibiotic, while 12.0% are multi-drug resistant (MDR). Genotype–phenotype correlations for ten tested drugs show >80% accuracy, and >90% specificity. Sensitivity above 90% was only achieved for ciprofloxacin and ceftazidime. Two lineages exhibit the majority of the MDR isolates. A total of 63 different AMR genes are detected, of which 30 are found in 17 different plasmids. Transmissible plasmids such as lncI-gamma/k, IncI1-I(Alpha), Col(pHAD28), IncFIB, IncHI2, and lncI2 that carry AMR genes associated with third-generation antibiotics are also identified. Finally, three new non-synonymous single nucleotide variations (SNVs) for nalidixic acid and eight new SNVs for nitrofurantoin resistance are predicted using genome-wide association studies, comparative genomics, and functional annotation. Our analysis provides for the first time the WGS-based

details of the circulating *S.* Typhimurium lineages and their antimicrobial resistance pattern in Peru.

Artigo de pesquisa 5

Referência bibliográfica: CARVALHO, R. ; ABURJAILE, F. ; **CANARIO, M.** ; NASCIMENTO, A. M. A. ; CHARTONE-SOUZA, E. ; DE JESUS, L. ; ZAMYATNIN, A. A. ; BRENIG, B. ; BARH, D. ; GHOSH, P. ; GOES-NETO, A. ; FIGUEIREDO, H. C. P. ; SOARES, S. ; RAMOS, R. ; PINTO, A. ; AZEVEDO, V.. **Genomic characterization of multidrug-resistant *Escherichia coli* BH100.** Frontiers in Microbiology, v. 11, p. 1-13, 2021. https://doi.org/10.3389/fmicb.2020.549254.

Contribuição: Análise filogenômica, escrita e revisão do manuscrito.

Abstract: The rapid emergence of multidrug-resistant (MDR) bacteria is a global health problem. Mobile genetic elements like conjugative plasmids, transposons, and integrons are the major players in spreading resistance genes in uropathogenic *Escherichia coli* (UPEC) pathotype. The *E. coli* BH100 strain was isolated from the urinary tract of a Brazilian woman in 1974. This strain presents two plasmids carrying MDR cassettes, pBH100, and pAp, with conjugative and mobilization properties, respectively. However, its transposable elements have not been characterized. In this study, we attempted to unravel the factors involved in the mobilization of virulence and drug-resistance genes by assessing genomic rearrangements in four BH100 sub-strains (BH100 MG2014, BH100 MG2017, BH100L MG2017, and BH100N MG2017). Therefore, the complete genomes of the BH100 sub-strains were achieved through Next Generation Sequencing and submitted to comparative genomic analyses. Our data shows recombination events between the two plasmids in the sub-strain BH100 MG2017 and between pBH100 and the chromosome in BH100L MG2017. In both cases, IS3 and IS21 elements were detected upstream of Tn21 family transposons associated with MDR genes at the recombined region. These results integrated with Genomic island analysis suggest pBH100 might be involved in the spreading of drug resistance through the formation of resistance islands. Regarding pathogenicity, our results reveal that BH100 strain is closely related to UPEC strains and contains many IS3 and IS21-transposase-enriched genomic islands associated with

virulence. This study concludes that those IS elements are vital for the evolution and adaptation of BH100 strain.

Artigo de pesquisa 6

Referência bibliográfica: GABRIELLE VIDAL DA SILVA, J. ; THOMAZ VIEIRA, A. ; SOUSA, T. J ; **VINICIUS CANÁRIO VIANA, M.** ; PARISE, D. ; SAMPAIO, B. ; LIMA DA SILVA, A. ; CLÁUDIO LIMA DE JESUS, L. ; KÁSSIO RIBEIRO MATOS LOUREIRO DE CARVALHO, P. ; DE CASTRO OLIVEIRA, L. ; FIGUEIRA ABURJAILE, F. ; MARTINS, F.; ROBERT NICOLI, J. ; GHOSH, P. ; BRENIG, B. ; AZEVEDO, V. ; CYBELLE PINTO GOMIDE, A. . **Comparative genomics and in silico gene evaluation involved in the probiotic potential of *Bifidobacterium longum* 51A.** GENE, v. 795, p. 145781, 2021. https://doi.org/10.1016/j.gene.2021.145781.

Contribuição: Escrita e revisão do manuscrito.

Abstract: The *Bifidobacterium longum* 51A strain of isolated from feces of a healthy child, has demonstrated probiotic properties by in vivo and in vitro studies, which may be assigned to its production of metabolites such as acetate. Thus, through the study of comparative genomics, the present work sought to identify unique genes that might be related to the production of acetate. To perform the study, the DNA strain was sequenced using Illumina HiSeq technology, followed by assembly and manual curation of coding sequences. Comparative analysis was performed including 19 complete *B. longum* genomes available in Genbank/NCBI. In the phylogenetic analysis, the CECT 7210 and 157F strains of *B. longum* subsp. infantis aggregated within the subsp. longum cluster, suggesting that their taxonomic classification should be reviewed. The strain 51A of *B. longum* has 26 unique genes, six of which are possibly related to carbohydrate metabolism and acetate production. The phosphoketolase pathway from *B. longum* 51A showed a difference in acetyl-phosphate production. This result seems to corroborate the analysis of their unique genes, whose presence suggests the strain may use different sources of carbohydrates that allow a greater production of acetate and consequently offer benefits to the host health.

Artigo de pesquisa 7

Referência bibliográfica: FERREIRA, L. C. ; MAUL, J. E. ; **VIANA, M. V. C.** ; DE SOUSA, T. J. ; DE CARVALHO AZEVEDO, V. A. ; ROBERTS, D. P. ; DE SOUZA, J. T.. **Complete genome sequence of the biocontrol agent *Serratia marcescens* strain N4-5 uncovers an assembly artefact.** Bacterial Fungal and Virus Molecular Biology, v. 52, p. 245-250, 2021. https://doi.org/10.1007/s42770-020-00382-2.

Contribuição: A montagem de genoma, filogenia do gene 16S, escrita e revisão do manuscrito.

Abstract: *Serratia marcescens* are gram-negative bacteria found in several environmental niches, including the plant rhizosphere and patients in hospitals. Here, we present the genome of *Serratia marcescens* strain N4–5 (=NRRL B-65519), which has a size of 5,074,473 bp (664-fold coverage) and contains 4840 protein coding genes, 21 RNA genes, and an average G + C content of 59.7%. N4–5 harbours a plasmid of 11,089 bp and 43.5% G + C content that encodes six unique CDS repeated 2.5× times totalling 13 CDS. Our genome assembly and manual curation uncovered the insertion of two extra copies of the 5S rRNA gene in the assembled sequence, which was confirmed by PCR and Sanger sequencing to be a misassembly. This artefact was subsequently removed from the final assembly. The occurrence of extra copies of the 5S rRNA gene was also observed in most complete genomes of *Serratia* spp. deposited in public databases in our comparative analysis. These elements, which also occur naturally, can easily be confused with true genetic variation. Efforts to discover and correct assembly artefacts should be made in order to generate genome sequences that represent the biological truth underlying the studied organism. We present the genome of N4–5 and discuss genes potentially involved in biological control activity against plant pathogens and also the possible mechanisms responsible for the artefact we observed in our initial assembly. This report raises awareness about the extra copies of the 5S rRNA gene in sequenced bacterial genomes as they may represent misassemblies and therefore should be verified experimentally.

Artigo de pesquisa 8

Referência bibliográfica: MORAIS-RODRIGUES, F. ; SILV'ERIO-MACHADO, R. ; KATO, R. B. ; RODRIGUES, D. L. N. ; VALDEZ-BAEZ, J. ; FONSECA, V. ; SAN, E. J. ;

GOMES, L. G. R. ; DOS SANTOS, R. G. ; **VINICIUS CANÁRIO VIANA, M.** ; DA CRUZ FERRAZ DUTRA, J. ; TEIXEIRA DORNELLES PARISE, M. ; PARISE, D. ; CAMPOS, F. F. ; DE SOUZA, S. J. ; ORTEGA, J. M. ; BARH, D. ; GHOSH, P. ; AZEVEDO, V. A. C. ; DOS SANTOS, M. A. . **Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression.** GENE, v. 726, p. 144168, 2020. https://doi.org/10.1016/j.gene.2019.144168.

Contribuição: Escrita e revisão do manuscrito.

Abstract: Methods based around statistics and linear algebra have been increasingly used in attempts to address emerging questions in microarray literature. Microarray technology is a long-used tool in the global analysis of gene expression, allowing for the simultaneous investigation of hundreds or thousands of genes in a sample. It is characterized by a low sample size and a large feature number created a non-square matrix, and by the incomplete rank, that can generate countless more solution in classifiers. To avoid the problem of the 'curse of dimensionality' many authors have performed feature selection or reduced the size of data matrix. In this work, we introduce a new logistic regression-based model to classify breast cancer tumor samples based on microarray expression data, including all features of gene expression and without reducing the microarray data matrix. If the user still deems it necessary to perform feature reduction, it can be done after the application of the methodology, still maintaining a good classification. This methodology allowed the correct classification of breast cancer sample data sets from Gene Expression Omnibus (GEO) data series GSE65194, GSE20711, and GSE25055, which contain the microarray data of said breast cancer samples. Classification had a minimum performance of 80% (sensitivity and specificity), and explored all possible data combinations, including breast cancer subtypes. This methodology highlighted genes not yet studied in breast cancer, some of which have been observed in Gene Regulatory Networks (GRNs). In this work we examine the patterns and features of a GRN composed of transcription factors (TFs) in MCF-7 breast cancer cell lines, providing valuable information regarding breast cancer. In particular, some genes whose $\alpha i *$ associated parameter values revealed extreme positive and negative values, and, as such, can be identified as breast cancer prediction genes. We indicate that the PKN2, MKL1, MED23, CUL5 and GLI genes demonstrate a tumor suppressor profile, and that the MTR, ITGA2B, TELO2, MRPL9, MTTL1, WIPI1, KLHL20, PI4KB, FOLR1 and SHC1 genes demonstrate an oncogenic profile.

We propose that these may serve as potential breast cancer prediction genes, and should be prioritized for further clinical studies on breast cancer. This new model allows for the assignment of values to the αi ∗ parameters associated with gene expression. It was noted that some αi ∗ parameters are associated with genes previously described as breast cancer biomarkers, as well as other genes not yet studied in relation to this disease.

Artigo de pesquisa 9

Referência bibliográfica: PROFETA, R. ; SEYFFERT, N. ; TIWARI, S. ; **VIANA, MARCUS V. C.** ; JAISWAL, A. K. ; CAETANO, A. C. ; BÜCKER, D. H. ; TAVARES DE OLIVEIRA, L. ; SANTOS, R. ; GALA-GARCIA, A. ; KATO, R. B. ; PADILHA, F. F. ; LIMA-VERDE, I. B. ; GHOSH, P. ; BARH, D. ; GÓES-NETO, A. ; FIGUEIREDO, HENRIQUE C. P. ; SOARES, S. C. ; MEYER, R. ; BRENIG, B. ; RAMOS, P. I. P. ; AZEVEDO, V. ; CASTRO, T. L. P. . **Comparative genomics with a multidrug-resistant *Klebsiella pneumoniae* isolate reveals the panorama of unexplored diversity in Northeast Brazil.** GENE, v. 1, p. 145386, 2020. https://doi.org/10.1016/j.gene.2020.145386.

Contribuição: Supervisão, escrita e revisão do manuscrito.

Abstract: The emergence of community acquired infections increases the public health concern on *K. pneumoniae* and closely related bacteria among which antimicrobial resistance spreads. We report a multidrug-resistant *K. pneumoniae* isolate, B31, of a patient infected in the community and admitted to an intensive care unit in Northeast Brazil. Antimicrobial susceptibility and genome information were thoroughly investigated to characterize B31 in front of 172 sequenced strains of different countries. Assigned to the Sequence Type 15, which is globally spread, B31 presented extended spectrum beta-lactamase, tigecycline and ciprofloxacin resistance. Genome sequencing revealed most resistance genes being carried by plasmids with high dissemination potential. The absence of main virulence factors, like yersiniabactin and colibactin, apparently suggests a mild pathogenic strain which, on the contrary, persisted and caused severe infection in a previously healthy patient. The present study contributes to unveil the unclear genomic scenario of virulent and multidrug-resistant *K. pneumoniae* in Brazil.

Artigo de pesquisa 10

Referência bibliográfica: HURTADO, R. ; CARHUARICRA, D. ; SOARES, SIOMAR C. ; **VIANA, M. V. C.** ; AZEVEDO, V. ; MATURRANO, L. ; ABURJAILE, FLAVIA . **Pan-genomic approach shows insight of genetic divergence and pathogenic-adaptation of *Pasteurella multocida*.** GENE, v. 670, p. 193-206, 2018. https://doi.org/10.1016/j.gene.2018.05.084.

Contribuição: Análise filogenética, escrita e revisão do manuscrito.

Abstract: *Pasteurella multocida* is a gram-negative, non-motile bacterial pathogen, which is associated with chronic and acute infections as snuffles, pneumonia, atrophic rhinitis, fowl cholera and hemorrhagic septicemia. These diseases affect a wide range of domestic animals, leading to significant morbidity and mortality and causing significant economic losses worldwide. Due to the interest in deciphering the genetic diversity and process adaptive between *P. multocida* strains, this work aimed was to perform a pan-genome analysis to evidence horizontal gene transfer and positive selection among 23 *P. multocida* strains isolated from distinct diseases and hosts. The results revealed an open pan-genome containing 3585 genes and an accessory genome presenting 1200 genes. The phylogenomic analysis based on the presence/absence of genes and islands exhibit high levels of plasticity, which reflects a high intraspecific diversity and a possible adaptive mechanism responsible for the specific disease manifestation between the established groups (pneumonia, fowl cholera, hemorrhagic septicemia and snuffles). Additionally, we identified differences in accessory genes among groups, which are involved in sugar metabolism and transport systems, virulence-related genes and a high concentration of hypothetical proteins. However, there was no specific indispensable functional mechanism to decisively correlate the presence of genes and their adaptation to a specific host/disease. Also, positive selection was found only for two genes from sub-group hemorrhagic septicemia, serotype B. This comprehensive comparative genome analysis will provide new insights of horizontal gene transfers that play an essential role in the diversification and adaptation mechanism into *P. multocida* species to a specific disease.

Artigo de pesquisa 11

Referência bibliográfica: DO CARMO, FILLIPE L. R. ; SILVA, W. M. ; TAVARES, GUILHERME C. ; IBRAIM, IZABELA C. ; CORDEIRO, BARBARA F. ; OLIVEIRA, EMILIANO R. ; RABAH, HOUEM ; CAUTY, CHANTAL ; DA SILVA, SARA H. ; **CANÁRIO VIANA, MARCUS V.** ; CAETANO, ANA C. B. ; DOS SANTOS, ROSELANE G. ; DE OLIVEIRA CARVALHO, RODRIGO D. ; JARDIN, JULIEN ; PEREIRA, FELIPE L. ; FOLADOR, EDSON L. ; LE LOIR, YVES ; FIGUEIREDO, HENRIQUE C. P. ; JAN, GWÉNAËL ; AZEVEDO, VASCO . **Mutation of the Surface Layer Protein SlpB Has Pleiotropic Effects in the Probiotic *Propionibacterium freudenreichii* CIRM-BIA 129.** Frontiers in Microbiology, v. 9, p. 1-22, 2018. https://doi.org/10.3389/fmicb.2018.01807.

Contribuição: Montagem de genoma, escrita e revisão do manuscrito.

Abstract: *Propionibacterium freudenreichii* is a beneficial Gram-positive bacterium, traditionally used as a cheese-ripening starter, and currently considered as an emerging probiotic. As an example, the *P. freudenreichii* CIRM-BIA 129 strain recently revealed promising immunomodulatory properties. Its consumption accordingly exerts healing effects in different animal models of colitis, suggesting a potent role in the context of inflammatory bowel diseases. This anti-inflammatory effect depends on surface layer proteins (SLPs). SLPs may be involved in key functions in probiotics, such as persistence within the gut, adhesion to host cells and mucus, or immunomodulation. Several SLPs coexist in *P. freudenreichii* CIRM-BIA 129 and mediate immunomodulation and adhesion. A mutant *P. freudenreichii* CIRM-BIA 129ΔslpB (CB129ΔslpB) strain was shown to exhibit decreased adhesion to intestinal epithelial cells. In the present study, we thoroughly analyzed the impact of this mutation on cellular properties. Firstly, we investigated alterations of surface properties in CB129ΔslpB. Surface extractable proteins, surface charges (ζ-potential) and surface hydrophobicity were affected by the mutation. Whole-cell proteomics, using high definition mass spectrometry, identified 1,288 quantifiable proteins in the wild-type strain, i.e., 53% of the theoretical proteome predicted according to *P. freudenreichii* CIRM-BIA 129 genome sequence. In the mutant strain, we detected 1,252 proteins, including 1,227 proteins in common with the wild-type strain. Comparative quantitative analysis revealed 97 proteins with significant differences between wild-type and mutant strains. These proteins are involved in various cellular process like

signaling, metabolism, and DNA repair and replication. Finally, in silico analysis predicted that slpB gene is not part of an operon, thus not affecting the downstream genes after gene knockout. This study, in accordance with the various roles attributed in the literature to SLPs, revealed a pleiotropic effect of a single slpB mutation, in the probiotic *P. freudenreichii*. This suggests that SlpB may be at a central node of cellular processes and confirms that both nature and amount of SLPs, which are highly variable within the *P. freudenreichii* species, determine the probiotic abilities of strains.

Artigo de pesquisa 12

Referência bibliográfica: PARISE, D. ; PARISE, M. T. D. ; **VIANA, MARCUS V. C.** ; MUÑOZ-BUCIO, A. V. ; CORTÉS-PÉREZ, Y. A. ; ARELLANO-REYNOSO, B. ; DÍAZ-APARICIO, E. ; DORELLA, F. A. ; PEREIRA, F. L. ; CARVALHO, A. F. ; FIGUEIREDO, H. C. P. ; GHOSH, P. ; BARH, D. ; GOMIDE, A. C. P. ; AZEVEDO, V. A. C. . **First genome sequencing and comparative analyses of *Corynebacterium pseudotuberculosis strains* from Mexico.** Standards in Genomic Sciences, v. 13, p. 21, 2018. https://doi.org/10.1186/s40793-018-0325-z.

Contribuição: Supervisão, escrita e revisão do manuscrito.

Abstract: *Corynebacterium pseudotuberculosis* is a pathogenic bacterium which has been rapidly spreading all over the world, causing economic losses in the agricultural sector and sporadically infecting humans. Six *C. pseudotuberculosis* strains were isolated from goats, sheep, and horses with distinct abscess locations. For the first time, Mexican genomes of this bacterium were sequenced and studied in silico. All strains were sequenced using Ion Personal Genome Machine sequencer, assembled using Newbler and SPAdes software. The automatic genome annotation was done using the software RAST and in-house scripts for transference, followed by manual curation using Artemis software and BLAST against NCBI and UniProt databases. The six genomes are publicly available in NCBI database. The analysis of nucleotide sequence similarity and the generated phylogenetic tree led to the observation that the Mexican strains are more similar between strains from the same host, but the genetic structure is probably more influenced by transportation of animals between farms than host preference. Also, a putative drug target was predicted and in silico analysis of 46 strains showed two gene clusters

capable of differentiating the biovars equi and ovis: Restriction Modification system and CRISPR-Cas cluster.

Artigo de revisão 2

Referência bibliográfica: ALMEIDA, M. O. ; **VIANA, MARCUS** ; CERQUEIRA, J. C. ; ABURJAILE, F. F. ; ZAMYATNIN JR, A. A. ; AZEVEDO, V. ; CARVALHO, R. D. O. . **Novel insights in Bacterial Vaginosis etiology through genomic approaches.** Anais da Academia Brasileira de Ciências v. 93, p. e20200945, 2021. https://doi.org/10.1590/0001-3765202120200945.

Contribuição: Escrita e revisão do manuscrito.

Abstract: Bacterial vaginosis (BV) has been considered as dysbiosis state whose etiology is not fully understood. This condition affects a large number of women of reproductive age and its study has been highly relevant due to the growing association of BV with and gynecological and obstetric complications and diseases, in addition to a greater susceptibility to sexually transmitted diseases, including HIV. The vaginal microbiota composition presents high variability among different ethnic groups of women, although, generally, the prevalence of lactobacilli species has been reported. Several studies suggest they may play a protective role, especially *Lactobacillus crispatus* whose population is typically present in low proportions in women with BV. This review article describes the contributions and limitations of genomic approaches in elucidating protective characteristics and mechanisms associated with colonization and persistence of lactobacilli strains. Although some genetic features were associated with resilience of *L. crispatus* during BV, furher studies are required to uncover their functions.

Artigo de revisão 3

Referência bibliográfica: ABURJAILE, F. F. ; **CANÁRIO, M.** ; CERQUEIRA, J. C. ; JESUS, L. C. L. ; SILVA, T. F. ; CARVALHO, R. D. O. ; AZEVEDO, V. A. C. **Probiotic potential of novel Brazilian *Lactobacillus crispatus* strains.** Genetics and Molecular Research, v. 20, p. GMR18900, 2021. https://doi.org/10.4238/gmr18900.

Contribuição: Escrita e revisão do manuscrito.

Abstract: *Lactobacilli* are the predominant bacterial species colonizing the vaginal surfaces of healthy women, where they play a protective role against opportunistic and polymicrobial infections, such as bacterial vaginosis. Several Lactobacillus species, especially *L. crispatus*, have been prospected for probiotic applications due to their potential antimicrobial and anti-inflammatory capacities. During the last decade, several genomic studies have been investigating the genetics of *L. crispatus* strains in an effort to identify novel probiotic strains and evaluate their potential for improving human and animal health. This mini review highlights the main genes associated with *L. crispatus* protective mechanisms in four novel strains of this species that we recently isolated from healthy Brazilian women of reproductive age. Among the probiotic features of these strains, the roles of a pyruvate oxidase-encoding gene, lactate synthesis related enzymes, bacteriocin genes, and genomic islands, are reviewed, and the next steps for confirming their activity are indicated.

APÊNDICE D – Trabalhos apresentados em eventos

1. CASTILLO, R. E. H. ; **VIANA, M. V. C.** ; CARHUARICRA, D. ; SOARES, S. C. ; MATURRANO, L. ; AZEVEDO, V. A. C. ; ABURJAILE, F. F. . Evolutionary analysis based on the *Pasteurella multocida* genome from Veterinary isolates. **X-Meeting eXperience 2020**, 2020. Disponível em: https://drive.google.com/file/d/18aoRaBQy-T3JSiPNJOm4_6TduJFn32vl/view
2. CARVALHO, R. D. O. ; ABURJAILE, F. F. ; SANTOS, R. P. S. ; **VIANA, M. V. C.** ; AZEVEDO, V. A. C. . Unraveling potential probiotic features in the genome of *Lactobacillus paragasseri*. **X-Meeting eXperience 2020**, 2020. Disponível em: https://drive.google.com/file/d/18aoRaBQy-T3JSiPNJOm4_6TduJFn32vl/view
3. **VIANA, M. V. C.**; KUMAVATH, RANJITH ; IMCHEN, MADANGCHANOK ; AZEVEDO, V. ; WATTAM, ALICE REBECCA . Microbial communities from mangroves soil and the influence of conversion to shrimp ponds on carbon stock and greenhouse gas emissions. **Aquaculture 2019**. 2019. Disponível em: https://www.was.org/Meeting/pdf/AQ2019BluePages.pdf
4. CERQUEIRA, J. C. ; SANTOS, R. P. S. ; SILVA, A. L. ; CASTILLO, R. E. H. ; ALMEIDA, M. O. ; CASTRO, T. ; RODRIGUES, D. L. N. ; BAEZ, J. L. V. ; COSTA, F. R. ; GOMIDE, A. C. P. ; FIGUEIREDO, H. C. P. ; WATTAM, A. R. ; SILVA, A. ; AZEVEDO, V. A. C. ; **VIANA, M. V. C.** . Taxonomy and comparative genomics of a *Corynebacterium ulcerans* strain isolated from Pig, previously identified as *C. pseudotuberculosis*. **X-Meeting 2019**, 2019. p. 96-96. Disponível em: https://drive.google.com/file/d/1cj0e5a1lpceSrNuthuVDOTEJnhl2hSZ-/view.
5. CASTILLO, R. E. H. ; CERQUEIRA, J. C. ; SANTOS, R. P. S. ; **VIANA, M. V. C.** ; AZEVEDO, V. A. C. . Genomic and epidemiologycal analyses of *Manhheimia haemolytica* strains. **X-Meeting 2019**, 2019. p. 118-118. Disponível em: https://drive.google.com/file/d/1cj0e5a1lpceSrNuthuVDOTEJnhl2hSZ-/view.
6. SANTOS, R. P. S. ; JESUS, L. C. L. **; VIANA, M. V. C.** ; CERQUEIRA, J. C. ; DRUMOND, M. M. ; MANCHA-AGRESTI, P. ; BRENIG, B. ; AZEVEDO, V. A. C. . Genomic characterization of *Lactobacillus delbrueckii* CIDCA 133: a potential probiotic strain. **X-Meeting 2019**, 2019. p. 128-128. Disponível em: https://drive.google.com/file/d/1cj0e5a1lpceSrNuthuVDOTEJnhl2hSZ-/view.

7. **VIANA, M. V. C.**; JAISWAL, A. K. ; SILVA, A. L. ; RIBEIRO, B. S. ; SANTOS, R. P. S. ; JAMAL, S. B. ; GUIMARÃES, LUIS C. ; SOARES, SIOMAR C. ; TIWARI, S. ; WATTAM, A. R. ; AZEVEDO, V. ; SILVA, A. . In silico identification of drug and vaccine targets in *Corynebacterium ulcerans*. **X-Meeting 2018**, 2018.
8. RIBEIRO, B. S. ; CAETANO, ANA C. B. ; SILVA, A. L. ; DOS SANTOS, ROSELANE G. ; **VIANA, M. V. C.** ; SANTOS, R. P. S. ; SEYFFERT, N. ; LE LOIR, YVES ; BRENIG, B. ; AZEVEDO, V. ; CASTRO, T. . Complete genome sequence of *Staphylococcus aureus* strain O17. **X-Meeting 2018**, 2018.
9. SANTOS, R. P. S. ; SEYFFERT, N. ; TIWARI, S. ; **VIANA, M. V. C.** ; JAISWAL, A. K. ; CAETANO, ANA C. B. ; RAMOS, P. I. ; SOARES, SIOMAR C. ; GALA-GARCIA, A. ; AZEVEDO, VASCO ; CASTRO, T. . Comparative genomic analysis with a multidrug resistant *Klebsiella pneumoniae* strain recently isolated from a patient in Brazil. **Gene Time Conference**. 2018.
10. CAETANO, ANA C. B. ; DOS SANTOS, ROSELANE G. ; PARISE, DOGLAS ; **VIANA, M. V. C.** ; SANTOS, R. P. S. ; TIWARI, S. ; GALA-GARCIA, A. ; SEYFFERT, N. ; AZEVEDO, VASCO ; CASTRO, T. . Identification of core, accessory and exclusive genes of *Staphylococcus aureus* isolated from ovine mastitis. **Gene Time Conference**. 2018.

APÊNDICE E – Participação em eventos

1. X-Meeting eXperience 2020.
2. 111th Annual Meeting - National Shellfisheries Association. 2019.
3. Aquaculture 2019.
4. LIA BACT-INFLAM 2019.
5. X-Meeting 2019.
6. 3rd Brazilian Student Council Symposium. 2018.
7. X-Meeting 2018.

APÊNDICE F – Organização de eventos

1. 2nd Associated International Laboratory Meeting (LIA 2019) - Bact-Inflam Conference. UFMG, Belo Horizonte, MG, Brasil. 2019.

APÊNDICE H – Disciplinas ministradas

1. Pós-Graduação (2019.1): Aplicação da Bioinformática para a Ecologia e Genética (BIG890V). 16h.
2. Graduação (2018.2, Medicina Veterinária): Genética e Evolução (BIG601). 30h.
3. Graduação (2019.2, Biomedicina): Genética (BIG157). 15h.
4. Graduação (2019.2, Farmácia): Genética (BIG157). 15h.
5. Graduação (2020.2, Biomedicina): Genética (BIG157). 15h.
6. Graduação (2020.2, Farmácia): Genética (BIG157). 15h.
7. Graduação (2021.1, Biomedicina): Genética (BIG157). 15h.
8. Graduação (2021.1, Farmácia): Genética (BIG157). 15h.

APÊNDICE I – Cursos e estágios ministrados

1. Curso: VIANA, M. V. C.; AZEVEDO, VASCO. Genômica comparativa de procariotos. 2018.
2. Estágio: Estágio em genômica de procariotos. Discente: Paulo Barros de Abreu Junior. Universidade Federal de Minas Gerais. 120h.

APÊNDICE J – Coorientações

1. Alessandra Lima da Silva. Seleção diversificadora em *Corynebacterium pseudotuberculosis*. 2018. Dissertação (Mestrado em Bioinformática) - Universidade Federal de Minas Gerais.
2. Marcelle Oliveira de Almeida. Comparative genomics of *Lactobacillus crispatus*: analysis of strain-specific antimicrobial genetic factors. 2021. Tese (Doutorado em Genética) - Universidade Federal de Minas Gerais.

APÊNDICE K – Bancas

1. Dissertação de mestrado: AZEVEDO, V. A. C.; PINTO, ANNE CYBELE; NICOLI, J. R.; LOBO, F. P.; **VIANA, M. V. C.**. Participação em banca de Jéssica Gabrielle Vidal da Silva. Genômica Comparativa e Avaliação In Silico De Genes Envolvidos No Potencial Probiótico De *Bifidobacterium Longum* 51A. 2018. Dissertação (Mestrado em Bioinformática) - Universidade Federal de Minas Gerais.

2. Qualificação de doutorado LAGE, A. P.; **VIANA, M. V. C.**; FOSTER, J.. Participação em banca de Carine Rodrigues Pereira. Genetic epidemiology of *Brucella abortus* isolated from cattle in Brazil. 2021. Exame de qualificação (Doutorando em Ciências Veterinárias) - Universidade Federal de Lavras.

APÊNDICE L – Prêmios

1. Johnnie Castro Montealegre Student Awards, FUCOBI. 2019.
2. 3º lugar em concurso para Professor de Educação Superior, Universidade Estadual de Minas Gerais. 2019. https://www.jornalminasgerais.mg.gov.br/?dataJornal=2019-11-02.
3. 1° colocado no processo seletivo para entrada no doutorado em Bioinformática (2018.3), Universidade Federal de Minas Gerais. 2018.
4. Best Poster Award of X-Meeting 2018, AB3C. 2018.

APÊNDICE M – Pós-Doutorados

1. Área: Genômica. Instituição: Universidade Federal do Pará (UFPA). Agência Financiadora: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Supervisor: Vasco Ariston de Carvalho Azevedo. Duração: 6 meses. 2018.
2. Área: Genômica. Instituição: Universidade Federal do Pará (UFPA). Agência Financiadora: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Supervisor: Artur Luiz da Costa da Silva. Duração: 12 meses. 2018-2019.
3. Área: Genômica. Instituição: Universidade Federal do Pará (UFPA). Agência Financiadora: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Supervisor: Artur Luiz da Costa da Silva. Duração: 14 meses. 2019-2020.
4. Área: Genômica. Instituição: Institut National de la Recherche Agronomique (INRAE). Agência Financiadora: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Supervisores: Jean-Marc Chatel, Vasco Ariston de Carvalho Azevedo. Duração: 12 meses. 2021-2022.
5. Área: Genômica. Universidade Federal de Minas Gerais (UFMG). Agência Financiadora: Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). Supervisor: Vasco Ariston de Carvalho Azevedo. Duração: 12 meses. 2022-2023.