UNIVERSIDADE FEDERAL DA BAHIA

INSTITUTO DE CIÊNCIAS DA SAÚDE

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

DANIELE ALMEIDA ALVES

AVALIAÇÃO DA UTILIZAÇÃO DE ÍNDICES DE SIMILARIDADE GENÔMICA GLOBAL PARA CLASSIFICAÇÃO DE ESPÉCIES PATOGÊNICAS EMERGENTES DO GÊNERO *Corynebacterium*

Salvador

2020

DANIELE ALMEIDA ALVES

AVALIAÇÃO DA UTILIZAÇÃO DE ÍNDICES DE SIMILARIDADE GENÔMICA GLOBAL PARA CLASSIFICAÇÃO DE ESPÉCIES PATOGÊNICAS EMERGENTES DO GÊNERO *Corynebacterium*

Dissertação de mestrado submetido ao Programa de Pós-Graduação em Biotecnologia do Instituto de Ciências da Saúde da Universidade Federal da Bahia como requisito parcial para obtenção do grau de Mestre em Biotecnologia

Orientador: Dr. Luis Gustavo Carvalho Pacheco (UFBA)

Coorientador: Dr. Eric Roberto Guimarães Rocha Aguiar (UESC)

Salvador

2020

Ficha catalográfica elaborada pelo Sistema Universitário de Bibliotecas
(SIBI/UFBA), com os dados fornecidos pelo(a) autor(a)

Alves, Daniele Almeida
Avaliação da utilização de índices de similaridade genômica global para classificação de espécies patogênicas emergentes do gênero C*orynebacterium*. / Daniele Almeida Alves. -- Salvador, 2020.
66 f. : il

Orientador: Luis Gustavo Carvalho Pacheco.
Dissertação (Mestrado – PPGBiotecnologia) -- Universidade Federal da Bahia, Instituto de Ciências da Saúde, 2020.

1. Bibliotecas. 2. Bibliotecas Universitárias. I. Nome, Orientador. II. Título.

DANIELE ALMEIDA ALVES

AVALIAÇÃO DA UTILIZAÇÃO DE ÍNDICES DE SIMILARIDADE GENÔMICA GLOBAL PARA CLASSIFICAÇÃO DE ESPÉCIES PATOGÊNICAS EMERGENTES DO GÊNERO *Corynebacterium*

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Biotecnologia, Instituto de Ciências da Saúde, da Universidade Federal da Bahia

n° 98

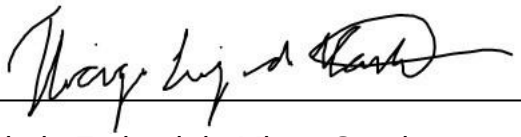
Aprovada em 08 de outubro de 2020

Banca Examinadora

Eric Roberto Guimarães Rocha Aguiar – Coorientador 

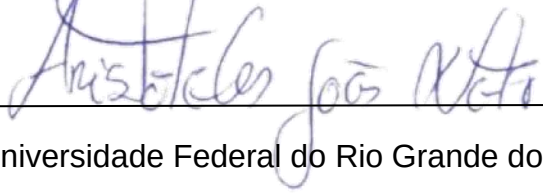
Doutor em Bioinformática pela Universidade Federal de Minas Gerais

Universidade Estadual de Santa Cruz

Thiago Luiz de Paula Castro 

Doutor em Genética pela Universidade Federal de Minas Gerais

Instituto de Ciências da Saúde/ Universidade Federal da Bahia

Aristóteles Góes Neto 

Doutor em Botânica pela Universidade Federal do Rio Grande do Sul

Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais

AGRADECIMENTOS

Primeiramente quero agradecer a Deus por me proporcionar perseverança durante toda a minha vida;

À Universidade Federal da Bahia pelas oportunidades conferidas a mim durante os anos de formação;

Ao meu orientador, professor Dr. Luis Pacheco, por me aceitar como orientanda, pela receptividade, confiança e valiosas contribuições nessa etapa tão importante na minha carreira profissional. Obrigada por me manter motivada durante todo o processo;

Ao meu coorientador professor Dr. Eric Aguiar, pela confiança, incentivo e por dedicar inúmeras horas para sanar as minhas dúvidas e me colocar na direção correta, me ajudando sempre que necessário;

Aos Grupo de Estudos em Genômica Funcional e Biologia Sintética (GenoFun) e ao Grupo de Pesquisa em Bioinformática de Vírus, coordenados pelos meus orientadores e aos meus colegas do grupo;

Ao professor Dr. Thiago Luiz de Paula Castro, por aceitar o convite para compor a banca de avaliação final;

Ao professor Dr. Aristóteles Góes Neto por também ter aceitado o convite para compor a banca de avaliação final;

A todos os mestres que contribuíram com a minha formação acadêmica e profissional até aqui;

À Maria Brandão, por me ajudar a compreender meu verdadeiro potencial;

Aos meus pais e meu irmão por me ajudar a seguir em busca da realização dos meus sonhos;

E a todos meus amigos que me apoiaram em cada momento dessa trajetória, trazendo alegria nos momentos difíceis.

ALVES, Daniele Almeida. Avaliação da utilização de Índices de Similaridade Genômica Global para classificação de espécies patogênicas emergentes do gênero *Corynebacterium*. 66 f. 2020. Dissertação (Mestrado) – Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, 2020.

RESUMO

Índices de Similaridade Genômica Global têm sido amplamente utilizados nos últimos anos para classificação taxonômica de bactérias. Entre estes, a Identidade Média de Nucleotídeos por BLAST (ANIb) é amplamente considerada como o índice mais preciso para a circunscrição de espécies bacterianas, quando se considera um limite 95-96% de identidade. No entanto, o uso exclusivo de ANIb para esta identificação pode gerar resultados confusos para alguns grupos bacterianos intimamente relacionados, como muitos patógenos de *Corynebacterium* spp. Nesse trabalho desenvolvemos e avaliamos o desempenho de um classificador de espécies desenvolvido internamente, com base nos valores de correlação entre diferentes índices de parentesco do genoma, para classificar sequências genômicas do grupo *Corynebacterium diphtheriae*. Para isso, 213 sequências genômicas correspondentes a três espécies de *Corynebacterium* intimamente relacionadas foram recuperadas do NCBI Genoma DB: 188 classificados como *C. diphtheriae* e 03 isolados de *C. diphtheriae* subsp. *lausannense*; 10 como *Corynebacterium belfantii*; 01 como *Corynebacterium rouxii* e 11 classificados como *C. diphtheriae* obtidos do Arquivo Europeu de Nucleotídeos. Os padrões de uso de tetranucleotídeos (TETRA) e ANIb foram calculados no servidor Web JSpecies através de comparações par-a-par entre os genomas avaliados. As matrizes resultantes foram mescladas para gerar uma matriz com os valores concatenados de ANIb e TETRA para cada bactéria, representando uma forma de impressão digital, que foi então utilizada para calcular os valores de correlação de Spearman entre os genomas através de uma estratégia própria desenvolvido no ambiente estatístico R. Análise de sequência multilocus utilizando os genes: *atpA*, *gyrA*, *dnaE*, *dnaK*, *fusA*, *leuA* e *rpoB* e análises de decomposição dividida foram usados para confirmar as relações entre as várias espécies. No total, 45.369 comparações genoma a genoma compuseram a matriz de impressão digital das bactérias que foi usada para construir um dendrograma com clados bem definidos (> 95% de confiança de *bootstrap*). Os grupos contendo *C. belfantii* e *C. rouxii* foi claramente distinguido por esta estratégia, em oposição ao uso de ANIb sozinho que não foi capaz de diferenciar as espécies *C. diphtheriae* subsp. *lausannense* CHUV2995 e *C. belfantii*. Adicionalmente, observamos que nossos resultados são corroborados pela MLSA, evidenciando a classificação errada no NCBI. Com os resultados supracitados nós concluímos que o classificador desenvolvido internamente que integra diferentes índices foi a ferramenta mais eficiente para a circunscrição de espécies no grupo *C. diphtheriae*, quando comparada para ANIb sozinho. Antecipamos que esta nova estratégia pode ser extrapolada para melhorar a identificação baseada no genoma de outros patógenos bacterianos clinicamente importantes.

Palavras-chave: patógenos emergentes, taxonomia genômica, Identidade média de nucleotídeos, padrões tetranucleotídicos, *Corynebacterium* spp

ALVES, Daniele Almeida. Avaliação da utilização de Índices de Similaridade Genômica Global para classificação de espécies patogênicas emergentes do gênero Corynebacterium. 66 f. 2020. Dissertação (Mestrado) – Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, 2020.

ABSTRACT

Overall genome-relatedness indexes (OGRIs) have been extensively used in recent years for taxonomic classification of bacteria. Among these, Average Nucleotide Identity by BLAST (ANIb) is widely regarded as the most accurate index for bacterial species circumscription, when considering a species boundary of ca. 95-96% identity. However, the sole use of ANIb for species identification may render confusing results for some closely related bacterial groups, such as many pathogenic *Corynebacterium* spp. In this work we develop and evaluate performance of an *in-house* developed species-classifier, based on the correlation values between different genome relatedness indexes, to correctly classify genomic sequences from the *Corynebacterium diphtheriae* group. For that, 213 genomic sequences corresponding to three closely related *Corynebacterium* species were retrieved from NCBI's Genome DB: 188 classified in NCBI Taxonomy as *Corynebacterium diphtheriae*, including the reference strain NCTC11397 and 03 isolates of *C. diphtheriae* subsp. Lausannense; 10 as *Corynebacterium belfantii*; 01 as *Corynebacterium rouxii* and 11 classified as *C. diphtheriae* obtained from the European Nucleotide Archive. Tetranucleotide usage patterns (TETRA) and average nucleotide identities by BLAST (ANIb) were calculated through the JSpecies Web server application and compared all-vs-all. Resulting matrices were then merged to generate a single fingerprint matrix, which was used to calculate the Spearman's correlation values among bacterial genomes using an *in-house* script developed on R software. MLSA (genes: *atpA*, *gyrA*, *dnaE*, *dnaK*, *fusA*, *leuA* e *rpoB*) and split-decomposition analyses were used to confirm relationships between the various species. In total, 45,369 genome-to-genome comparisons composed the fingerprint matrix that was used to build a dendogram with well-defined clades (> 95% boostrap confidence). The groups containing *C. belfantii* and *C. rouxii* were clearly distinguished by this strategy, as opposed to the use of ANIb alone that was unable to differentiate the *C. diphtheriae* subsp. *lausannense* strain CHUV2995 and *C. belfantii*. Additionally, we observed that our results are corroborated by the MLSA, highlighting the wrong classification in the NCBI. With the aforementioned results we concluded that the classifier developed internally that integrates different OGRIs was the most efficient tool for the circumscription of species in the group *C. diphtheriae*, when compared to ANIb alone. We anticipate that this new strategy can be extrapolated to improve the genome based on identification of other clinically important bacterial pathogens.

Keywords: Emerging pathogens, Genomic taxonomy, Average Nucleotide Identity, Tetranucleotide patterns, *Corynebacterium* spp

LISTA DE IMAGENS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

A classificação dos organismos é um tema de corriqueiro debate entre os cientistas causando controvérsias sobre o assunto. Taxonomistas buscam estabelecer premissas para auxiliar a circunscrição das unidades biológicas, incluindo os procariontes (RICHTER, ROSSELÓ-MORA, 2009). Na microbiologia, a maioria dos cientistas concordam que o objetivo da taxonomia é classificar o microrganismo representando uma hierarquia estabelecendo um sistema que represente as relações taxonômicas como uma ordem na natureza (KÄMPFER, GLAESER, 2012).

Inicialmente a classificação de procariontes era baseada apenas em testes fenotípicos tendo como consequências classificações confusas. A integração de dados genômicos nas descrições das espécies foi um importante avanço que contribuiu para a compreensão atual das classificações taxonômicas. Este avanço se deu através do desenvolvimento da hibridização DNA-DNA (DDH) em 1960, técnica que permitiu comparações entre genomas (ROSELLÓ-MÓRA, AMANN, 2015; RICTHER, ROSELLÓ-MÓRA, 2009). Contudo a inexistência de limites claros para comparação entre espécies foi sempre um entrave nesse tipo de análise. Estudos comparativos indicaram que o valor de 70% de similaridade proposto como ponto de corte não poderia ser um valor absoluto, sugerindo um limiar entre 60 a 70%, reforçando que apenas a sequência completa de DNA deve ser o padrão de referência para determinar a classificação das espécies. Esse método tem sido criticado devido a sua complexidade e demora na realização. O segundo avanço que revolucionou os estudos taxonômicos procarióticos foi a reconstrução genealógica baseada nos genes de RNA ribossômicos, em particular o *16S* rRNA (ROSELLÓ-MÓRA, AMANN 2015; KÄMPFER, GLAESER, 2011).

A substituição da DDH como padrão ouro na classificação procariótica tem sido proposta, e muitos esforços tem sido realizados para desenvolver um método substitutivo com valores análogos à DDH (CHUN et al., 2018; KIM et al., 2014). A utilização do *16S* é considerado um marco nos estudos taxonômicos procarióticos, diante da possibilidade em estabelecer um sistema taxonômico

hierárquico com base em um marcador molecular prático (KÄMPFER, GLAESER, 2011). No entanto, resoluções filogenéticas discriminadas apenas com o *16S* rRNA tem sido amplamente discutida, pois algumas espécies compartilham similaridade >99% ainda que estejam separadas por DDH (KIM et al., 2014). Como alternativa à abordagem do *16S,* a Análise de Sequência Multilocus (MLSA) pode ser utilizada para visualizar uma reconstrução filogenética mais robusta, já que para essa abordagem são sugeridos genes ortólogos codificadores de proteínas (ROSELLÓ-MÓRA, AMANN, 2015).

Em razão dos genes codificadores de proteínas únicas não refletirem relações filogenéticas gerais, a MLSA tem como proposta superar o viés causado por filogenias baseadas em sequências de genes únicos. Nesse método fragmentos internos de vários genes codificadores de proteínas são alinhados e concatenados para calcular árvores filogenéticas. Os requisitos básicos para a MLSA são a seleção de genes, verificação da qualidade da sequência, comprimento dos fragmentos dos genes e cálculo das árvores filogenéticas. Embora a MLSA esteja sendo cada vez mais aplicada para obter um maior poder de resolução entre as espécies dentro de um gênero, há um ponto crítico para as análises que é a seleção dos genes, pois não há um consenso sobre a quantidade de genes (GLAESER, KÄMPFER, 2015).

Nas situações em que as espécies apresentam alta semelhança na sequência do *16S* com as espécies filogenéticas vizinhas, genes como *rpoB*, *gyrB*, e *recA* são sugeridos para apoiar ainda mais a autenticidade dos dados do genoma. Entretanto não há um limiar universal nem um número comum e um conjunto de genes propostos para MLSA (ROSELLÓ-MÓRA, AMANN 2015). Diferentemente da MLSA, as análises comparativas do genoma são mais econômicas e menos subjetivas (GLAESER, KÄMPFER, 2015).

Nesse contexto, a genômica ganhou destaque pelas suas características de especificidade, reprodutibilidade e confiabilidade para inferir relações filogenéticas entre procariontes. Os avanços nas tecnologias de sequenciamento do DNA e a ampla disponibilidade do genoma têm facilitado análises da amplitude do genoma com análises comparativas que podem fornecer resultados mais precisos e análogos à DDH. Nas comparações *in-silico* Índices De Similaridade Genômica Global (do inglês - Overall genome-relatedness indexes -

OGRIs) foram desenvolvidos tendo como principal aplicação na taxonomia bacteriana o cálculo da relação genômica geral entre duas espécies, servindo como estrutura para o conceito de espécie (CIUFO et al., 2018; LEE et al., 2016). Estes índices permitem o cálculo da similaridade entre duas sequências genômicas sem ter que selecionar genes, realizar passos de anotação, tendo como vantagem objetividade, reprodutibilidade, rapidez e ser fácil de implementar (KIM et al., 2014).

Dentre os vários índices, a Identidade Média de Nucleotídeos por BLAST (do inglês *Average Nucleotide Identity by BLAST*– ANIb) tem sido a métrica mais utilizada e sugerida como a melhor escolha para determinar relacionamento entre as espécies, bem como confirmar a identificação (CIUFO et al., 2018). O ANIb é definido como uma medida de pares de similaridade entre duas sequências do genoma em que a sequência do genoma de consulta é fragmentada *in silico* em sequências de 1020 bp de comprimento, e esses fragmentos são então comparados contra o outro genoma para encontrar regiões homólogas (YOON et al., 2017). Além do ANIb, a frequência de fragmentos de quatro nucleotídeos no genoma (TETRA) é sugerida como um parâmetro complementar na identificação de genomas em nível de espécie (LEE et al., 2016; ROSSELLÓ-MÓRA, AMANN, 2015).

Os cálculos estatísticos de frequências oligonucleotídicas entre as sequências são reconhecidos como uma alternativa em que apresenta rapidez, facilidade em sua implementação, além de não ser necessário realizar alinhamento das sequências. Essas frequências carregam um sinal específico da espécie, porém as razões evolutivas ainda não foram esclarecidas (PRIDE et al., 2003).

A identificação correta da espécie bacteriana causadora da infecção, com determinação do perfil de sensibilidade a antimicrobianos é um dos pilares de um programa de gerenciamento de sucesso previsto pela Agência Nacional de Vigilância Sanitária (GVIMS/GGTES/ANVISA, 2017). No entanto, a identificação correta destes patógenos com base nos perfis bioquímicos é um desafio, pois são necessários diversos testes bioquímicos que não estão disponíveis no *API Coryne*, um sistema padronizado para a identificação de bactérias corineformes que utiliza 21 testes bioquímicos miniaturizados para detecção de atividades

enzimáticas e fermentação de carboidratos, além de uma base de dados on-line específica, gerando resultados após 24 horas (BERNARD, 2012; CDC, 2015).

Bactérias corineformes são cada vez mais reconhecidas como patógenos oportunistas sendo relatadas conjuntamente como um dos principais grupos bacterianos causadores de infecções multirresistentes em pacientes cronicamente doentes (FORBES, 2017).

Nesse contexto, o presente trabalho utilizou OGRIs para o delineamento de espécies e identificação específica de novos isolados através de cálculos de similaridade entre os genomas das linhagens de interesse e das linhagens de referência tendo *Corynebacterium diphtheriae* como grupo controle. Validamos a estratégia de correlação baseada em uma assinatura específica para cada espécie do estudo utilizando valores de ANI e TETRA para classificação rápida e confiável de espécies bacterianas do gênero *Corynebacterium* patogênicas emergentes causadoras de infecções multirresistentes a antibióticos.

1.1 Fundamentação teórica

1.1.1 Infecções emergentes por Corinebactérias em humanos

Doenças infecciosas são significativamente uma ameaça à saúde pública há centenas de anos e atormentam a humanidade desde o seu início. Estas doenças são comumente causadas por microrganismos e as bactérias estão entre os principais agentes infecciosos (MCFEE, 2018). Surtos de doenças infecciosas emergentes e reemergentes continuam sendo um problema em todo o mundo. A disseminação global destes agentes, inclusive os resistentes a antibióticos, representam uma ameaça à segurança da saúde pública (ZUMLA, HUI, 2019; NII-TREBI, 2017). Há pouco mais de uma década as doenças infecciosas já eram responsáveis por 26% das mortes anuais em uma população global de 6,2 bilhões (MORENS, FOLKERS, FAUCI, 2008). Na nossa história mais recente, eventos marcantes no campo das doenças infecciosas vêm acontecendo como pandemia da Síndrome Respiratória Aguda Grave (SARS) (2002-2004), o surto de doença do vírus Ebola na África Ocidental (2013-2016), o vírus Zika nas Américas e sudeste da Ásia (2016–2018), difteria na Venezuela (2016-2017) e no Iêmen (2017-2018), entre outras (ZUMLA, HUI, 2019).

Segundo a Organização Mundial de Saúde (OMS) uma doença que aparece pela primeira vez em uma população, ou até mesmo já tenha existido e está aumentando em passo acelerado ocasionando novos casos ou em nova extensão geográfica é considerada como emergente (NII-TREBI, 2017). A difteria é uma infecção bacteriana que está na lista doenças infecciosas prioritárias que ameaçam a segurança global da saúde (ZUMLA, HUI, 2019). Esta doença é causada por bactérias toxigênicas do gênero *Corynebacterium*, principalmente *C. diphtheriae*, e raramente *Corynebacterium ulcerans* e *Corynebacterium pseudotuberculosis* intimamente relacionadas (SHARMA et al., 2019).

As espécies não diftéricas deste gênero têm emergido como patógenos oportunistas, estando amplamente distribuídos no meio ambiente. Além disso, são conhecidos por seu comportamento comensal na pele e mucosas se tornando um foco de atenção, pois há relatos crescentes de seu isolamento em pacientes com

várias infecções, sendo capazes de causar doenças graves que não podem ser prevenidas com vacinas (SHARMA et al., 2019). Dentre as infecções causadas por estes microrganismos estão endocardite, pneumonia, osteomielite, sepse, artrite séptica e diversas infecções em sítios de implantação de aparatos médicos como próteses e cateteres venosos (YANAI et al. 2018; BERNARD, 2012).

Embora *C. diphtheriae*, agente causador da difteria, continue sendo o patógeno mais significativo do gênero e o homem o único hospedeiro natural conhecido, *C. pseudotuberculosis* e *C. ulcerans* são transmitidos para humanos pelo contato com animais doentes e são capazes de produzir a toxina diftérica causando doenças com sintomas semelhantes aos da difteria (BERNARD, 2012). Recentemente infecções por *C. ulcerans* tem sido mais comum que *C. diphtheriae* no Reino Unido, e na Europa este patógeno vem sendo cada vez mais relatado em casos com sintomas clínicos típicos de difteria, inclusive em indivíduos imunizados (SHARMA et al., 2019).

As espécies não diftéricas, comumente encontradas como constituintes da microbiota residente da pele e mucosas em seres humanos têm sido relatadas conjuntamente como um dos principais grupos bacterianos causadores de infecções multirresistentes em pacientes cronicamente doentes em diversos países desenvolvidos e em desenvolvimento, incluindo o Brasil (GILBERT et al., 2018; BERNARD, 2012). Em um estudo realizado recentemente (2014-2016) em um hospital universitário no Japão, prontuários de todos os pacientes foram avaliados e mais de 40% dos pacientes foram diagnosticados com bacteremia, sendo *C. striatum* causador de pneumonia e infecções no trato urinário, a espécie mais frequentemente identificada em isolados clínicos. Apesar de os comensais cutâneos bacterianos apresentarem virulência relativamente baixa, há relatos recentes que a produção de biofilme por *C. striatum* resistente a antimicrobianos é um novo fator de virulência e está relacionado a surtos nosocomiais (YANAI et al., 2018).

1.1.2 Multirresistência a antibióticos

Em 1928 a penicilina, primeiro antibiótico comercializado, foi descoberta por Alexander Fleming, e juntamente com esta descoberta veio o reconhecimento da

resistência aos medicamentos pelos microrganismos, assim como o compartilhamento da resistência uns com os outros (CDC, 2020). A resistência aos medicamentos pode ser causada por eventos como mutação, obtenção de genes por meio de transformação ou infecção com plasmídeos (MORENS, FOLKERS, FAUCI, 2008). A resistência a medicamentos também pode ser resultado de mutações no genoma de um patógeno devido a danos genéticos que podem ser causados por exposição a produtos químicos e agentes antimicrobianos levando ao surgimento de variantes do patógeno podendo causar novas doenças, ocasionando eventos como epidemias (NII-TREBI, 2017).

As infecções causadas por microrganismos resistentes a antibióticos proporcionam dificuldade e as vezes até impossibilidade de tratamento, representando mundialmente um dos mais emergentes problemas para a saúde pública (CDC, 2020). O combate a essa ameaça é uma prioridade de saúde pública, e requer uma abordagem global colaborativa. Os países têm autonomia para tomar medidas para diminuir a resistência antimicrobiana implementando programas eficazes de gerenciamento, como preconizado pela Diretriz Nacional para Elaboração de Programa de Gerenciamento do Uso de Antimicrobianos em Serviços de Saúde, elaborada em 2017 pela Agência Nacional de Vigilância Sanitária (GVIMS/GGTES/ANVISA, 2017).

Devido ao aumento do uso de antibióticos de amplo espectro o surgimento de cepas multirresistentes vêm sendo relatados em diversos estudos. No estudo realizado recentemente por Yanai e colaboradores (2018) foram identificados 66 casos de infecções por *Corynebacterium* spp. Sendo a identificação final como *C. striatum*, *Corynebacterium jeikeium* e *Corynebacterium argentoratense* sendo a maioria dos isolados multirresistentes à penicilina, imipenem / cilastatina, eritromicina, clindamicina e levofloxacina. Mais recentemente foi relatado que *C. diphtheriae* apresentou resistência ou sensibilidade reduzida a penicilina, cefotaxima, tetraciclina e cloranfenicol, e que os fenótipos desta espécie têm probabilidade de diminuir o efeito da terapia antimicrobiana. Adicionalmente, foi sugerido que a vulnerabilidade reduzida à penicilina e eritromicina induz a formação de biofilme e hidrofobicidade da superfície celular (SHARMA et al., 2019).

1.1.3 O gênero *Corynebacterium*

O gênero *Corynebacterium* pertence à família Corynebacteriaceae, compreendendo bactérias Gram-positivas, com morfologia de bacilos, geralmente aeróbicas e alto conteúdo G+C incluindo atualmente aproximadamente 111 espécies. Além de *C. diphtheriae*, as espécies *C. ulcerans* e *C. pseudotuberculosis* também produzem a toxina diftérica, codificada por um gene (*tox*), que é transportado por um prófago e adquirido por transferência horizontal de genes causando doenças semelhantes a difteria em humanos (BADELL et al., 2020; SHARMA et al., 2019; BERNARD, 2012).

O patógeno *C. diphtheriae* é o principal agente causador da difteria, que é uma infecção potencialmente fatal em humanos, uma causa significativa de morbidade e mortalidade global. *C. diphtheriae* é uma espécie geneticamente heterogênea e com base nas características bioquímicas tem sido tradicionalmente subdividido em quatro *biovares*: *gravis*, *mitis*, *belfanti* e *intermedius*, sendo este último quase nunca relatado na literatura recente (BADELL et al., 2020; DAZAS et al., 2018; SANGAL et al., 2013). Exceto belfanti, a nomenclatura dos biovares se dá de acordo com o agravo da doença, sendo os seguintes níveis: leve para mitis, intermediário para intermedius e grave para gravis (SHARMA et al., 2019).

A separação dos biovares tem sido usada para diferenciar as cepas na microbiologia clínica, porém a separação entre os biovares é considerada complexa e a base genômica não é clara, pois as cepas de *C. diphtheriae* podem ser geneticamente mais distantes dentro de um biovar do que entre eles (TAGINI et al., 2018). Análises genômicas comparativas de 17 cepas indicaram que não há correlação entre a separação bioquímica com a variação do conteúdo gênico envolvido nas categorias metabólicas relevantes que estão potencialmente envolvidas na discriminação biovar. Além disso é proposto que a caracterização molecular de cepas seja adotada em detrimento da análise fenotípica (SANGAL et al., 2013).

Nosso grupo de pesquisa tem trabalhado com diversas abordagens de taxonomia genômica para auxiliar a identificação de espécies patogênicas emergentes de *Corynebacterium* spp. a fim de fornecer novas ferramentas para auxiliar programas efetivos de administração de antimicrobianos no gerenciamento

de infecções causadas por essas bactérias. Dentre os trabalhos já realizados, Santos et al., (2018) realizou uma revisão abrangente da literatura com uma abordagem de bioinformática com o propósito de identificar a base genômica que contribui para as variabilidades bioquímicas observadas nos métodos de identificação fenotípica dessas bactérias.

A escolha das reações bioquímicas alvo se deu através de resultados conflitantes apresentados em estudos de identificação fenotípica de seis espécies patogênicas emergentes e reemergentes incluindo os biovares de *C. diphtheriae*. O trabalho apresentou um achado importante que foi a identificação de uma enzima com atividade aleatória apenas em algumas cepas específicas de *C. diphtheriae*, sugerindo como informação auxiliar no entendimento das habilidades diferenciais de utilizar glicogênio e amido entre os biovares, da mesma maneira que mostrou que transferência horizontal de genes desempenha um papel na variabilidade bioquímica dos isolados (SANTOS et al., 2018).

Alguns isolados do biovar belfanti foram propostos quase simultaneamente como subespécie *lausannense* e como nova espécie *C. belfantii* por diferentes autores. Diante de uma apresentação clínica muito particular e achados de broncoscopia em um paciente, Tagini e colaboradores (2018) investigaram uma cepa de *C. diphtheriae* (CHUV2995) propondo-a juntamente com as cepas CMCNS703 e CCUG 5865 como subespécie *lausannense* de *C. diphtheriae*, sendo que a subespécie reagrupa apenas as cepas biovar belfanti. Em contrapartida, Dazas e colaboradores (2018) propuseram o nome *C. belfantii* para o grupo de cepas anteriormente consideradas como *C. diphtheriae* biovar belfanti, visto que este biovar representa um ramo claramente demarcado dos biovares mitis e gravis, assim como a incapacidade de redução de nitrato apresentada pelo biovar belfanti.

Badell e colaboradores (2020) compararam seis isolados atípicos, inicialmente identificados como *C. diphtheriae* biovar belfanti, com a cepa tipo de *Corynebacterium belfantii* (FRC0043) proposta por Dazas e colaboradores (2018), cepas de *C. diphtheriae*, e as cepas de referência de *C. ulcerans* e *C. pseudotuberculosis*. Uma árvore filogenética baseada na sequência genômica revelou três principais clados, sendo que os isolados atípicos formaram um clado separado por um nível de divergência de nucleotídeos que está bem abaixo do limiar 95 – 96%, ponto de corte recomendado para o delineamento de espécies. Os

isolados agrupados neste clado foram propostos como uma nova espécie *Corynebacterium rouxii,* inclusive a cepa FRC0190, previamente identificada como *C. diphtheriae*. As principais características para distinguir *C. rouxii* do complexo de *C. diphtheriae* foram os biomarcadores específicos na espectrometria de massas no MALDI-TOF, o conteúdo GC%, apresentação atípica no *API Coryne* para a reação negativa com maltose dos seis isolados em que os resultados não era nem tão amarelo quanto as cepas tipicamente positivas, nem tão roxo quanto as cepas negativas.

1.1.4 Testes de identificação

A abordagem polifásica é um consenso adotado para a circunscrição de espécies para fins taxonômicos. Atualmente essa abordagem consiste em um conjunto de parâmetros que objetiva a concordância ao integrar diferentes tipos de dados em uma classificação de contradições mínimas. Monofilia, coerência genômica e fenotípica são as três principais premissas para uma classificação precisa de uma espécie (ROSSELLÓ-MÓRA, AMANN, 2015).

O gênero *Corynebacterium* inclui várias espécies bacterianas patogênicas importantes para a saúde humana e animal. Essas bactérias são comumente relatadas como taxonomicamente confusas pela sua variedade de perfis bioquímicos, o que dificulta uma identificação correta com base nos testes bioquímicos, pois requerem diversos testes que não estão disponíveis no sistema *API Coryne*, sistema padronizado para a identificação de bactérias corineformes que utiliza 21 testes bioquímicos miniaturizados para detecção de atividades enzimáticas e fermentação de carboidratos (SANTOS et al., 2018; BERNARD, 2012).

Apresentamos a Tabela 1 com as informações disponíveis sobre a base geral para diferenciação bioquímica de *C. dihphtheriae*, com base nos trabalhos realizados por Sangal (2013) e Santos (2016) e seus colaboradores.

Tabela 1 Base geral para diferenciação bioquímica de *C. diphtheriae*

	Mitis	Gravis	Intermedius	Belfanti
Redução de nitrato	+	+	+	-
Lipofilismo	-	-	+	-
Glicogênio	-	+	+	-
Amido	+[1]	+	+	-
Hemólise	+[2]	+[2]	-	?

[1] Raramente podem usar amido

[2] Algumas cepas podem ser fracamente hemolíticas

? Informações ainda desconhecidas na literatura

Fonte: Elaborada pelos autores com base em Sangal et al., (2013) e Santos et al., (2016)

Conforme apresentado na Tabela 1, é possível observar que as características gerais não são bem definidas, de forma que não há uma exclusividade de um perfil bioquímico para um determinado *biovar*. Além disso, os bancos de dados associados são atualizados com pouca frequência e não fornecem discriminação suficiente entre as espécies recém-descritas. As análises manuais consomem tempo e exigem grande quantidade de material biológico, o que pode ser uma desvantagem. Métodos moleculares, como sequenciamento dos genes *rpoB* e *16S* rRNA, demonstraram ter valor complementar, mas não são práticos para uso rotineiro devido ao seu alto custo. E ainda há relatos que essas espécies intimamente relacionadas não podem ser adequadamente diferenciadas por técnicas moleculares (BERNARD, 2012).

Sabendo que algumas espécies de *Corynebacterium* spp. são difíceis de identificar e requerem testes moleculares adicionais, foi desenvolvido também pelo nosso grupo de pesquisa um método de PCR multiplex (mPCR) baseado em genes específicos para cada espécie para uma diferenciação eficiente de isolados clínicos de *C. striatum*, *C. amycolatum* e *C. xerosis*. Fazendo uma comparação com trabalhos realizados anteriormente usando PCR para a identificação dessas corinebactérias, foi mostrada a necessidade de análises complementares, bem como foi ilustrada a utilidade do recém-desenvolvido mPCR com *primers* específicos como uma abordagem complementar rápida para a resolução de identificações ambíguas ou incorretas de espécies destas espécies. Além disso foi sugerida a

disponibilidade de um ensaio de mPCR que também poderia incluir a espécie *C. minutissimum* (SANTOS et al., 2016).

Outro teste utilizado atualmente na identificação microbiana clínica é a espectrometria de massas do tipo MALDI-TOF. Mesmo essa abordagem sendo considerada rápida, barata e precisa, as avaliações recentemente publicadas mostraram identificações ambíguas e limitações na variedade de espécies nos bancos de dados, que também não são atualizados com frequência adequada (ALIBI et al., 2015; BERNARD, 2012).

Atualmente, a genômica representa um método alternativo interessante quando outros testes simples não estão disponíveis (TAGINI et al., 2018; ALIBI et al., 2015). Os métodos genotípicos estão desempenhando um papel fundamental na identificação e classificação filogenética em vários níveis taxonômicos. O sequenciamento do genoma total é considerado o método mais confiável para identificação da espécie estudada. A identificação de genes espécie específicos é uma das vantagens apresentadas por esse método. Porém o alto custo de aquisição dos equipamentos necessários e requerimento de qualificação dos responsáveis pela execução e interpretação dos testes restringe tal teste a laboratórios de referência (GLAESER, KÄMPFER, 2015; KÖSER et al., 2012).

O gene codificador do RNA ribossômico *16S* rRNA foi e ainda é aplicado como gene marcador universal para análise evolutiva de bactérias cultiváveis e não cultiváveis. No entanto, o *16S* rRNA é bem conhecido por sua limitada resolução filogenética que dificulta a utilidade em nível de espécie ou subespécie. Sequências genômicas são recomendadas nas análises que tem como objetivo propósitos taxonômicos, em vez da utilização de hibridização convencional DNA-DNA e da filogenia baseada no *16S* rRNA (CHUN et al., 2018).

Na maioria das vezes, uma nova cepa é primeiro alocada filogeneticamente com base na análise do gene *16S* rRNA a nível de gênero. Buscando uma melhor resolução entre as espécies dentro de um gênero a MLSA é cada vez mais aplicado para este fim. Fragmentos internos de genes codificadores de proteínas são sequenciados e posteriormente usados para construir árvores filogenéticas. Como cada filogenia da sequência gênica única reflete apenas a evolução desse único gene, o que novamente pode não refletir a relação filogenética “verdadeira”,

fragmentos de DNA de várias sequências gênicas são concatenados e árvores filogenéticas são calculadas com base nisso. Assume-se que as árvores baseadas nas sequências alinhadas concatenadas refletem melhor a relação "verdadeira" dos táxons bacterianos (GLAESER, KÄMPFER, 2015).

Vários estudos relatam análises filogenéticas do gênero *Corynebacterium*, mas apenas através de um número limitado de cepas. Entre esses, foi feito um estudo de 56 espécies (KHAMIS, RAOULT, LA SCOLA, 2004) que analisou a filogenia com dois genes separados (*16S* e *rpoB*) para tentar identificar o melhor método para definir espécies do gênero. Um estudo posterior dos mesmos autores mostrou que uma sequência parcial de *rpoB* forneceu o melhor método para identificação de corinebactérias. Eles sugeriram que a realização de uma análise filogenética em todo o gênero usando uma análise mais ampla, incluindo muitos outros genes, resolveria algumas das diferenças observadas entre os estudos anteriores e forneceria mais informações sobre a evolução do gênero (OLIVEIRA et al., 2017).

1.1.5 Análises genômicas comparativas

O rápido progresso do sequenciamento de nova geração (NGS) está atualmente facilitando a análise da amplitude do genoma, sendo sugerida como uma alternativa ao DDH. Além disso, em comparação com o MLSA, a análise comparativa com os dados de sequenciamento do genoma disponíveis em bancos de dados é mais econômica e específica do que selecionar genes. A alta velocidade e o baixo custo do sequenciamento do genoma abrem as portas para comparações *in-sílico* que lembram o DDH (CHUN et al., 2018).

Diferentes parâmetros para comparar genomas com a finalidade de circunscrever espécies têm sido desenvolvidos, sendo a Identidade Média de Nucleotídeos (ANI) o mais reconhecido na comunidade científica. O recente avanço nas tecnologias de sequenciamento de DNA e sua ampla utilização, culminou nos últimos anos com a geração de mais de 140 mil projetos de sequenciamento genômico completo de organismos bacterianos, sendo cerca de 17 mil desses destinados ao sequenciamento de linhagens de referência (microrganismo-tipo) para várias

espécies, de acordo com os bancos de dados *GenBank* e GOLD *Genomes* (CHUN et al., 2018; MUKHERJEE et al., 2019).

Na versão mais atual do *Genome Taxonomy Database* (Release 05-RS95, 04 de outubro de 2020) constam 194.600 genomas de Bacteria e Archaea, representando 30.238 espécies diferentes de Bacteria (PARKS et al., 2019). Essa enorme disponibilidade de dados genômicos completos é agora uma rica fonte de informação para estudos taxonômicos de espécies bacterianas. A análise genômica comparativa apresenta vantagens em relação a hibridização de DNA, pois é mais econômica e específica, o que pode tornar a DDH uma técnica ultrapassada. Diante da disponibilidade de sequenciamento e fácil acesso ao genoma pelos laboratórios gerais de microbiologia, os OGRIs são recomendados a fim de minimizar os problemas apresentados pela DDH (CIUFO et al., 2018; LEE et al., 2016).

Richter e Rosselló-Móra (2009) avaliaram as frequências de assinatura de tetranucleotídeos como um recurso genômico independente de alinhamento para circunscrição de espécies. Eles observaram que a utilização de códons de cada tipo de genoma determina uma ocorrência de frequência característica para cada uma das 256 combinações de grupos de sequências de tetranucleotídeos. Com isso, é esperado que os genomas intimamente relacionados apresentem uma distribuição parecida do uso dessas assinaturas. Tal expectativa foi confirmada ao observar que ao plotar cada frequência tetranucleotídica correspondente, os genomas intimamente relacionados podem mostrar valores de correlação muito altos, em que os valores plotados seguem uma linha clara e quando os genomas mostram um certo grau de divergência, os valores plotados mostram maior dispersão e a correlação tende a diminuir.

Na mini revisão realizada por Rosselló-Móra e Amann (2015) em que os autores argumentam que o principal ponto de discordância na literatura é a definição, ou seja, a maneira como as espécies são circunscritas por meio de caracteres observáveis, é recomendado que o DDH não seja mais utilizado, e a MLSA é considerada pouco competitiva, trazendo destaque para análises comparativas com genomas inteiros como ferramenta para a descrição de novas espécies. Recomenda-se ainda que seja realizada a comparação genômica por um ou mais parâmetros, atentando ao *cuttof* ANIb > 95% - 96% e TETRA > 0.998 - 0.999 a fim

de garantir circunscrições de espécies que deverão ser verificadas de acordo com a abordagem polifásica.

Chun et al., 2018 propõem uma nova abordagem para classificação de organismos bacterianos, combinando análise de *16S* rRNA e OGRIs, sugerindo padrões mínimos para o uso de dados do genoma para a taxonomia de procariontes a nível de espécie. Os autores recomendam que MLSA seja realizada para complementar a filogenia baseada no *16S* rRNA. No entanto, esse é um método pouco preciso diante da possível seleção arbitrária dos genes, já que não há um consenso sobre o número mínimo de genes para realizar tal análise e nem quais genes.

2 Justificativa

A incidência de difteria clássica mediada por toxinas tem apresentado uma queda considerável devido à vacinação generalizada. No entanto, existe uma necessidade de vigilância continuada, pois novos casos continuam sendo relatados mundialmente. O número de espécies do gênero *Corynebacterium* identificadas como causadoras de infecções oportunistas e hospitalares em humanos vêm crescendo e com isso, a necessidade de estudos e obtenção de conhecimento sobre as corinebactérias para os microbiologistas clínicos e profissionais da saúde.

Diversos estudos têm demonstrado a necessidade da identificação correta e específica de espécies bacterianas patogênicas emergentes causadoras de infecções multirresistentes. A identificação especifica da espécie e/ou gênero dirige a estratégia de medicação, o que tem levado a taxas satisfatórias de recuperação daqueles pacientes que receberam o tratamento apropriado (YANAI et al., 2018).

Em razão da dificuldade em realizar identificação desses patógenos apenas com testes fenotípicos, e alguns testes moleculares serem inviáveis e/ou inespecíficos para uso rotineiro em laboratórios, comparações baseadas no genoma total vêm sendo cada vez mais indicadas para tais análises, pois têm melhor poder de resolução na circunscrição de espécies (KÖSER et al., 2012; GLAESER & KÄMPFER, 2015). Nesse contexto, os OGRIs têm permitido uma rápida verificação da afiliação de espécies de dados genômicos públicos, permitindo a identificação de envios taxonomicamente errados nas deposições de genomas de bancos de dados públicos (RICHTER et al., 2016).

Nesse estudo foi desenvolvida uma estratégia que concatena os valores de ANI e TETRA, gerando uma assinatura única ‘*fingerprint’* para auxiliar a classificação específica, rápida e confiável de espécies bacterianas patogênicas emergentes causadoras de infecções multirresistentes, em associação aos métodos moleculares previamente desenvolvidos.

3 Objetivos

3.1 Geral

Desenvolver e validar uma estratégia para classificação rápida e confiável de espécies bacterianas patogênicas emergentes do gênero *Corynebacterium* baseada em uma assinatura combinada das métricas de ANI e frequência de tetranucleotídeos.

3.2 Específicos

- Avaliar a classificação de espécies patogênicas emergentes do gênero *Corynebacterium*;
- Comparar as análises filogenéticas convencionais e a classificação baseada na estratégia de *fingerprint* concatenando ANIb e frequência de tetranucleotídeos;
- Avaliar a estratégia de ANI e TETRA na definição de espécies intimamente relacionadas a *C. diphtheriae*;
- Avaliar os impactos das variações genômicas como, por exemplo, conteúdo G-C% e qualidade dos genomas (número de *contigs* e tamanho) na classificação taxonômica das espécies.

4 Materiais e métodos

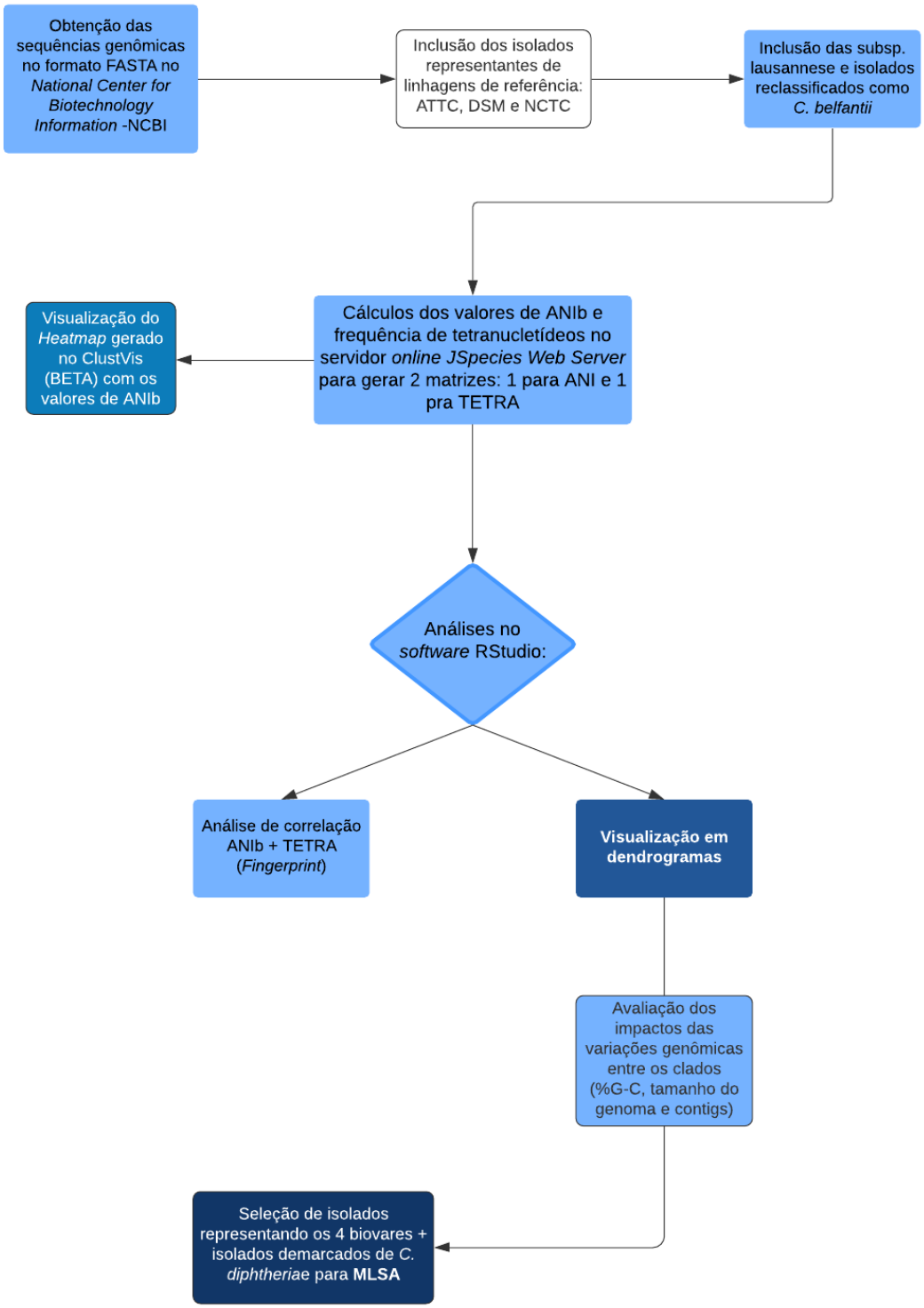


Figura 1 Fluxograma da metodologia utilizada nesse trabalho.

A metodologia utilizada neste estudo seguiu etapas de aquisição das sequências de *C. diphtheriae* disponíveis no NCBI, seguido da inclusão gradativa dos genomas das espécies *C. belfantii, C. rouxii* e as subespécies *lausannense* de *C. diphtheriae*. Posteriormente as sequências foram carregadas no JSpeciesWS para calcular ANI e TETRA e com esses valores realizar a construção dos dendrogramas no RStudio. Os valores do ANI foram utilizados para construção a parte de um *heatmap*. Com os dendrogramas avaliamos os impactos das variações genômicas como, por exemplo, conteúdo G-C% e qualidade dos genomas (número de contigs e tamanho) na classificação taxonômica das espécies. Adicionalmente realizamos a MLSA para comparar a nossa estratégia. Essa metodologia é apresentada resumidamente na Figura 1.

4.1 Aquisição das sequências de *C. diphtheriae* e isolados reclassificados

As sequências genômicas de *C. diphtheriae* utilizadas neste trabalho foram obtidas a partir do banco de dados de genomas do *National Center for Biotechnnology Informarion* - NCBI (https://www.ncbi.nlm.nih.gov). As sequências foram adquiridas no formato FASTA resultante da busca com a palavra-chave *Corynebacterium diphtheriae* após selecionar a opção *Genome* ao lado do campo de busca.

Com o limite de até 200 genomas no nosso projeto no JSpeciesWS (http://jspecies.ribohost.com/jspeciesws/), ambiente em que realizamos o cálculo dos valores de ANIb e TETRA, utilizamos como critério de inclusão em nossas análises as linhagens de referência com seus identificadores ATCC, DSM e NCTC e a representatividade de isolados dos *biovares* disponíveis. Incluímos bactérias bem caracterizadas a fim de verificar como a ferramenta se comporta utilizando dados de referência, nesse caso considerados dados controle. Além destes, adicionamos os isolados reclassificados como *C. belfantii* que inicialmente estavam classificadas como *C. diphtheriae*.

Adicionalmente, incluímos também os genomas dos isolados utilizados nos trabalhos de Tagini et al., (2018); Dazas et al., (2018); Pivot et al., (2019) e Badell et al., (2020). Para as sequências genômicas destes isolados que não estavam com o genoma montado utilizamos o *European Nucleotide Archive,* banco de dados em

que as sequências geradas nos estudos de Pivot et al., (2019) e Dazas et al., (2018) foram depositadas. Realizamos as buscas com os números dos projetos PRJEB28372 e PRJEB22103 fazendo a aquisição das sequências no formato Fastq, e posteriormente fizemos *upload* no serviço de montagem na plataforma de análise de genomas bacterianos PATRIC (DAVIS et al., 2020). Utilizamos o Serviço de Análise Abrangente do Genoma no PATRIC, onde é disponibilizada entre outros serviços a análise simplificada que aceita leituras brutas e executa uma análise abrangente, incluindo montagem e anotação de genomas a partir de dados de sequenciamento de nova geração. O serviço disponibiliza a escolha dos parâmetros para o meta-serviço em que o usuário seleciona a melhor estratégia para os seus dados. Fizemos o teste com as estratégias: Automática, *Spades* e *Unicycler* verificando posteriormente nos dendrogramas se esses genomas apresentavam alguma diferença no agrupamento com métodos diferentes de montagem.

4.2 Cálculo dos valores de ANIb e TETRA

Dentre os vários algoritmos que já incorporam o calculo de ANIb, nesse trabalho foi utilizado *JSpecies Web Server*, uma versão mais atualizada do servidor, disponível em: http://jspecies.ribohost.com/jspeciesws/#home, sendo um banco de dados de referência *on-line* de todos os genomas completos e preliminares publicados - cerca de 32.000 - com resultados pré-calculados de OGRIs (RICHTER et al., 2016). A frequência de tetranucleotídeos também foi calculada utilizando o servidor JSpeciesWS, em que cada genoma é comparado com todos os outros genomas estudados, gerando matrizes com os resultados dessas comparações que foram analisadas com a linguagem R através do ambiente RStudio.

4.3 Dendrogramas gerados pela *fingerprint* concatenando ANIb e TETRA

Utilizando *scripts* próprios analisamos os resultados em tabulares de ANIb e TETRA em que ambas as planilhas foram lidas gerando uma tabela única. Os valores de ANIb e TETRA de cada espécie foram concatenados e depois de normalizados, utilizadas como uma assinatura única ‘*fingerprint*’ de cada bactéria. Essas assinaturas foram utilizadas para gerar dendrogramas de relacionamento entre as espécies baseado em correlação de *Pearson.*

Inicialmente avaliamos a estratégia de classificação com 8 sequências genômicas e um valor de k = 5, onde k é o parâmetro que define o número máximo de *clusters* a se procurar. Nesta análise inicial utilizamos o método de distância "*euclidiana*" para calcular a relação entre os genomas e como método de agrupamento o método "*average*" através das funções inclusas no ambiente, *dist* e *hclust*.

Para avaliação da correlação entre os valores de ANIb e TETRA, os valores foram plotados como gráfico de dispersão e correlação computada pelo método *Spearman* usando a função incorporada '*cor.test*' em R associada ao pacote ggplot2 (WICKHAM, 2016). A normalidade dos dados foi avaliada usando R através da função incorporada '*shapiro.test*'.

A confiança de cada clado é calculada utilizando o pacote *pvclust* no RStudio, utilizando distância *euclidiana*, o agrupamento hierárquico através do método *Ward.D* e 1000 de replicatas de *bootstrap*. O número de clados com confiança (k) será definido baseado no número de clados com valor de *bootstrap* maior que 80%. O dendrograma final foi construído com os pacotes *factoextra* (KASSAMBARA; MUNDT, 2016), *ggdendro* (VRIES; RIPLEY, 2013) e *ggplot2* (WICKHAM, 2016) no ambiente R.

Com a visualização dos dendrogramas também avaliamos se alteração na quantidade de genomas alterava o agrupamento. Para isso incluímos gradativamente genomas com variações nos parâmetros (conteúdo G-C%, qualidade dos genomas - número de *contigs* e tamanho do genoma) avaliando a acurácia da ferramenta e verificando se os clados apresentavam diferença significativa para esses parâmetros.

4.4 Heatmap

O *heatmap* foi gerado na plataforma ClusVis Beta: https://biit.cs.ut.ee/clustvis/), uma ferramenta online para visualizar o *clustering* de dados multivariados. A planilha com os valores de ANIb foi importada para a ferramenta e posteriormente avaliamos o *heatmap* resultante da análise.

4.5 Reconstrução com *16S* e *rpoB* e Análise de Sequência Multilocus

A análise individual com *16S* rRNA e *rpoB* foi realizada no MEGA X aplicando o método de associação *Neighbor-Joining* com 34 sequências nucleotídicas. Esse método é fundamentado no agrupamento de vizinhos e tem como princípio encontrar pares de unidades taxonômicas operacionais que minimizem o comprimento total do ramo em cada estágio de agrupamento.

Para fazer a avaliação comparativa da Análise de Sequência Multilocus (MLSA) com a estratégia de correlação adicionamos 62 sequências nucleotídicas de 6 genes de manutenção sugeridos no banco de dados da nomenclatura internacional PubMLST (https://pubmlst.org/cdiphtheriae/info/protocol.shtml) para *C. diphtheriae*, nosso modelo de validação por ser um organismo amplamente estudado e com muitos genomas bem montados.

Após a aquisição dessas sequências no PATRIC (https://www.patricbrc.org/), foi criado um arquivo agrupando as sequências de todos os isolados para cada gene. Posteriormente foi realizado um alinhamento múltiplo no programa SeaView Versão 4 (GOUY, GUINDON, GASCUEL, 2010) utilizando o programa externo ClustalW. O SeaView é um editor de alinhamentos de múltiplas sequências que permite adicionar ou remover uma ou várias lacunas em uma ou várias sequências simultaneamente. Por fim, as sequências foram concatenadas para ser carregado um arquivo único no formato PHYLIP no ambiente PhyML (http://www.atgc-montpellier.fr/phyml-sms/) com seleção de modelo inteligente (SMS).

O SMS é uma ferramenta que usa estratégias para evitar o teste de todos os modelos e opções, simplificando alguns cálculos para economizar tempo de computação. SMS é um algoritmo simples, rápido e preciso para estimar grandes filogenias por máxima verossimilhança (LEFORT, LONGUEVILLE, GASCUEL, 2017).

4.6 Identificação genômica na plataforma Type (Strain) Genome Server (TYGS)

A plataforma TYGS (MUKHERJEE et al., 2017) permite a rápida análise de classificação e identificação genômica, fornecendo também valores de DDH e diferenças no conteúdo de G-C%. Utilizamos essa plataforma a fim de verificar a atual identificação genômica e comparar a estratégia de correlação.

Os dados da sequência do genoma são carregados no servidor: https://tygs.dsmz.de, para uma análise taxonômica baseada em todo o genoma. A análise TYGS é subdividida nas seguintes etapas: determinação de estirpes de tipos estreitamente relacionados; comparação pareada de sequências de genoma, inferência filogenética e agrupamento de espécies e subespécies de tipo. Nessas etapas os genomas são comparados contra todos os genomas de linhagens do tipo disponíveis no banco de dados TYGS, por meio do algoritmo MASH. Essas distâncias são usadas para determinar os 10 genomas de linhagem do tipo mais próximo para cada um dos genomas do usuário.

5 Resultados e Discussão

5.1 Sequências genômicas de *C. diphtheriae* e espécies reclassificadas

Atualmente (04 de outubro de 2020) existem 225 genomas de *C. diphtheriae* disponíveis nos bancos de dados do NCBI. Além dos 187 isolados de *C. diphtheriae*, adicionamos os 10 isolados reclassificados como *C. belfantii* e os isolados relacionados na Tabela 2 que estão em constante reclassificação conforme proposições pelos autores mencionados na coluna “Referências”. Dentre os vizinhos filogenéticos mais próximos de *C. diphtheriae* selecionamos *C. pseudotuberculosis* C231 que compartilhou ANIb=71% com a cepa de referência NCTC11397. Apresentamos também nesta tabela as informações dos conteúdos genômicos, parâmetros utilizados para avaliar se as variações genômicas interferem na clusterização.

Tabela 2 **Isolados em situação de reclassificação que foram adicionados às análises.**

Isolado	Espécie	Tamanho do genoma	GC (%)	Contigs	Localização geográfica	Referências
CCUG 5865	*C. diphtheriae*	2598402	53.63	132	London	Tagini et al., 2018
FRC 0043	*C. belfantii*	2609417	53.62	156	Ausente	Dazas et al., 2018
CMCNS703	*C. diphtheriae*	2725068	53.66	289	India: Vellore	Tagini et al., 2018
CHUV2995	*C. diphtheriae*	3060363	53.94	1	Switzerland	Tagini et al., 2018
FRC0190	*C. rouxii*	2451019	53.22	1	France	Badell, et al., 2020
FRC0074	*C. belfantii*	2609846	53.68	178	France	Pivot et al., 2019
FRC0318	*C. belfantii*	2614788	53.67	175		Pivot et al., 2019
FRC0382	*C. belfantii*	2611059	53.67	178		Pivot et al., 2019
FRC0381	*C. belfantii*	2608771	53.67	178		Pivot et al., 2019
FRC0223	*C. belfantii*	2715971	53.75	187	France	Pivot et al., 2019
FRC0455	*C. belfantii*	2609980	53.67	173		Pivot et al., 2019
FRC0301	*C. belfantii*	2701083	53.89	178	France	Dazas et al., 2018
06_4305	*C. belfantii*	2599771	53.74	190	France	Dazas et al., 2018
FRC0250	*C. belfantii*	2697075	53.69	167	France	Dazas et al., 2018
05_3187	*C. belfantii*	2713786	53.76	177	France	Dazas et al., 2018
00_0744	*C. belfantii*	2706250	53.88	176	France	Dazas et al., 2018
FRC0436	*C. diphtheriae*	2463590	53.67	65	France	Dazas et al., 2018

5.2 Dendrogramas gerados pela *fingerprint* concatenando ANIb e TETRA

Inicialmente testamos a estratégia de correlação avaliando o agrupamento com 8 genomas, conforme apresentado na Figura 2. Nesta análise inicial já foi possível observar uma inconsistência nos agrupamentos em que o isolado *C. belfantii* 631 se agrupa no mesmo clado que *C. diphtheriae* bv. mitis *strain* 2686, enquanto a cepa *C. diphtheriae* bv. mitis *strain* ISS 3319 fica em um clado separado. Adicionalmente observa-se também a demarcação do isolado CHUV 2995, agrupamento similar ao de grupos externos quando comparado com CMCNS 703 e CCUG 5865, sendo esses 3 últimos propostos como subespécie única *lausannense* de *C. diphtheriae* por Tagini et al., (2018).

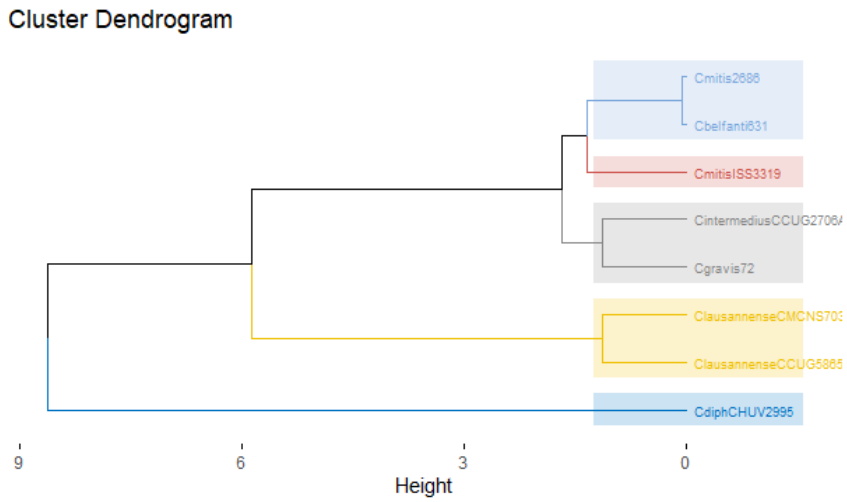


Figura 2 Dendrograma resultante da análise de correlação utilizando valores de ANIb e TETRA no *software* R incluindo 8 genomas. Nesta análise o valor de agrupamentos máximo aceito foi (k) = 5. *C. diphtheriae* subsp. *lausannense* CdipCHUV2995 é proposta como cepa tipo da subespécie *lausannense* e apresentou uma distância maior quando comparada aos outros isolados da mesma subespécie (ClausannenseCMCNS703 e ClausannenseCCUG5865), comportamento similar no agrupamento com inserção de genoma de grupos externos nas análises.

Dazas et al., (2018) caracterizaram os isolados do biovar belfanti por sequenciamento genômico e análises bioquímicas e quimiotaxonômicas propondo a reclassificação dos isolados do biovar belfanti. Dentre as justificativas dos autores, são apresentadas a incapacidade de reduzir o nitrato como uma característica bioquímica essencial que distingue *C. belfantii* de *C. diphtheriae* e o tamanho do genoma, pois os isolados de belfanti têm um genoma maior (2.7 Mb em média) e mais fragmentado (N50 médio, 37.5 kb) em comparação com os isolados de mitis e gravis (tamanho médio: 2.45 Mb, N50 médio: 163 kb). Além disso, as análises

filogenéticas indicam que este biovar representa um ramo claramente demarcado de *C. diphtheriae*.

Analisamos as variações nos conteúdos genômicos entre esses isolados que adicionamos à esta análise inicial e conforme dados apresentados na Tabela 3 podemos observar que o isolado adicionado em nossas análises reclassificado como *C. belfantii* não apresentam as variações correspondentes às propostas por Dazas et al., (2018) para esses parâmetros. Em contraste os isolados propostos como subespécies *lausannense* apresentam maior desvio do padrão apresentado pelos isolados de *C. diphtheriae*.

Tabela 3 Variações genômicas dos isolados analisados no dendrograma apresentado na Figura 2. O isolado 631 é um representante dos isolados do biovar reclassificado como *C. belfantii.*

Isolado	Tamanho do genoma	Contigs	GC%
C. diphtheriae bv. mitis *strain 2686*	2400684	55	53.6
C. diphtheriae bv. mitis str. ISS 3319	2366093	33	53.53
C. diphtheriae bv. belfanti strain 631	2397667	41	53.61
C. diphtheriae bv. intermedius strain CCUG 2706A	2367471	37	53.58
C. diphtheriae bv. gravis strain 72	2441855	28	53.48
C. diphtheriae subsp. *lausannense* CMCNS703	2725068	289	53.66
C. diphtheriae subsp. *lausannense* CCUG 5865	2598402	132	53.63
C. diphtheriae subsp. *lausannense* CHUV 2995	3060363	1	53.94

Para avaliar a acurácia da estratégia incluímos os genomas de todos os isolados reclassificados do biovar belfanti, das subespécies *lausannense* e duas cepas tipo de *C. diphtheriae*. De forma interessante, observamos na Figura 3 que os isolados formalmente reclassificados como *C. belfantii* formam um clado monofilético com as cepas tipos de *C. diphtheriae*, enquanto os isolados agrupados no clado sombreado em amarelo formam um clado parafilético com as cepas diftéricas.

Os isolados avaliados por Dazas et al., (2018) não correspondem àqueles previamente identificados como *C. diphtheriae* bv. Belfanti que apresentam o genoma com média de 2.404264bp, diferentemente dos iniciados com identificadores FRC, disponível no projeto PRJEB22103 e que apresentam tamanho do genoma em média 2.649816bp.

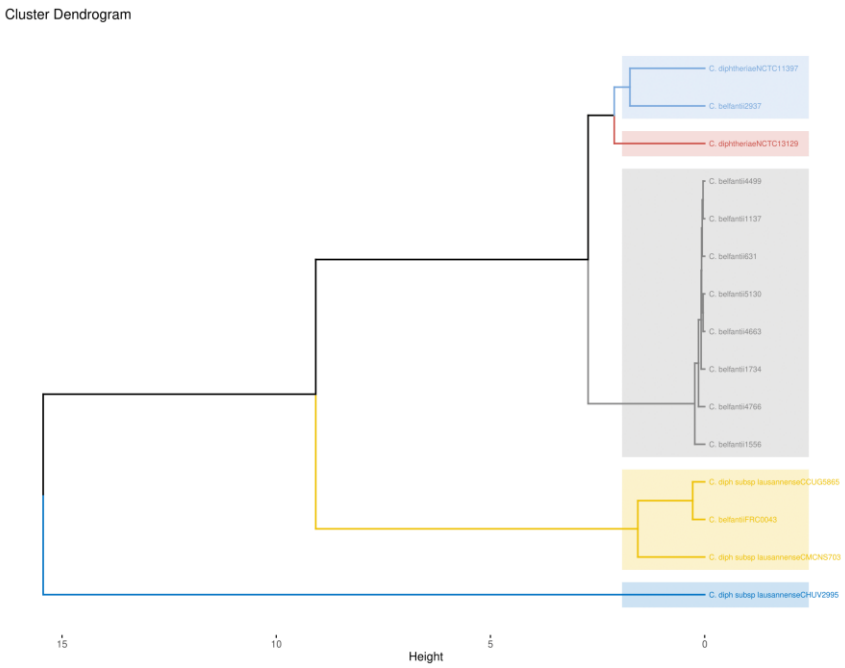


Figura 3 Análise com 15 genomas incluindo os isolados formalmente reclassificados como *C. belfantii*. A visualização deste dendrogramas permite observar a inconsistência no agrupamento do isolado 2937 de *C. belfantii*, enquanto a cepa tipo desta espécie FRC0043 forma um clado com o isolado CCUG5865 da subespécie *lausannense* de *C. diphtheriae*. Adicionalmente observa-se também a demarcação do isolado CHUV 2995 das demais subespécies propostas como *lausannense*, clusterização semelhante a grupos externos.

Embora os isolados do biovar belfanti tenham sido propostos como uma nova espécie *C. belfantii* (Dazas et al., 2018), nossos resultados sugerem que apenas a cepa tipo FRC 0043 pode ser reclassificada, pois além de apresentar valor de ANIb abaixo do ponte de corte até então estabelecido para circunscrição de espécies (ANIb = 94.92%) apresenta também o valor médio de 0.997 para a frequência de tetranucleotídeos quando comparada à cepa de referência *C. diphtheriae* NCTC11397, ficando consistentemente demarcada dos isolados das diftéricas em nossas análises.

Adicionalmente, a cepa FRC 0043 apresentou valor médio de ANIb = 94.79% e frequência de TETRA = 0.998 entre os isolados reclassificados para a espécie *C. belfantii.* Esse valor é considerado como limite, pois valores entre 0.989 e 0.999 não são considerados como bons e somente genomas com alta similaridade apresentam valor > 0.999, correspondendo a valores de ANIb >94%, lembrando que essa frequência tetranucleotídica é exclusiva para cada genoma, carregando um sinal específico (ROSSELLÓ-MORA; AMANN, 2015). Diferentemente da cepa tipo (FRC

0043), os isolados do biovar apresentam uma média de ANIb = 98.40% e TETRA = 0.999 com a cepa *C. diphtheriae* NCTC 11397.

Esses resultados nos permite sugerir que houve um equívoco ao reclassificar todos isolados do biovar como *C. belfantii,* tendo em vista também que além das características gerais não serem bem definidas de forma que não há uma exclusividade de um perfil bioquímico para um determinado biovar*,* a separação bioquímica é complexa e filogeneticamente confusa (BERNARD, 2012; SANTOS et al., 2018).

É possível observar também que as subespécies *lausannenses* não se agrupam conforme a atual proposição. Estes isolados foram propostos como subespécie de *C. diphtheriae* por Tagini et al., (2018) ao realizar análise genômica comparativa utilizando ANI e observar que CHUV 2995 compartilha ANI mediano = 95.25% com todas as cepas diftéricas e em contraste, CHUV compartilha ANI >99% com CMCNS703 e CCUG5865. Além disso, eles realizaram análise com o gene *16S* rRNA, em que foi observado a conservação do gene entre todas as cepas com identidade >99% e por isso eles sugerem que estes isolados estão intimamente relacionados e que o clado demarcado deve ser classificado como subespécie de *C. diphtheriae*. No entanto, o *16S* é bem conhecido pela sua limitada resolução filogenética na utilização para circunscrição de espécie ou subespécies (NA et al., 2018).

No trabalho de Tagini et al., (2018) foi adicionado *C. ulcerans* BR-AD22 como grupo externo e foi realizada uma reconstrução filogenética com máxima verossimilhança para confirmar a monofilia da nova subespécie conforme proposta por eles. Foi utilizado um *core genome* comum a todas as cepas de *C. diphtheriae*, e também para *C. ulcerans*. Eles afirmam que a árvore filogenética demonstra consistência com os cálculos do ANI, além de afirmar que a cepa CHUV2995 se agrupa com CCUG 5865 e CMCNS703. Com isso, eles defendem que as cepas mencionadas representam monofilia e os membros da subespécie *diphtheriae* formam outro clado distinto.

Também adicionamos o genoma de *C. ulcerans* BR-AD22 em nossas análises em que ao adotar a nossa estratégia de correlação de ANI e frequência de tetranucleotídeos pudemos perceber uma diferença nos clusters quando

comparados ao trabalho realizado por Tagini et al. (2018), que utilizaram uma abordagem com riscos arbitrários, conforme apêndice A.

Badell et al., (2020) sugerem CHUV2995 como *C. belfantii*, pois os valores entre CHUV e FRC 0043 (cepa tipo de *C. belfantii*) é 99.3% e CHUV foi posicionado dentro do ramo filogenético de *C. belfantii*. Complementarmente, destacamos a atual discussão sobre o ponto de corte generalizado do ANIb, tendo em vista que esse valor por si só não tem refletido bem a circunscrição das espécies bacterianas. Palmer et al., (2020), discutem a integração do valor de ANI na taxonomia polifásica e afirmam que os métodos ANI não forneceram resultados consistentes em relação à co-especificidade dos isolados que eles analisaram incluindo três classes nas Proteobactérias e sugerem a necessidade inicial de determinar um valor de corte apropriado para um conjunto de táxons específico.

Observamos na Figura 4 que os isolados da espécie *C. belfantii* (identificados com iniciadores FRC) formam consistentemente um clado distinto, assim como CHUV 2995 e FRC0190. Diferentemente daqueles isolados do biovar belfanti que se mantém nos clados de *C. diphtheriae*. A frequência de tetranucleotídeos é considerado um parâmetro complementar para identificar genomas a nível de espécie e comparando os resultados dos trabalhos mencionados anteriormente e a clusterização visualizada em nossos dendrogramas percebemos como a frequência de tetranucleotídeos contribui significativamente na clusterização.

O estudo realizado por Pivot et al., (2019) investigou a possibilidade de transmissão cruzada de *C. diphtheriae* biovar belfanti entre 4 pacientes de um centro de Fibrose Cística. Eles encontraram 5 isolados sendo que nenhum era toxigênico e todos pertenciam ao biovar belfanti. O MLST mostrou que os cinco isolados pertenciam ao mesmo tipo de sequência e a variação da sequência do genoma inteiro entre os 5 isolados revelou apenas 62 SNPs entre eles, estando intimamente relacionados, mostrando que eles pertencem a uma única cepa.

Figura 4 Dendrograma utilizando 102 genomas e máxima distância *'euclidiana'* de 2. Neste dendrogramas observa-se a consistência do agrupamento no clado sombreado em vermelho com os isolados de *C. belfantii,* e a demarcação de CHUV2995 e FRC0190. É possível observar também a demarcação dos isolados WD37 e CMCTH1443 da espécie *C. diphtheriae*, sendo este último identificado no NCBI como contaminação.

Os isolados avaliados no trabalho de Pivot et al., (2019) foram adicionados as nossas análises e estão identificados na Figura 4 com os iniciadores FRC, sombreados em vermelho. Eles realizaram análise filogenética baseada em SNPs, usando a cepa FRC0223, considerada a mais estreitamente relacionada como grupo externo. O resultado revelou três subtipos, sendo que dentro dos subtipos, apenas 18 SNPs separavam FRC0318 e FRC0381 e apenas 4 SNPs separavam FRC0382 e FRC0455.

Assim como no trabalho de Pivot et al., (2019) em nossas análises o nível médio de identidade nucleotídica desses isolados com a cepa tipo *C. belfantii* foi de 99.4%, enquanto com a cepa tipo de *C. diphtheriae* NCTC11397 foi de aproximadamente 95%. Os autores afirmam que os cinco isolados pertencem a nova espécie *C. belfantii*. De acordo com eles, a variação do SNP descoberta pela análise genômica reflete a evolução da cepa desde o último ancestral comum dos cinco isolados e a distinção dos 3 subtipos pode refletir a transmissão direta entre eles.

Para avaliar os resultados dos agrupamentos somente com ANI fizemos uma matriz de correlação com esses valores, incluindo representantes de *C. diphtheriae*, *C. belfantii*, *C. rouxii.*

5.3 Matriz de correlação ANIb apresentada como *Heatmap*

Usando apenas a correlação do ANIb entre os isolados (116 genomas) apresentamos na Figura 5 foi realizada uma clusterização hierárquica dos valores visualizado em forma de *heatmap*. Neste *heatmap* podemos observar que *C. diphtheriae* subsp. *lausannense* CHUV 2995 se agrupa aos demais isolados identificados com FRC e com os isolados propostos como subespécies *lausannense*, conforme sinalização em vermelho.

No entanto, essa alta correlação apresentada no mapa de calor não reflete a inconsistência observada nos nossos dendrogramas e nem as particularidades desses isolados. Avaliando os valores de ANIb e TETRA individualmente percebemos que apesar de CHUV ter compartilhado uma média de ANIb = 99.18% com os isolados da espécie *C. belfantii*, o valor de TETRA apresentou uma média de 0.994.

Recentemente Badell et al., (2020) propuseram o nome *C. rouxii* sp. nov. para o isolado identificado como *C. diphtheriae* FRC 0190, demonstrando que FRC0190 forma um clado único com os outros isolados de *C. rouxii* formando uma espécie independente. Este isolado que inicialmente foi classificado como *C. diphtheriae* se manteve demarcado em todos os nossos dendrogramas, ressaltando a acurácia da estratégia de correlação. Outro isolado que nos permite sugerir que o ANIb sozinho apresenta uma performance subótima para a circunscrição de espécies e os atuais valores aceitos precisam ser ajustados é o isolado CMCTH 1443, destacado em amarelo na Figura 5, que além de apresentar tamanho do genoma = 2772004bp e conteúdo G-C 51.49% é identificada a contaminação na anotação do seu genoma.

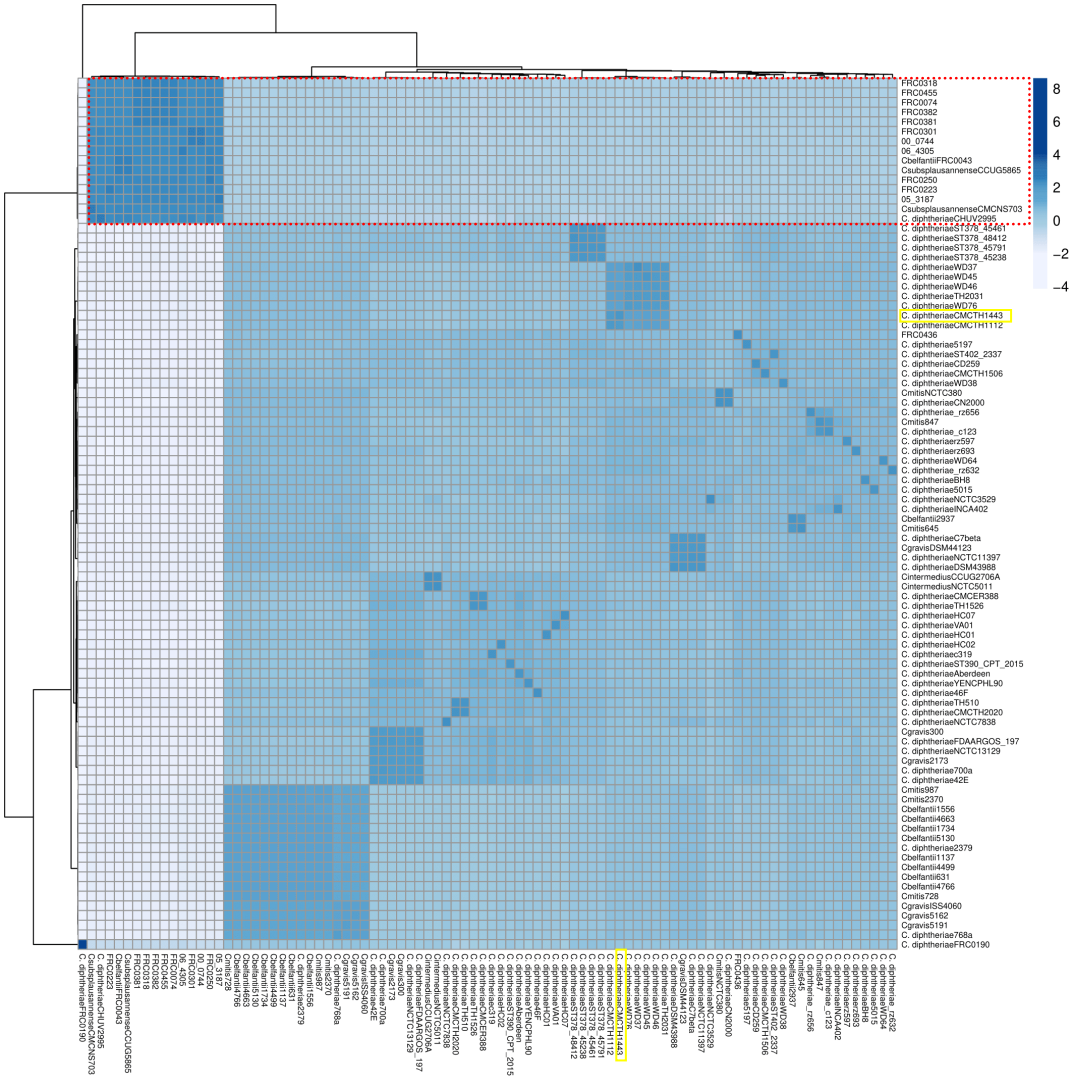


Figura 5 Heatmap com 166 genomas, incluindo *C. belfantii*, *C. rouxii* e *C. diphtheriae*. O isolado CMCTH1443 destacado em amarelo corrobora a dificuldade em identificar corretamente os isolados de *C. diphtheriae* apenas com ANI, visto que este isolado representa contaminação.

No trabalho de Badell et al., (2020) é apresentada uma reconstrução filogenética com os isolados das espécies *C. rouxii* (chamados isolados maltose atípicos), *C. diphtheriae*, *C. belfantii*, *C. ulcerans* e *C. pseudotuberculosis*. Na reconstrução são visualizados 5 clados, e o isolado CHUV 2995 está incluso no clado da espécie *C. belfantii.* Segundo os autores a estirpe do tipo *lausannense* enquadra-se em *C. belfantii* e os isolados atípicos compreendem um clado geneticamente homogêneo entre si.

O isolado FRC 0190 forma um clado genômico distinto, separado por um alto nível de divergência de nucleotídeos do limite atual estabelecido para o limite de espécies genômicas (~ 94-96%). Já o isolado CHUV 2995 apresenta ANI = 99.3% com *C. belfantii*, sendo proposto como sinônimo heterotípico posterior de *C. belfantii*. Os autores afirmam ainda que a análise filogenética com os genes *rpoB* e *16S* rRNA foi consistente com a distinção dos isolados atípicos. Contudo, ressaltamos mais uma vez a limitada resolução filogenética a nível de espécies com esses marcadores e a necessidade da utilização de sequências de genoma para fins taxonômicos. Nas figuras 6 e 7 podemos observar também a demarcação deste isolado.

Em ambos dendrogramas visualizados nas figuras 6 e 7 as observam-se as demarcações dos isolados FRC 0190, CHUV 2995, CMCTH 1443, WD37, CMCNS 703, CCUG 5865 e os isolados da espécie *C. belfantii* com os identificadores FRC (exceto a cepa FRC 0436 que é *C. diphtheriae* e foi adicionada para verificar a qualidade da montagem de genomas que utilizamos no PATRIC) utilizada no trabalho de Badell et al., (2020). Em todas as nossas análises mesmo com variação de quantidade de genomas, alternância dos grupos externos e alteração no valor da distância informada na estratégia, esses isolados se mantiveram demarcados.

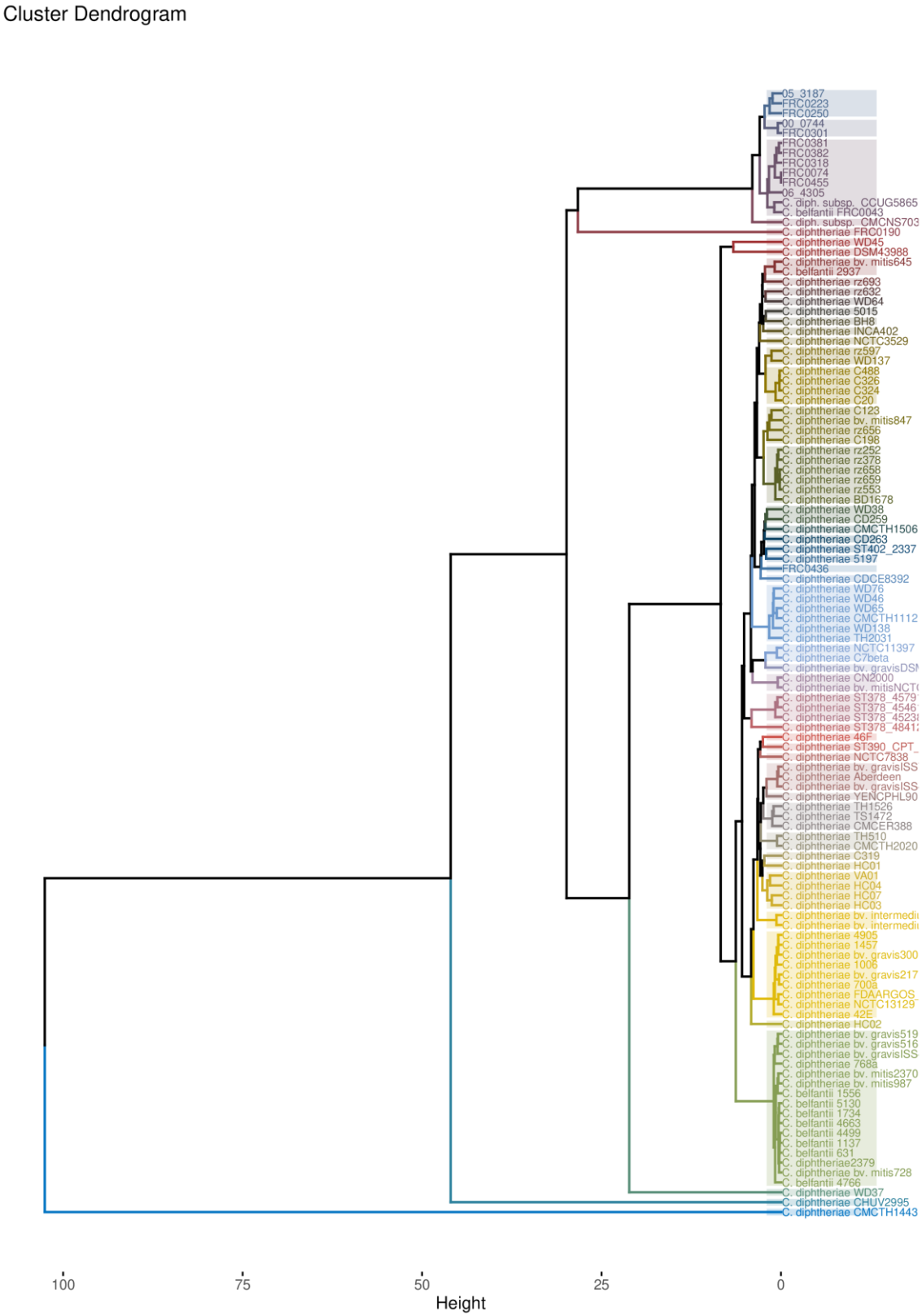


Figura 6 Dendrograma com 114 genomas, com distância *euclidiana* de 2 e sem grupo externo. Nas análises sem grupo externo é possível observar que os clados dentro da espécie *C. diphtheriae* apresentam maior resolução.

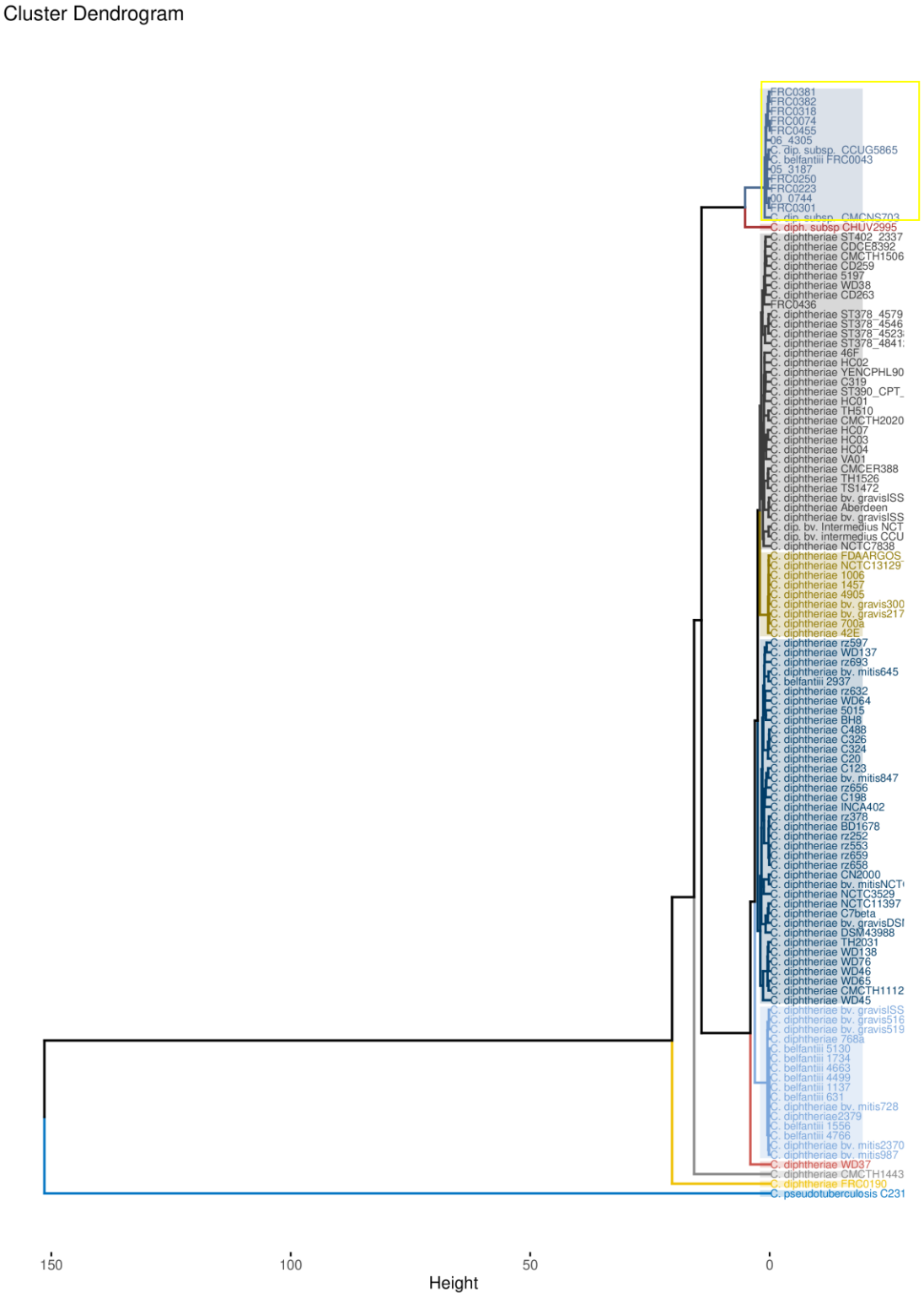


Figura 7 Dendrograma com 115 genomas, com distância *euclidiana* de 2. Nesta análise incluímos as linhagens francesas identificadas pelos iniciadores FRC, isolados dos biovares de *C. diphtheriae*, as subespécies *lausannense*; os isolados propostos por Dazas como *C. belfantii*, e a nova proposta de *C. rouxii* (BADEL, 2020). Nesta análise visualizamos os agrupamentos com adição de grupo externo.

Na Figura 6 não foi adicionado o grupo externo e podemos perceber que há uma maior subdivisão entre os clados. Não obstante, nota-se também que as análises sem o grupo externo permitem a visualização de grupos clonais. Na Figura 7 com *C. pseudotuberculosis* C231 como grupo externo é possível observar uma melhor definição de grupos, que corrobora as análises filogenéticas e filogenômicas, não havendo nenhum representante que não fosse esperado nos clados de *C. diphtheriae*. Claramente observamos um clado demarcado agrupando todos os isolados de *C. belfantii* e diferente do que foi proposto por Badell et al., (2020). CHUV 2995 não se agrupa em nenhuma das análises utilizando a estratégia de correlação entre ANIb e TETRA.

As recentes classificações para CHUV 2995 têm se baseado no valor de ANI e similaridade do *16S* rRNA. Nas discussões sobre a reclassificação para este isolado apresentadas por Tagini (2018) e Badell (2020) e colaboradores têm sido fortemente sugeridas a inclusão em *C. belfantii*. Tagini sugere que a linhagem 1 de *C. diphtheriae* (compreendendo a maioria das linhagens) seja renomeada para *Corynebacterium diphtheriae* subsp. diphtheriae e a linhagem 2 de *C. diphtheriae* (reagrupando somente as linhagens biovar belfanti) seja renomeada para *Corynebacterium diphtheriae* subsp. *lausannense*.

No trabalho de Tagini et al., (2018) os autores relatam que CHUV 2995 juntamente com os outros isolados da subespécie *lausannense* não apresentou nenhum código genético para os operons clássicos (ou associados a ele) de *C. diphtheriae*, bem como carecem de genes que codificam a redutase do nitrato. Além disso, os genomas das subespécies *lausannenses* são significativamente maiores (CHUV2995: 3.06Mb; CCUG 5865: 2.6Mb; CMCNS703: 2.73Mb) do que os genomas diftéricos. Eles realizaram o teste *t* comparando o conteúdo G-C% entre os dois clados e o resultado não foi significativo ($p = 0.1555$), mas CHUV 2995 apresenta o conteúdo G-C= 53.94% diferentemente das cepas diftéricas que apresentam uma média 53.54 ± 0.16%.

Com a visualização dos dendrogramas e a identificação dos isolados consistentemente demarcados, avaliamos se existia correlação entre o agrupamento de cada clado com a variação no conteúdo genômico. Buscando estabelecer

melhores parâmetros de distância em nossa estratégia de forma que a visualização no dendrograma represente maior acurácia nos agrupamentos, alternamos o valor da distância entre os clados. Na Figura 8 observamos o dendrograma gerado com a distância *euclidiana* máxima entre os elementos do clado igual a 2 (d=2) e com inclusão do grupo externo. Comparando-a com a Figura 9, em que aumentamos a distância (d=4) podemos observar uma maior subdivisão entre os isolados de *C. diphtheriae* mesmo tendo inclusão do grupo externo.

Nas análises com *C. pseudotuberculosis* C231 e a distância no nosso script = 4 nomeamos como Clado 1, 2, 3 e clados separados para aqueles que não se agruparam, conforme Figura 9**.** Com isso avaliamos se existia diferenças significativas nos parâmetros que justificassem essa subdivisão. Para isso avaliamos individualmente os parâmetros de cada um dos isolados destes clados de e essas análises são apresentadas nos gráficos da Figura 10.

Fizemos várias análises alternando o grupo externo, a distância entre os clados e a quantidade dos genomas de *C. diphtheriae*. Os dendrogramas gerados podem ser observados nos apêndices A à C. Com os valores de GC%, tamanho do genoma e número de contigs comparamos cada clado avaliando se existia diferença entre estes parâmetros que justificassem um isolado do biovar Mitis ficar em um clado e outro isolado em outro por exemplo. O grupo externo escolhido para ser adicionado às nossas análises foi *C. pseudotuberculosis* C231 que compartilhou ANIb médio = 71.08%. *Mycobacterium tuberculosis* ATCC35801, *Rhodococcus equi* 103S, *C. accolens* ATCC49725, *C. simulans* PES1, *C. minutissimum* ATCC23348 e *C. aurimucosum* 1237CAUR compartilharam valores abaixo de 70%.

A Figura 10A apresenta as variações no tamanho dos genomas para os clados 1, 2, 3 e os isolados identificados como clados separados, e a Figura 10B apresenta as variações entre conteúdo GC% e o número de contigs. Destacamos que mesmo o clado 3 apresentando predominantemente representatividade dos isolados do biovar gravis não foi possível observar diferença significativa entre os parâmetros quando comparados aos outros clados.

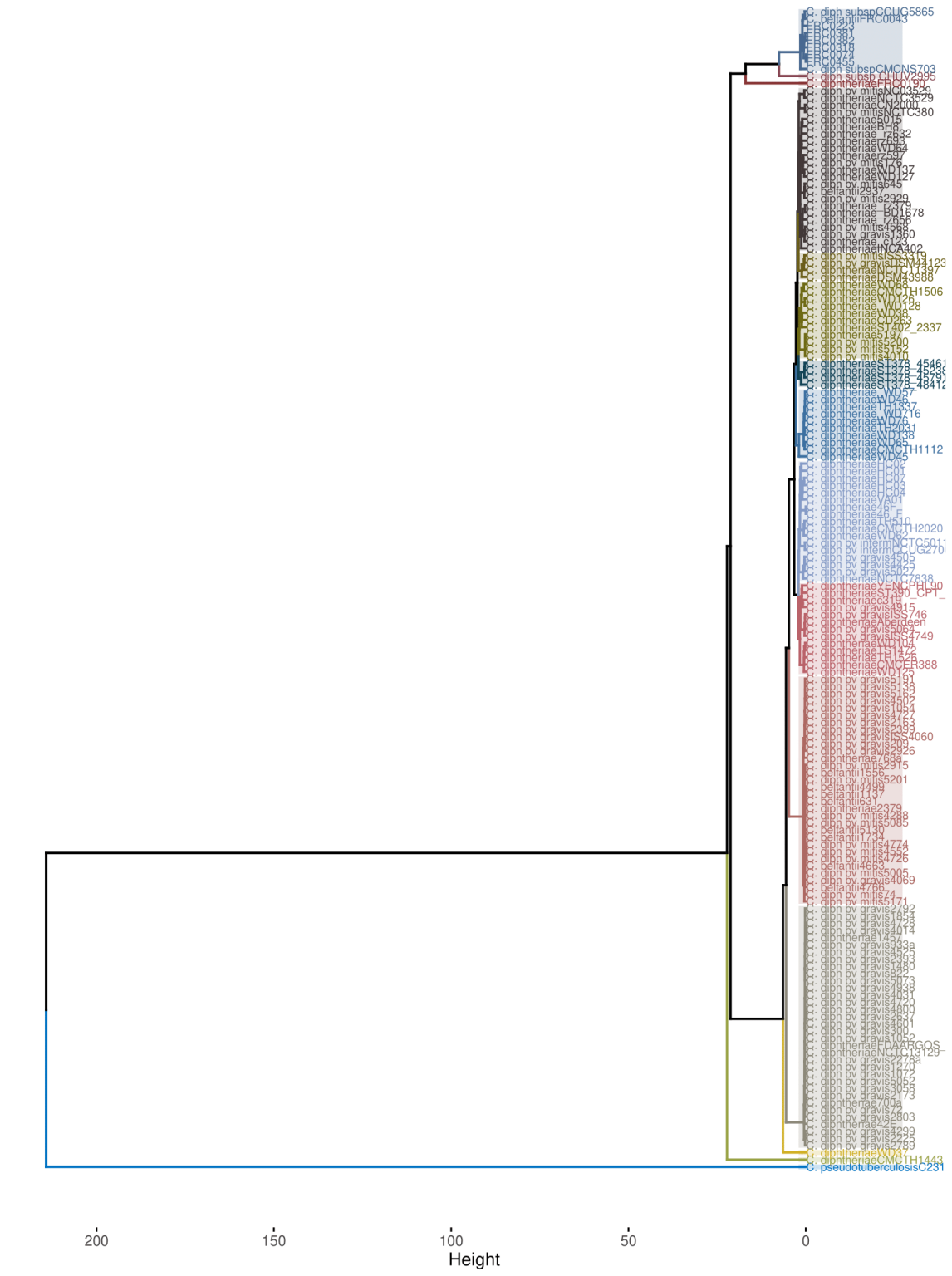


Figura 8 Dendrograma com 162 genomas e distância 'euclidiana' de 2. Nesta análise aumentamos a quantidade de genomas e verificamos que dentro da espécie *C. diphtheriae* os agrupamentos apresentam maior subdivisão quando comparado às análises com distância igual a 4.

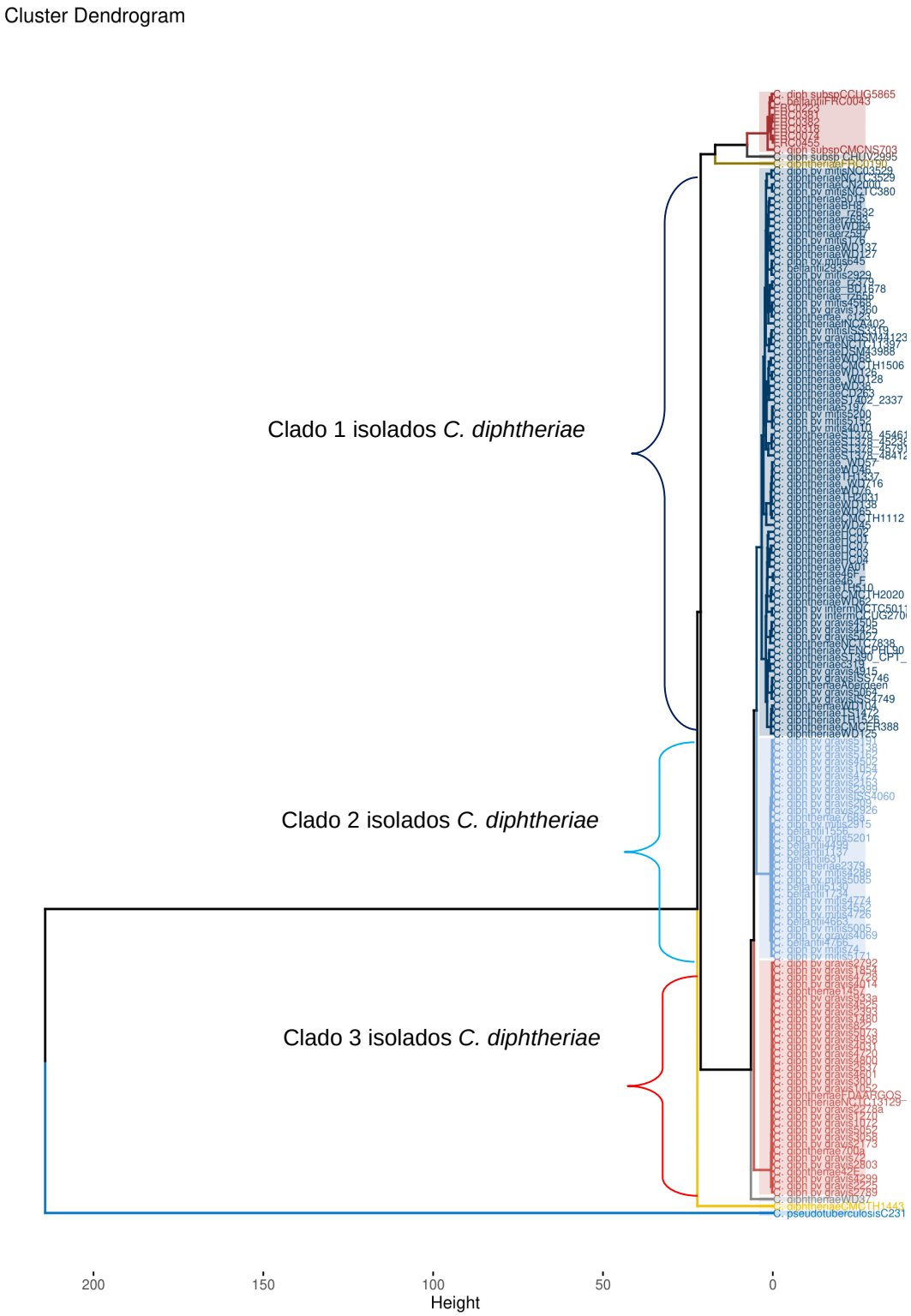


Figura 9 Dendrograma com 162 genomas e distância 'euclidiana' igual a 4. Neste dendrogramas com a distância *euclidiana* modificada observa-se que os isolados dentro da espécie *C. diphtheriae* forma menos clados quando comparada ao dendrogramas com a distância =2. Nomeamos os clados como 1, 2, 3 e clados demarcados para avaliar a influência das variações gênicas entre os clados.

Figura 10 Comparação entre as variações dos conteúdos genômicos no dendrograma com 162 genomas e identificados na Figura 9 como clado 1, 2, 3 e clados separados.

Ainda analisando os isolados descritos como *C. diphtheriae* em ambos dendrogramas é possível observar que *C. diphtheriae* CMCTH 1443 se mantém

demarcado das demais mesmo com grupo externo em que se observa consistentemente um agrupamento mais uniforme. Diante dessa demarcação avaliamos as suas características e constatamos que este resultado reforça a acurácia da ferramenta de correlação entre ANIb e TETRA, pois está registrada a contaminação da mesma, e o valor do GC é igual a 51.49%, enquanto *C. diphtheriae* apresenta em média 53.54%. Esta é uma identificação potencialmente não confiável, já que dentro das espécies o conteúdo não varia mais que 1%.

O isolado *C. diphtheriae* WD37 também se manteve em um clado demarcado das diftéricas nas análises com e sem grupo externo. No entanto, não há registro de contaminação e os valores dos conteúdos genômicos variam pouco quando comparados com as diftéricas. Comparando *C. diphtheriae* NCTC 11397 com WD37 os valores são respectivamente: tamanho do genoma: 2463666bp e 2601759bp; contigs: 1 e 389 e GC% 53.52 e 54.14.

Com exceção à cepa FRC 0043, os isolados reclassificados como *C. belfantii* apresentaram uma média de ANIb = 98% com as cepas de referência de NCTC3529; NCTC7838; NCTC11397 e NCTC13129. Para a frequência de tetranucelotídeos o valor foi = 0.999227. Em contraste, FRC 0043 apresentou média para ANIb = 94.66% e para TETRA = 0.99801 com as mesmas cepas de referências mencionadas anteriormente. A representação dessas divergências pode ser observada no dendrograma ao visualizar o agrupamento com os isolados juntamente ao com identificadores FRC e a demarcação do clado com *C. belfantii* nos levando a questionar tais reclassificações.

Comparando os agrupamentos visualizados nos dendrogramas através da alteração na quantidade de genomas, apresentamos na Figura 11 a visualização com 115 genomas e a distância = 4. Nesta figura observamos que ao aumentar a distância estabelecida para o agrupamento, as pequenas diferenças apresentadas entre os isolados não são suficientes para a subdivisão dentro da espécie. No entanto, é possível sugerir que este valor promove a melhor separação a nível de espécie.

Neste dendrograma com a distância = 4 ainda observamos os isolados WD37 e CMCTH1443 separados das cepas de *C. diphtheriae*. A cepa WD37 embora não apresente registro de contaminação como CMCTH 1443, apresenta baixa qualidade

no genoma, tamanho do genoma = 2601759bp e o conteúdo GC =54.14%. Dentro dos clados de *C. diphtheriae* os valores apresentaram média de 2.4Mb e 53.54% para tamanho do genoma e conteúdo GC, respectivamente. Outro isolado que apresentou valores divergentes da média para *C. diphtheriae* é o isolado WD45, apresentando conteúdo GC = 52.98% e o tamanho do genoma 2583616bp. No apêndice B observa-se que o isolado WD45 não se agrupa aos demais isolados de *C. diphtheriae*.

Rosseló-Moraá e Amman (2015) recomendam o abandono do DDH, e afirmam que a MLSA não é considerada competitiva para análise de coerência genômica, bem como a comparação dos genomas por um ou vários parâmetros sendo o ANIb >96% e TETRA >0.999 limites que garantem circunscrições de espécies que precisarão ser verificadas por estudos fenotípicos. Porém, eles afirmam que o uso de ANIb para genomas distantes é questionável, pois diante da complementaridade ser baixa apenas pequenas partes dos genomas são comparadas.

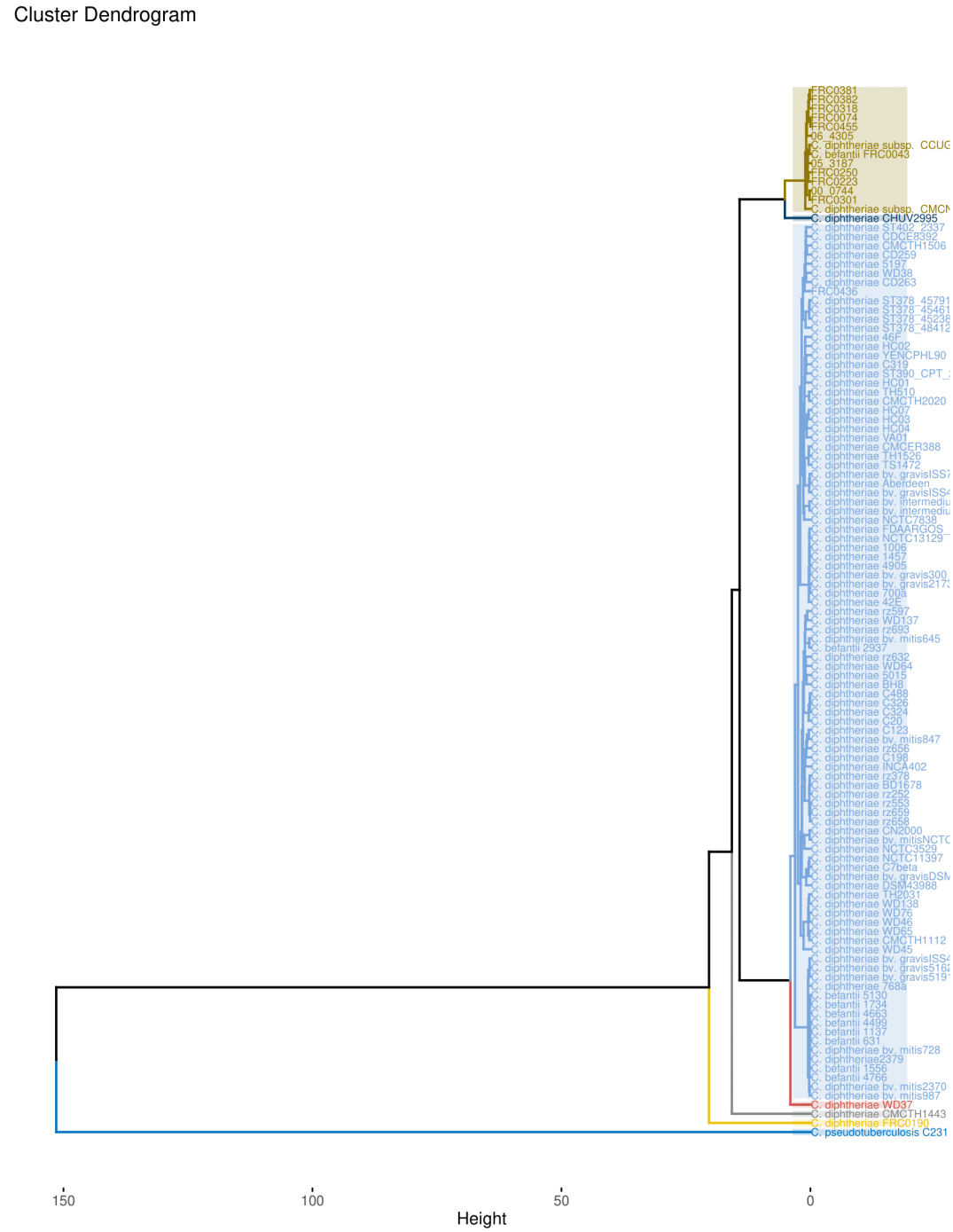


Figura 11 Análise com 115 genomas e distância 'euclidiana' igual a 4. Neste dendrograma observa-se que a mudança na quantidade de genomas juntamente à distância provocou a visualização de um clado único para a espécie *C. diphtheriae.* Na análise com 162 genomas e essa mesma distância os clados dentro de *C. diphtheriae* apresentaram mais subdivisão.

5.4 Análise das relações filogenéticas utilizando *16S* rRNA, *rpoB* e MLSA

Na maioria das vezes, uma nova cepa é primeiro alocada filogeneticamente com base na análise do gene *16S* rRNA a nível de gênero. No entanto, reconstruções filogenéticas com árvores monogênicas utilizando o gene *16S* rRNA como marcador filogenético apresentam como desvantagem uma resolução insuficiente a nível de espécie. Resoluções filogenéticas mais altas, especialmente a nível de gênero, podem ser obtidas através da realização de análises filogenéticas adicionais baseadas em genes que codificam para proteínas (GLAESER; KÄMPFER, 2015; ROSSELLÓ-MÓRA; AMANN, 2015).

Em contraste com o gene *16S* rRNA, os genes codificadores de proteínas apresentam como vantagem a taxa lenta de evolução, conferindo assim um melhor poder de resolução, especialmente no nível do gênero ou mesmo abaixo. No caso de espécies filogenéticas próximas genes como o *rpoB* são recomendados para apoiar ainda mais a autenticidade dos dados do genoma. Buscando uma melhor resolução filogenética entre as espécies dentro de um gênero a MLSA tem sido sugerida para este fim, pois pode resolver problemas filogenéticos entre espécies intimamente relacionadas que não podem ser diferenciadas com base no *16S* rRNA (GLAESER; KÄMPFER, 2015).

A MLSA minimiza o viés gerado em reconstruções filogenéticas monogênicas. Geralmente são analisados aqueles que codificam subunidades de enzimas presentes em todos os táxons, como por exemplo, a subunidade da DNA gyrase (*gyrB*), a subunidade beta da RNA polimerase (*rpoB*), recombinase A (*recA*), entre outros (GLAESER; KÄMPFER, 2015; ROSSELLÓ-MÓRA; AMANN, 2015). Para fazer a avaliação comparativa da MLSA com a estratégia de correlação testamos as reconstruções baseadas nas análises monogênicas e a MLSA utilizando os genes *ATP synthase alpha chain* (atpA); *DNA polymerase III alpha subunit* (dnaE), *Chaperone protein dnaK; Translation elongation factor G (fusA)*; *2-isopropylmalate synthase* (leuA) e *DNA-directed RNA polymerase beta subunit (rpoB).*

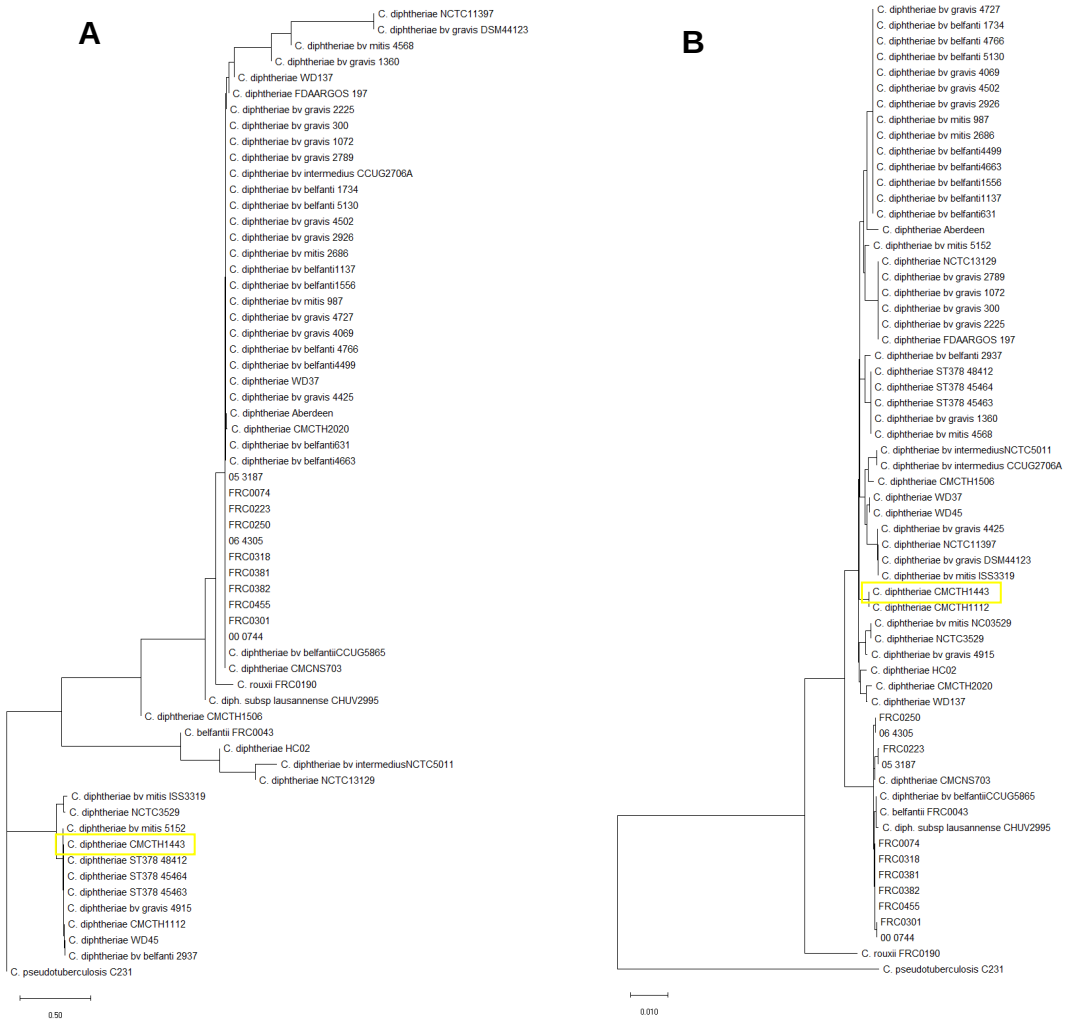


Figura 12 Reconstrução filogenética inferida usando o método Neighbor-Joining com os genes *16S* rRNA em A e *rpoB* em B. Esta análise envolveu 62 e 61 sequências nucleotídicas respectivamente, pois para o isolado *C. diphtheriae bv mitis* NC03529 não foi encontrado o gene *16S* rRNA bem anotado.

Conforme apresentado na Figura 12 A, podemos observar que as espécies *C. rouxii* e *C. belfantii* (representada pelos iniciadores FRC) não ficaram bem demarcadas na reconstrução filogenética monogênica com o gene *16S* rRNA. Diferente da clusterização visualizada na Figura 12 B, em que se observa uma melhor representação do isolado *C. rouxii*. Esses dados reforçam a limitação na resolução filogenética utilizando o *16S* rRNA para reconstruções abaixo do nível de gênero.

Avaliando a MLSA na Figura 13 é possível observar a subdivisão dos clados com os isolados de *C. belfantii* conforme sinalização em azul e dos isolados de *C. diphtheriae* em verde. Além disso, claramente observa-se a formação de um clado isolado para *C. rouxii* ratificando a proposição desta como nova espécie no trabalho de Badell et al., (2020).

Nessa figura podemos observar que a estratégia de correlação corrobora os agrupamentos visualizados na MLSA. Contudo, a análise de MLSA é um método custoso computacionalmente e difícil de desenvolver, já que não há um consenso sobre o número mínimo de genes a se utilizar e a abordagem de MLSA provou ser incômoda pela possibilidade de seleção arbitrária de genes Dessa forma, consideramos nossa estratégia uma ferramenta com maior acuracidade e com potencial para auxiliar a classificação de patógenos emergentes do gênero *Corynebacterium*, visto sua rapidez e facilidade de modificar os parâmetros, bem como não precisar recalcular todos os parâmetros, somente novos elementos que vão sendo incluídos.

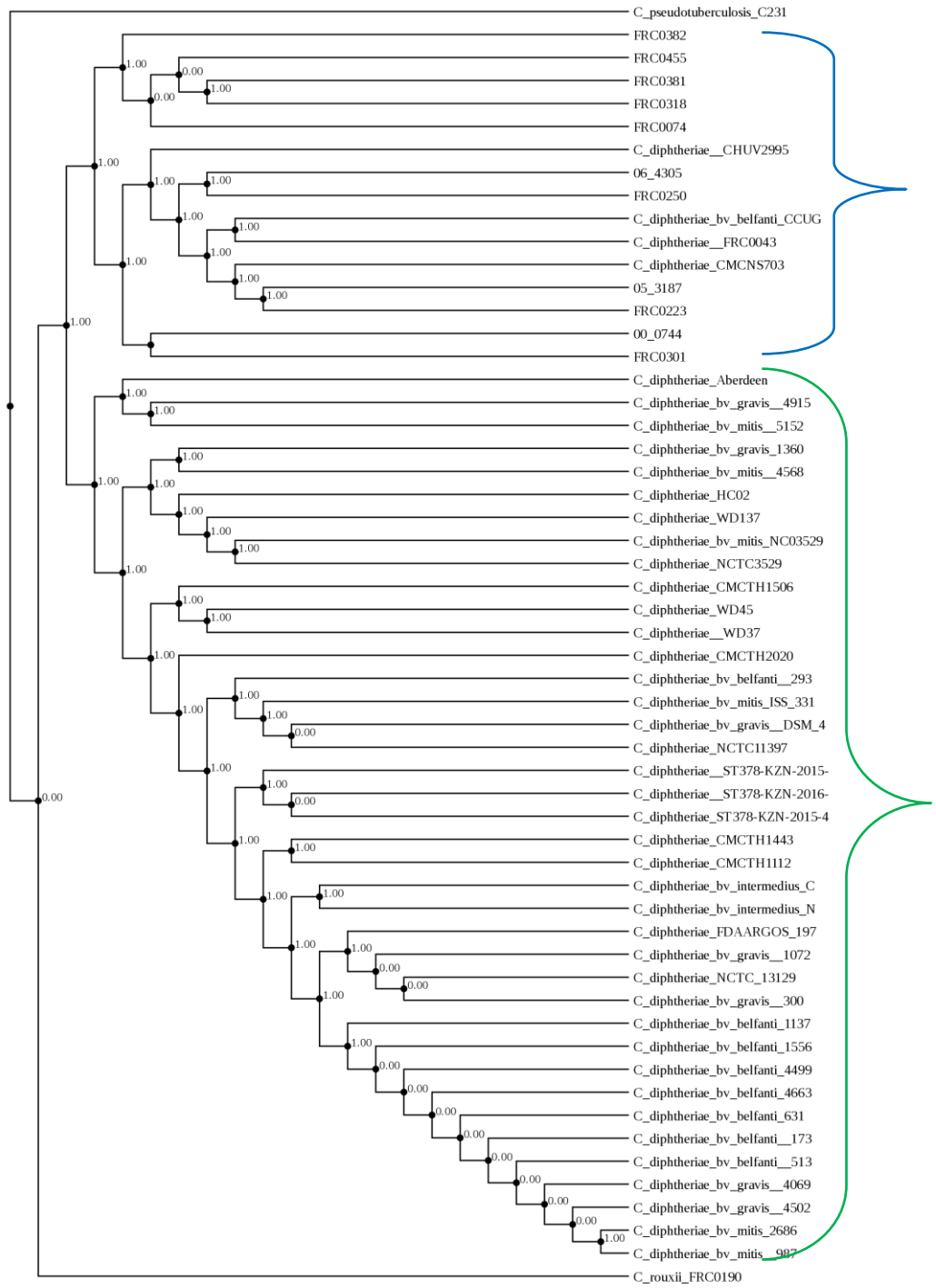


Figura 13 Árvore filogenética inferida com sequências dos genes *atpA*, *dnaE*, *dnaK*, *fusA*, *leuA* e *rpoB*. Esta árvore foi gerada pelo Smart Model Slection no ambiente PhyML com as sequências dos genes concatenadas após curadoria no SeaView.

5.5 Identificação genômica utilizando a plataforma TYGS

Além das análises apresentadas nos tópicos anteriores, nós ainda consultamos plataformas de identificação genômica, selecionando a Type (Strain) Genome Server para apresentar os resultados que obtivemos na consulta. A pesquisa apresentou como conclusão para os isolados FRC0043, CHUV2995, CCUG5865 e CMCNS703 a sua indicação de espécies conhecidas e identificadas como *C. belfantii*. Os isolados CMCTH1443, WD37, WD45 foram identificadas como *C. diphtheriae*. No entanto, CMCTH1443 apresenta a observação que os valores obtidos das comparações pareadas do genoma com a cepa tipo indicam um resultado de identificação potencialmente não confiável e, portanto, devem ser verificados pela similaridade da sequência do gene *16S* rRNA. Tais desvios fortes podem, em princípio, ser causados por contaminação por sequência, além da variação no GC% ser >1%.

O isolado WD37 não tem nenhuma observação adicional, mas como mencionado anteriormente o conteúdo GC% varia um pouco quando comparado a média apresentada por *C. diphtheriae*; e para WD45 também é sugerido a comparação com o *16S* rRNA. Já FRC0190 apresenta como conclusão potencial nova espécie, e como observação a informação que o banco é atualizado quase diariamente, recomendando também que pode ser feita a solicitação de uma análise extensa do gene *16S* rRNA.

6 Conclusões

- O classificador que integra diferentes resultados de OGRIs mostrou ser eficiente para a circunscrição de espécies quando comparada as análises filogenéticas clássicas assim como às análises que utilizam somente ANI;
- Necessidade de novas métricas, em razão da performance sub-ótima do ANIb sozinho;
- A estratégia da assinatura digital concatenando ANI e TETRA mostrou sensibilidade a algumas variações genômicas como conteúdo GC% e tamanho do genoma;
- A distância *'euclidiana'* = 2 e adição do grupo externo foram os parâmetros que permitiram a visualização com melhor resolução nos dendrogramas que foi corroborada pela MLSA;
- Esta estratégia pode ser uma ferramenta complementar para a classificação de outros patógenos bacterianos clinicamente importantes.

7 Perspectiva

Prevemos que esta nova estratégia pode ser uma ferramenta complementar que pode ser extrapolada para identificação de outros patógenos bacterianos clinicamente importantes.

Referências Bibliográficas

ALIBI, S. *et* al. Identification of clinically relevant *Corynebacterium* strains by Api Coryne, MALDI-TOF-mass spectrometry and molecular approaches. **Pathol Biol (Paris),** v. 63, p. 153-157, Ago, 2015. http://dx.doi.org/10.1016/j.patbio.2015.07.007

ANVISA. Diretriz Nacional para Elaboração de Gerenciamento do Uso de Antimicrobianos em Serviços de Saúde. 2017. Disponível em: <http://portal.anvisa.gov.br/documents/33852/271855/Diretriz+Nacional+para+Elaboração+de+Programa+de+Gerenciamento+do+Uso+de+Antimicrobianos+em+Serviços+de+Saúde/667979c2-7edc-411b-a7e0-49a6448880d4>

BADELL, E. *et al. Corynebacterium rouxii* sp. nov., a novel member of the diphtheriae species complex. **Research in Microbiology,** v; 171, p. 122-127, jun, 2020. https://doi.org/10.1016/j.resmic.2020.02.003

BERNARD, K. The genus Corynebacterium and other medically relevant coryneform-like bacteria. **Journal of Clinical Microbiology**, v. 50, p. 3152–3158, 2012. doi: https://doi.org/10.1128/JCM.00796-12

CDC, 2015. **National Action Plan for Combating Antibiotic-resistant Bacteria**. Washington, DC. March 2015. Disponível em < https://www.cdc.gov/drugresistance/us-activities/national-action-plan.html#:~:text=The%20National%20Action%20Plan%20provides,of%20evidence%2Dbased%20stewardship%20strategies> Acesso em 22 de agosto de 2020.

CDC, 2020. **About Antibiotic Resistance**. Disponível em <https://www.cdc.gov/drugresistance/about.html>

CHUN, J. *et al*. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. **Int J Syst Evol Microbiol**, v. 68, p. 461–466, jan, 2018. doi 10.1099/ijsem.0.002516

CIUFO, S. *et al*. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. **Int J Syst Evol Microbiol**, v. 68, p. 2386-2392, jul, 2018. doi: 10.1099/ijsem.0.002809.

DAVIS J. et al. **The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities**. Nucleic Acids Res. 2020 Jan 8;48(D1):D606-D612. doi: 10.1093/nar/gkz943. PMID: 31667520. PMCID: PMC7145515.

DAZAS, M. *et al*. Taxonomic status of Corynebacterium diphtheriae biovar Belfanti and proposal of *Corynebacterium belfantii* sp. nov. **Int J Syst Evol Microbiol**, v. 68, p. 3826-3831, Out, 2018. doi: 10.1099/ijsem.0.003069

FORBES, B. Did I Hear You Correctly? The Organism Identified Was Corynebacterium diphtheriae. **Clinical Microbiology Newsletter, v.** 39, n. 5, p. 35-41 mar, 2017. https://doi.org/10.1016/j.clinmicnews.2017.02.001

GLAESER, S.; KÄMPFER, P. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. **Systematic and Applied Microbiology**, v. 38, p. 237–245, jun, 2015. doi: http://dx.doi.org/10.1016/j.syapm.2015.03.007

GOUY, M., GUINDON, S., GASCUEL, O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building, **Molecular Biology and Evolution**, v. 27, p. 221–224, Ed. 2, Fev, 2010, https://doi.org/10.1093/molbev/msp259

KÄMPFER, P.; GLAESER, S. Prokaryotic taxonomy in the sequencing era – the polyphasic approach revisited. **Environmental microbiology.** v. 14, n. 2, p. 291-317, fev, 2011; doi: 10.1111/j.1462-2920.2011.02615.x.

KASSAMBARA, A. MUNDT, F. "Factoextra: extract and visualize the results of multivariate data analyses." R package version 1(3), 2016

KHAMIS, A., RAOULT, D., LA SCOLA, B. *rpoB* gene sequencing for identification of *Corynebacterium* species. **J. Clin. Microbiol**, v. 42, p. 3925–3931. Set, 2004. doi: 10.1128/JCM.42.9.3925-3931.2004

KIM, M. *et al*. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. **International Journal of Systematic and Evolutionary Microbiology.** v. 64, p. 346-351, fev, 2014. https://doi.org/10.1099/ijs.0.059774-0

KÖSER, C. *et al*. Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. **PLoS Pathogens**, v. 8, Ago, 2012. https://doi.org/10.1371/journal.ppat.1002824

KUMAR S. *et al*. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. **Molecular Biology and Evolution** 35:1547-1549.

LEE, I. *et al*. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. **Int J Syst Evol Microbiol**, v. 66, p. 1100–1103, fev, 2016. doi: 10.1099/ijsem.0.000760

LEFORT, V., LONGUEVILLE, J., GASCUEL, O. SMS: Smart Model Selection in PhyML. **Molecular Biology and Evolution**, v. 34, p. 2422–2424, Set, 2017, https://doi.org/10.1093/molbev/msx149

MCFEE, R. EMERGING INFECTIOUS DISEASES – OVERVIEW. **Dis Mon**. v. 64, n. 5, p. 163-169, maio, 2018. doi: 10.1016 / j.disamonth.2018.01.002

MORENS, D.; FOLKERS, G.; FAUCI, A. Emerging infections: a perpetual challenge. **Lancet Infectious diseases**, v. 8, n. 11, p. 710–719, nov, 2008 https://doi.org/10.1016/S1473-3099(08)70256-1

MUKHERJEE, D. *et al*. Genomes OnLine database (GOLD) v.7: updates and new features. **Nucleic Acids Research,** v. 47, p D649–D659, Jan, 2019. https://doi.org/10.1093/nar/gky977

MUKHERJEE, S. *et al*. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. **Nature biotechnology.** v. 3, p. 676- 684, Jun, 2017. DOI: https://doi.org/10.1038/nbt.3886

NA, S. *et al*. UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. **J. Microbiol**, v. 56, p. 280-285, Jan, 2018. https://doi.org/10.1007/s12275-018-8014-6

NII-TREBI, N. Emerging and Neglected Infectious Diseases: Insights, Advances, and Challenges. **Biomed Res Int**, fev, 2017. doi: 10.1155/2017/5245021

OLIVEIRA, A. *et al.* Insight of Genus *Corynebacterium*: Ascertaining the Role of Pathogenic and Non-pathogenic Species. **Front Microbiol.** v. 8, Out, 2017. doi: 10.3389/fmicb.2017.01937

PALMER, M. *et al.* All ANIs are not created equal: implications for prokaryotic species boundaries and integration of ANIs into polyphasic taxonomy. **International Journal of Systematic and Evolutionary Microbiology,** v. 70, Abr, 2020. DOI: https://doi.org/10.1099/ijsem.0.004124

PARKS, *et al*. Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. **bioRxiv**, Nov, 2019. doi: https://doi.org/10.1101/771964

PIVOT, D. *et al.* Carriage of a Single Strain of Nontoxigenic *Corynebacterium diphtheriae* bv. Belfanti (*Corynebacterium belfantii*) in Four Patients with Cystic Fibrosis. **J Clin Microbiol**. V. 57, Maio, 2019. doi: 10.1128/JCM.00042-19

PRIDE, D. *et al*. Evolutionary implications of microbial genome tetranucleotide frequency biases. **Genome Res**, v. 13, p. 145-158, jan, 2003. doi: 10.1101/gr.335003

RICHTER, M. *et al*. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. **Bioinformatics**, v. 32, p. 929-31, Nov, 2016. doi: 10.1093/bioinformatics/btv68

RICHTER, M.; ROSSELLÓ-MÓRA, R. Shifting the genomic gold standard for the prokaryotic species definition. **Proc Natl Acad Sci USA,** v. 10, p. 19126-31, out, 2009. doi: 10.1073/pnas.0906412106

ROSSELLÓ-MÓRA, R.; AMANN, R. Past and future species definitions for Bacteria and Archaea. **Systematic and Applied Microbiology**, v. 38, p. 209–216, 2015. doi: http://dx.doi.org/10.1016/j.syapm.2015.02.001

SAITOU N., NEI M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, (1987), 4:406-425.

SANGAL, V. A lack of genetic basis for biovar differentiation in clinically important Corynebacterium diphtheriae from whole genome sequencing. **Infection, Genetics and Evolution**, v. 21, p.54-57, Jan, 2013. doi: http://dx.doi.org/10.1016/j.meegid.2013.10.019

SANTOS, A. *et al*. Searching whole genome sequences for biochemical identification features of emerging and reemerging pathogenic Corynebacterium species. **Functional and Integrative Genomics**, v. 18, p. 593-610, Set, 2018. https://doi.org/10.1007/s10142-018-0610-3

SANTOS, C. et al. Efficient differentiation of Corynebacterium striatum, *Corynebacterium amycolatum* and *Corynebacterium xerosis* clinical isolates by multiplex PCR using novel species-specific primers. **Journal of Microbiological Methods**, v. 142, p. 33–35, Ago, 2016 doi: https://doi.org/10.1016/j.mimet.2017.09.002

SHARMA, N. *et al*. Al. Diphtheria. **Nature Reviews Disease Primers**. v. 5, n.81, dez. 2019. https://doi.org/10.1038/s41572-019-0131-y

TAGINI, F. *et al*. Distinct Genomic Features Characterize Two Clades of *Corynebacterium diphtheriae*: Proposal of *Corynebacterium diphtheriae* Subsp. *diphtheriae* Subsp. nov. and *Corynebacterium diphtheriae* Subsp. *lausannense* Subsp. nov. **Front. Microbiol**, v. 17, Ago, 2018. DOI: https://doi.org/10.3389/fmicb.2018.01743

VRIES, A., RIPLEY, B. "**Ggdendro: tools for extracting dendrogram and tree diagram plot data for use with ggplot**." R package version 0.1-12, 2013. Recuperado de http://CRAN. R-project. org/package= ggdendro.

WICKHAM, H. **ggplot2: elegant graphics for data analysis**, Springer, 2016.

YANAI, M. *et al*. A. Retrospective evaluation of the clinical characteristics associated with *Corynebacterium* species bacteremia. **Brazilian Journal of Infectious Diseases**, v. 22, n. 1, p. 16–23, fev, 2018. https://doi.org/10.1016/j.bjid.2017.12.002

YOON, S. et al. A large-scale evaluation of algorithms to calculate average nucleotide identity. **Antonie Van Leeuwenhoek**. v. 110, n. 10, p. 1281-1286. out, 2017. doi: 10.1007/s10482-017-0844-4. Epub 2017 Feb 15. PMID: 28204908.

ZUMLA, A.; HUI, D. Emerging and Re-Emerging Infectious Diseases. **Infectious Disease Clinnics**, v. 36, n.4, p. 869-1158, dez, 2019. https://doi.org/10.1016/j.idc.2019.09.001

Apêndices

A

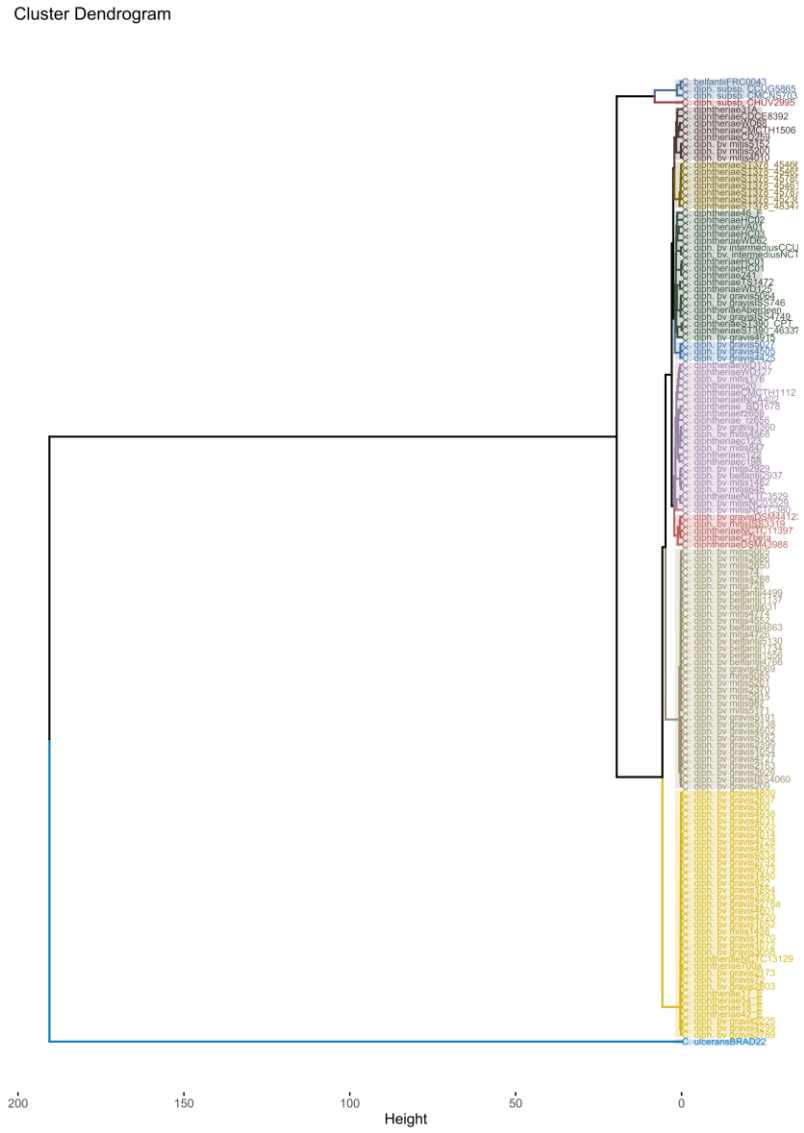


Figura 14 Análises com 140 genomas, incluindo *C. ulcerans* **como** grupo externo distância 'euclidiana' =2. Neste dendrograma observa-se a clusterização com outra espécie, e consistente demarcação dos isolados da subespécie *lausannense* dos isolados de *C. diphtheriae*, bem como a separação clara de CHUV 2995 de CCUG5865 e CMCNS703, propostas como subespécies.

B

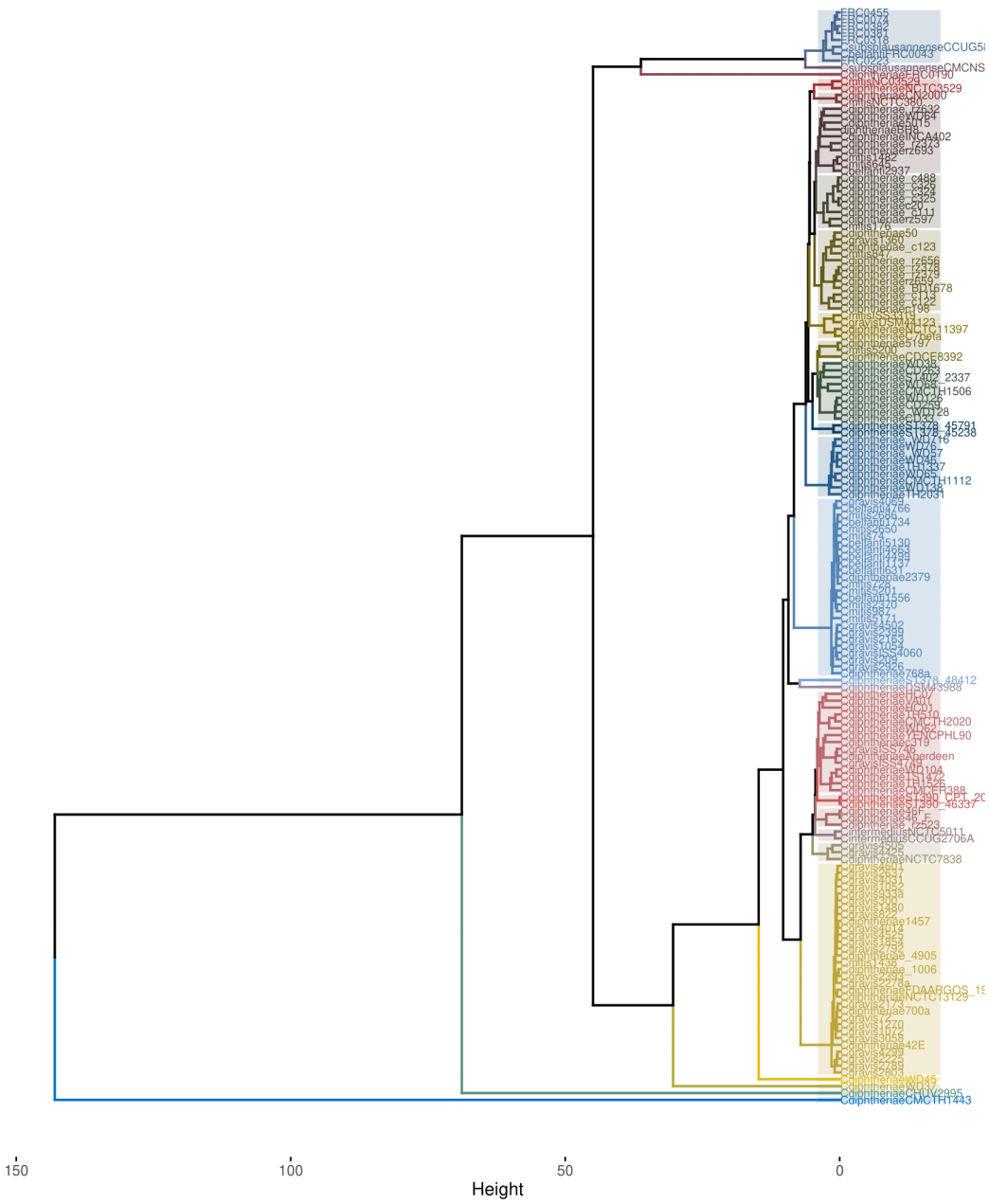


Figura 15 Dendrograma com 159 genomas e distância *'euclidiana'* = 2 sem grupo externo. Neste dendrogramas podemos observar que o isolado WD45 de *C. diphtheriae* formou um clado demarcados das demais diftéricas. Este isolado apresenta diferenças no tamanho do genoma e no conteúdo GC% (2583616bp e 52.98, respectivamente) em comparação aos demais representantes da espécie *C. diphtheriae*.

C

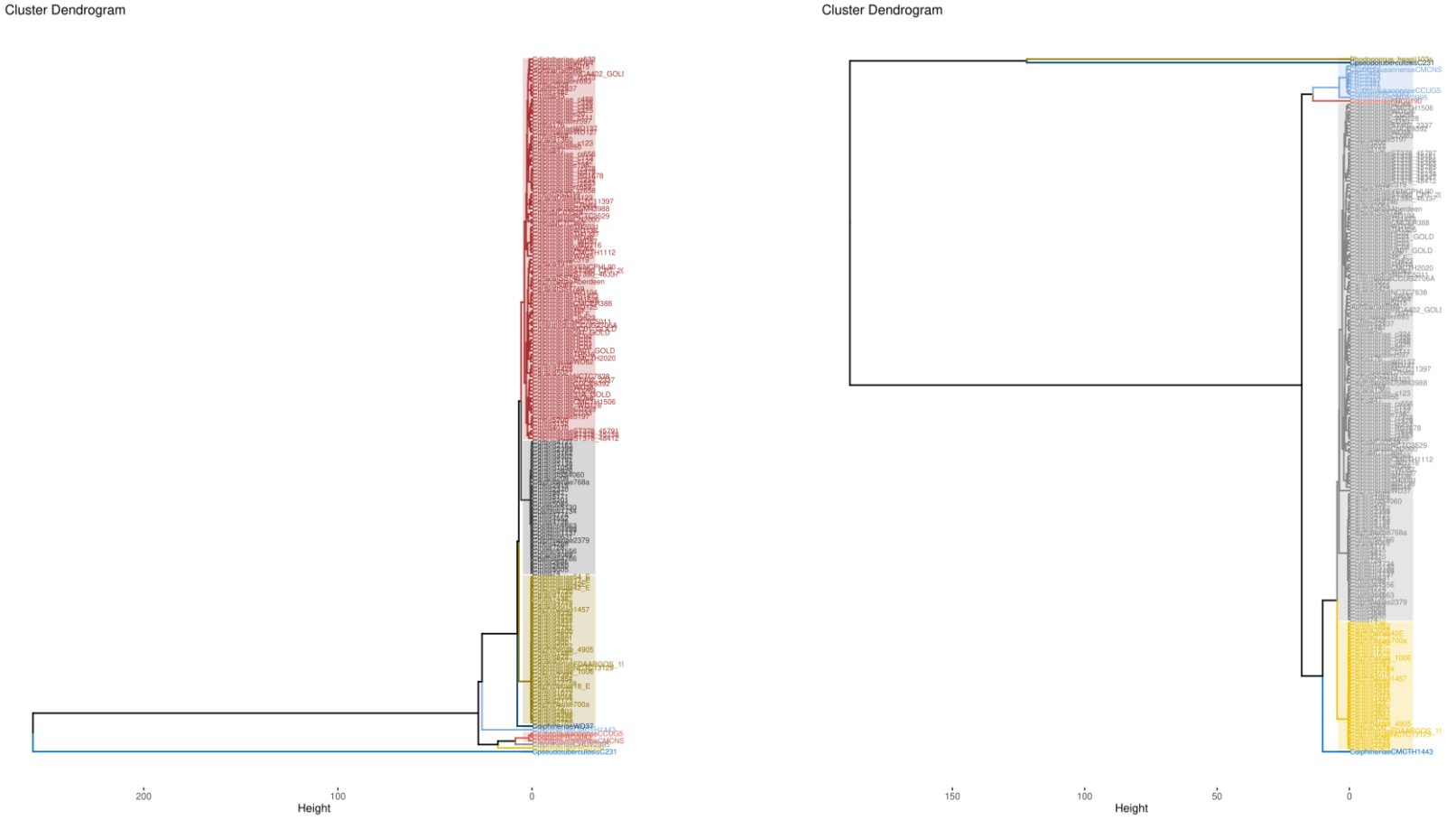


Figura 16 Dendrograma com 192 genomas (na esquerda) e 200 genomas (à direita). Nesta análise foram alteradas a quantidade dos genomas e as espécies como grupo externo. Aqui observa-se que embora sejam feitas modificações dos grupos externos, os isolados que apresentaram demarcação ao longo das nossas modificações se mantêm demarcados de *C. diphtheriae*, inclusive CMCTH1443 e CHUV também não se agrupa às demais *lausannense*.