UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Curso de Graduação em Biotecnologia

Trabalho de Conclusão de Curso

Análise pangenômica de *Curtobacterium* sp.

Christian Domingues Sanchez

Pelotas, 2019

Christian Domingues Sanchez

Análise pangenômica de *Curtobacterium* sp.

Trabalho de Conclusão de Curso apresentado ao Curso de Biotecnologia, do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Bacharel em Biotecnologia.

Orientador: Prof. Dr. Luciano da Silva Pinto
Co-Orientadora: Bela. Amanda Munari Guimarães

Pelotas, 2019

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

S211a Sanchez, Christian Domingues

Análise pangenômica de *Curtobacterium* sp. / Christian Domingues Sanchez ; Luciano da Silva Pinto, orientador ; Amanda Munari Guimarães, coorientadora. — Pelotas, 2019.

51 f. : il.

Trabalho de Conclusão de Curso (Bacharelado em Biotecnologia) — Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2019.

1. Genômica comparativa. 2. Genoma núcleo. 3. Metabolismo secundário. 4. Bactéria cosmopolita. I. Pinto, Luciano da Silva, orient. II. Guimarães, Amanda Munari, coorient. III. Título.

CDD : 575.21

Elaborada por Ubirajara Buddin Cruz CRB: 10/901

Christian Domingues Sanchez

Análise pangenômica de *Curtobacterium* sp.

Trabalho de Conclusão de Curso aprovado, como requisito parcial, para obtenção do grau de Bacharel em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 26 de novembro de 2019

Banca Examinadora:

Prof. Dr. Luciano da Silva Pinto (Orientador)
Doutor em Biotecnologia pela Universidade Federal de Pelotas
.....

Profª. Drª. Vanessa Galli
Doutora em Biologia Celular e Molecular pela Universidade Federal do Rio Grande do Sul
.....

Bela. Amanda Munari Guimarães
Biotecnologista pela Universidade Federal de Pelotas
.....

Dedico esse trabalho a minha família que sempre me apoiou em tudo e não mediu esforços para alcançar o nosso sonho

Agradecimentos

Primeiramente agradeço a minha família, ao meu pai, José Luiz e minha mãe Nelci, pelo apoio incondicional e pelos sacrifícios feitos por eles para podermos compartilhar desse momento. Vocês dois são o meu maior exemplo de pessoas, amáveis e apaixonados, então dedico esta conquista para vocês que são minha inspiração de vida e meus amores eternos. Amo vocês.

Agradeço a minha irmã, que sempre me puxou a orelha e é como uma segunda mãe, como ela mesmo diz, irmã mais velha é mãe e irmã, minha amiga e sempre incansável, essa conquista também é tua, obrigado pela motivação, dedico ao meu irmão Murilo, que ainda é pequeno, mas sempre apoiou o “mano”.

Agradeço a Caroline, que foi minha parceira e amiga durante esses anos e que me deu muito apoio e carinho, obrigado por tudo!

Agradeço ao meu professor e orientador Luciano, que abriu as portas desde o início do curso para o Laboratório de Bioinformática e Proteômica, no qual tenho orgulho de dizer que fiz parte! Sou muito grato pelas oportunidades ofertadas e por toda contribuição que ele teve na minha vida pessoal e acadêmica, pelos conselhos e puxões de orelha, muito obrigado professor.

Agradeço aos meus tutores nessa jornada de bioinformática Amanda Munari, minha co-orientadora e amiga, obrigado pelos conselhos e pelo carinho, menção honrosa ao Frederico que durante esses 4 anos me ensinou muito e que sou muito grato, ao Rafael por pelos ensinamentos e dedicação.

Agradeço a todos os meus amigos que fizeram parte desta jornada, colegas de laboratório e colegas de curso, todos vocês sabem o quanto fizeram parte dessa conquista!

Agradeço a todos os professores da Graduação em Biotecnologia pelos ensinamentos e pelo seu empenho, em especial a Professora Luciana Dode, que me ensinou a paixão pela extensão! A professora Vanessa, que além de ser o nome da minha turma, para mim é um exemplo de profissional e paixão pela pesquisa.

Agradeço a Universidade Federal de Pelotas e a todo curso de Biotecnologia, pelo excelente curso, também gostaria de agradecer aos funcionários, em especial a secretária da graduação Renata, que sempre é incansável com os alunos e muito dedicada, obrigado!

[...]Tenha felicidade bastante para fazê-la doce.
Dificuldades para fazê-la forte.
Tristeza para fazê-la humana.
E esperança suficiente para fazê-la feliz.[...]

Clarice Lispector

Resumo

Sanchez, Christian Domingues. **Análise Pangenômica de *Curtobacterium* sp**. 2019. 51f. Trabalho de Conclusão de Curso (Bacharel em Biotecnologia) – Curso de Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2019.

Curtobacterium sp. é um gênero de bactéria pertencente à ordem das Actinomycetales, são Gram-positivas e comumente encontradas no solo e em plantas. Esse taxon é considerado cosmopolita terrestre. O gênero possui aproximadamente 10 espécies e sua importância econômica se dá por causar a doença da Murcha de *Curtobacterium* em feijões (*Phaseolus* .spp*)*. Neste trabalho, a partir de 6 genomas de *Curtobacterium sp.* fez-se análises do pangenoma a fim de prospectar genes de possível interesse biotecnológico. Através da análise de pangenômica estrutural foi possível determinar o pangenoma, sendo esse do tipo fechado. Estes genomas foram anotados e a busca por grupos ortólogos gerou 2119 grupos pertencentes ao genoma núcleo, além de determinar o tamanho de 1.2 Mb para as regiões genômicas conservadas entre as cepas. Na análise metabólica foram encontradas rotas de processos celulares, de respostas à alteração do ambiente e de processamento de informações genéticas. Por meio da análise de vias metabólicas secundárias identificou-se possíveis mecanismos que nos ajudam a entender mais o mecanismo de patogenicidade e de adaptação a estresses.

Palavras-chave: Genômica comparativa; Genoma Núcleo; Metabolismo secundário; Bactéria cosmopolita.

Abstract

Sanchez, Christian Domingues. **Pangenomics analysis of *Curtobacterium* sp.** 2019. 51f. Trabalho de Conclusão de Curso (Bacharel em Biotecnologia) – Curso de Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2019.

Curtobacterium sp. is a genus of gram positive bacteria, belonging to the order Actinomycetales, commonly found in soil and plants. This taxon is considered a terrestrial cosmopolitan. This genus has approximately 10 families and is very economic relevant since is the causative agent of the bacterial wilt disease in beans (*Phaseolus* spp.). This aim of this study was to perform a pangenomic analysis with 6 genomes, for the purpose of prospecting possible genes of biotechnological interest. Through structural pangenomic analysis it was possible to determine the type of pangenome, which is closed. After annotation of the genomes, a search for ortholog groups generated 2119 groups belonging to the core genome and conserved genomic regions were 1.2 Mb between strains. In the metabolic analysis we found routes of cellular processes, responses to environmental change and processing of genetic information. Through the analysis of secondary metabolic pathways, we identified possible mechanisms that might help us understand the mechanism of pathogenicity and adaptation to stress.

Keywords: Comparative genomics; Core Genome; Secondary metabolism; Cosmopolitan bacterium.

Lista de Figuras

SUMÁRIO

1. Introdução

As bactérias do gênero *Curtobacterium* tem sua forma caracterizada por bastonetes retos, que são ligeiramente curvos ou em forma de cunho e curtos. Pertencendo ao filo Actonobacteria e a família Microbacteriacea, são aeróbicos obrigatórios, Gram-positivas, cujo tamanho varia de 0,3 a 0,6μm por 1 a 3μm. São bactérias móveis por um ou mais flagelos polares ou subpolares e não formam endósporo. Esse gênero foi primeiramente caracterizado como um gênero diferente em 1922, como bactéria causadora da Murcha da *Curtobacterium* e teve como sua primeira denominação de *Bacterium flaccumfaciens*, tendo ainda recebido outras denominações tais como: *Phytomonas flaccumfaciens; Pseudomonas flaccumfaciens; Corynebacterium flaccumfaciens; Corynebacterium flaccumfaciens pv. flaccumfaciens; Corynebacterium flaccumfaciens subsp. flaccumfaciens*.

O gênero pode ser fitopatogênico e esse táxon é considerado cosmopolita terrestre, tendo isolados principalmente do solo e de plantas. Como principal destaque por causar a Murcha de *Curtobacterium*, onde no Brasil há uma perda de 15% no rendimento de feijões devido à doença, também já foi descrita como infectando humanos, possivelmente de forma oportunista. Outra característica que podemos destacar é que está relacionada com a degradação de compostos orgânicos.

O advento da genômica moderna proporcionou com o uso de ferramentas de bioinformática, a identificação de várias características de interesse biotecnológico, analisando sua fisiologia, genômica funcional, bioquímica e também da patogênese de diversas doenças. Nesse sentido, o sequenciamento de nova geração (NGS, *Next Generation Sequencing*) contribuiu para o grande crescimento nos bancos públicos de genomas, como o NCBI GenBank (*National Center for Biotechnology Information, genetic sequence database*).

Devido a sua importância econômica para a cultura do feijão, já foram depositados até o momento 16 genomas do gênero *Curtobacterium* no total, onde 3 são completos e 13 genomas são rascunho, essa disponibilidade de bases de dados nos possibilita fazer o uso de análises de bioinformática para adquirir informações que possam ser relevantes pra busca de proteínas e genes que tenham potencial biotecnológico, como estudos comparativos que possibilitam compreender a plasticidade desses genomas, como a análise pangenômica.

A análise pangenômica busca compreender e também comparar diferentes isolados de um mesmo gênero ou espécie, tendo como objetivos identificar os aspectos evolutivos, interações entre patógeno-hospedeiro e aspectos adaptativos, propiciando ao Biotecnologista o acesso a informações detalhadas destes organismos e assim, a mineração de dados biológicos que levem à construção de estratégias para a erradicação da doença ou mesmo a identificação de novos alvos biotecnológicos para a geração de produtos de importância econômica. Nessa lógica, o presente trabalho, por meio de abordagens *in silico*, objetivou analisar o pangenoma e o genoma núcleo do gênero *Curtobacterium* com a finalidade de identificar genes e rotas metabólicas que possam caracterizar um melhor entendimento do gênero e possíveis produtos de aplicação biotecnológica.

2. Revisão de Literatura

2.1 Gênero Curtobacterium

Curtobacterium é um novo gênero que surgiu após sua diferenciação das *Corynebacterium*, através do perfil proteico via eletroforese de gel de acrilamida (NEILANDS, 1995) . E esse gênero está distribuído em 10 espécies, *Curtobacterium albidum; Curtobacterium ammoniigenes; Curtobacterium citreum; Curtobacterium flaccumfaciens; Curtobacterium ginsengisoli; Curtobacterium herbarum; Curtobacterium luteum; Curtobacterium plantarum.* A espécie *Curtobacterium flaccumfaciens* é de importância econômica, uma vez que causa a Murcha de *Curtobacterium* em feijoeiros, sendo então uma das únicas espécies de *Curtobacterium* associada com patogênese de plantas (COLLINS; JONES; SCHOFIELD, 1982), incluindo uma cepa capaz de formar biofilme em alfaces (DEES; BRURBERG; LYSØE, 2016).

Contudo, há outras espécies que já foram registradas realizando outros papéis ecológicos como simbiontes endofíticos (DE BOER; SADDLER; STEAD, 2003), reduzindo sintomas de doenças (LACAVA et al., 2007) . O gênero também pode infectar humanos de forma oportunista (BULGARI et al., 2011), pode ser também encontrada no solo (FUNKE; ARAVENA-ROMAN; FRODL, 2005) mesmo não esporulando (VIDAVER, 1982) . Como observado, essa espécie pode ser encontra

em diversos ambientes, sendo que o gênero já foi descrito nos 5 continentes (Figura 1).

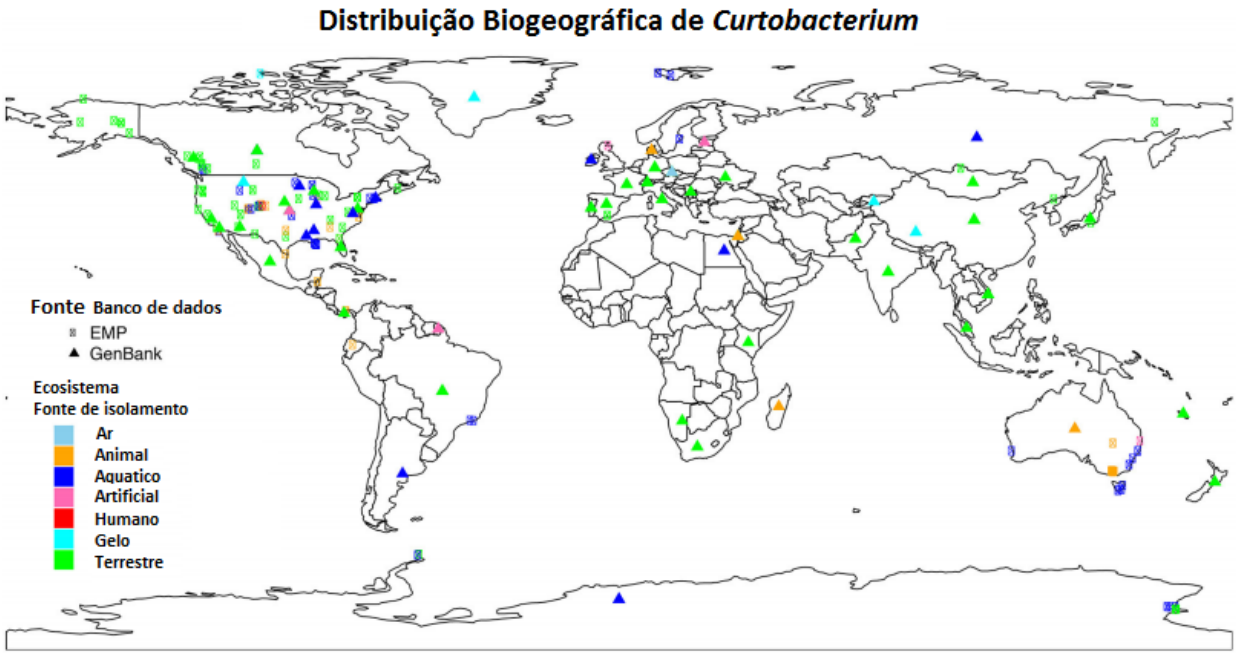

Figura 1. Distribuição geográfica de *Curtobacterium* compilada a partir de várias fontes de isolamento. As cores indicam os diferentes ecossistemas a partir dos quais a sequência foi isolada. A forma do símbolo indica o conjunto de dados do qual a sequência se originou. As sequências obtidas do GenBank (símbolos triangulares) foram principalmente aproximações, pois as coordenadas GPS detalhadas não estavam disponíveis. Adaptado de CHASE et al.,(2016[a]).

Em relação aos dados genômicos, muitas bactérias do gênero *Curtobacterium* ainda estão caracterizadas sem espécie no NCBI GenBank (BENSON et al., 2007). Sendo esse, um ponto importante, que é a falta de informações gerais, como fonte do isolado, local e informações básicas sobre os isolados, essa falta de informações acaba prejudicando a caracterização dos organismos. É possível encontrar algumas características dos genomas em artigos que foram publicados, como anúncios de genomas, dentre eles temos o isolado formador de biofilme em alfaces, que está identificada apenas pelo gênero, incluindo seu genoma (DEES; BRURBERG; LYSOE, 2016). Devido ao gênero apresentar grande plasticidade adaptativa, seu estudo importante devido à possibilidade de encontrarmos genes de interesse biotecnológico.

2.2 Curtobacterium flaccumfaciens pv flaccumfaciens

Apesar de o gênero ter bastante espécies catalogadas, muitas delas não são tão bem descritas quanto a espécie *C. flaccumfaciens* pv *flaccumfaciens*. A espécie causadora da Murcha de *Curtobacterium,* foi primeiramente relatada nos Estados Unidos da América, no estado americano de Dakota do Sul (HEDGES, 1922) COLLINS; JONES, 1984) . Essa espécie, também chamada de *Cff* na literatura, é de grande importância econômica uma vez que é responsável pela perda de produtividade de 15% nas plantações de feijões do Brasil (MANTEM, 2008).

A caracterização da doença se dá pela murcha e/ou flacidez de folíolos, que durante o dia ficam mais flácidos e a noite são mais túrgidos. Podendo ser assintomática. Essas bactérias colonizam os feixes vasculares das plantas causando então um problema vascular, sendo então uma síndrome de transporte, essa síndrome então causará manchas amarelas, queima das folhas e necrose (Figura 2), que são características comuns de crescimento bacteriano (SHERF; MACNAB, 1986). Apesar desses efeitos negativos, existem registros mostrando que tem efeitos positivos de ter a bactéria no solo, como o exemplo de um trabalho que cita a presença da *Cff*, que induziu resistência sistêmica em plantas de pepino, quando presentes na rizosfera (RAUPACH; KLOEPPER, 1998).

Figura 2. Folhas de feiojoeiro afetadas pela murcha. Necrose vascular com margens cloróticas em folhas de feijoeiro (cv. *Talash*) causada por *Cff.* Obtido de OSDAGHI et al., (2015).

A infecção por *Cff* já foi descrita na soja no Brasil (OSDAGHI et al., 2015), que é o segundo maior produtor de soja do mundo segundo (“Soja - Portal Embrapa”, 2019).

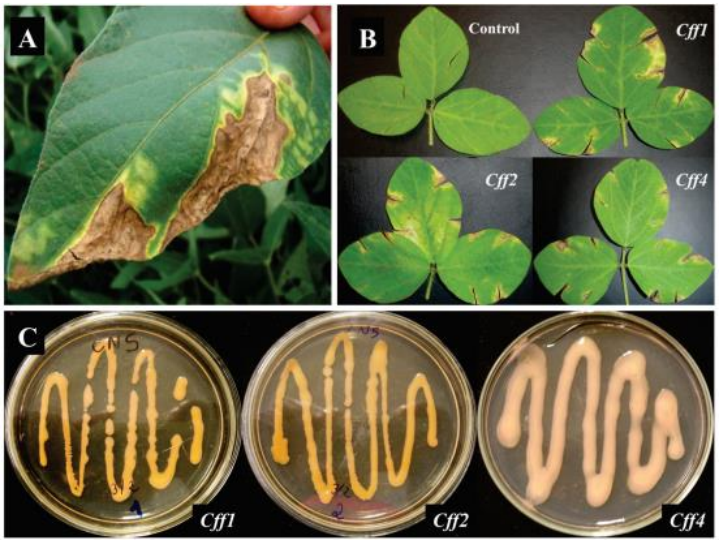

Figura 3. Folhas de soja infectadas com *Curtobacterium* e placas com meio seletivo. (A.) Tecido foliar clorótico e seco causado por *Curtobacterium flaccumfaciens* pv*. flaccumfaciens* (*Cff*). (B.) Sintomas de inoculação em tesoura com três isolados de *Cff.* (C.) colônias de três isolados de *Cff* no meio de cultura *Clavibacter nebraskensis* Seletivo médio (CNS), obtido de SOARES et al., (2013).

2.3 A importância da análise Genômica

O genoma é o ácido desoxirribonucleico (DNA) total presente em um organismo, com todas as informações que nele carrega, que é transmitido de geração em geração. O genoma é composto por genes, o gene é uma sequência específica do DNA que contém as instruções necessárias para a síntese de uma proteína ou molécula de RNA; sequências que ajudam na regulação da expressão; sequências não-codificadoras; ou seja, é toda a informação genética armazenada em um único organismo.

A genômica busca estudar as informações contidas nos genomas dos organismos, que são de grande importância, visto que possibilita caracterizar fatores muito importantes, como rotas metabólicas, processos celulares e genes relacionados a doenças (STAMATOYANNOPOULOS, 2004). Essa área da ciência está em crescente expansão e tem como origem da área o projeto de sequenciamento do genoma humano, que foi iniciado no final da década de 1980, mas se considera que a partir do descobrimento da dupla hélice do DNA até o desenvolvimento das técnicas presentes nos dias de hoje, como um dos campos da ciência moderna (STAMATOYANNOPOULOS, 2004).

A genômica é responsável pela caracterização dos genomas de organismos, tais como milhares de genes e proteínas que estão presentes no mesmo. A genômica então se divide em diversas subáreas como a genômica funcional, estrutural, metagenômica e epigenômica. Umas de mais destaque é a genômica funcional, que tem como principal foco compreender as interações dos genes com suas proteínas e suas funções. O desenvolvimento de ferramentas que permitem a predição e determinação de estruturas tridimensionais ajudam a genômica estrutural a se formar com a predição de organização e estrutura dos genes. A metagenômica é responsável pela verificação da diversidade de microrganismos presentes numa determinada amostra e a análise de organismos em ambientes específicos. A epigenômica busca compreender as modificações epigênéticas no material genético, como os silenciamentos de alguns genes. (LOCKHART; WINZELER, 2000).

2.4 Genômica de microrganismos procariotos

Os microrganismos procariotos são os mais adaptados, antigos e numerosos do planeta, são capazes de colonizar qualquer nicho biológico existente no planeta. Sua existência é essencial para manter o equilíbrio ecológico, além de terem grande importância como recurso genético para a biotecnologia e para o desenvolvimento de bioeconomia do século 21 (FIELDS, 2007). A utilização desses recursos permite o desenvolvimento de novos fármacos, inúmeras aplicações na área da saúde, indústria, meio ambiente e agricultura (VAN DIJK et al., 2014). A bioinformática em conjunto com a biologia molecular, permitem a prospecção *in silico,* a partir de dados genômicos, obtendo então produtos biotecnológicos como enzimas, polímeros, antibióticos, agentes terapêuticos, melhorando também a prevenção e prognostico de doenças emergentes (MCINERNEY et al., 2008).

A caracterização microrganismos procariotos é de grande relevância, visto que a obtenção de seu material genético comparado com um eucarioto, principalmente o tamanho do genoma e sua estabilidade genética, ajudam no entendimento de sua genética e fisiologia, também como eventos evolutivos, sua bioquímica, no entendimento da sua patogênese e também no desenvolvimento de compostos biotecnológicos. (MCINERNEY et al., 2008)

Esta área conta com uma grande base de dados, um dos melhores exemplos é o Genome do NCBI (www.ncbi.nlm.nih.gov), no qual constam aproximadamente

15.800 entradas de Procariotos. Já no Genomes Online Database (GOLD) (www.genomesonline.org), atualmente, estão registrados mais de 68 mil projetos de genomas de procariotos. Com essa grande quantidade de dados sendo gerada e disponibilizada, possibilitando então o surgimento de novas linhas de pesquisa, como análise comparativa dos genomas bacterianos, buscando identificar mecanismos adaptativos, aspectos evolutivos e possíveis produtos biotecnológicos.

2.5 Sequenciamento de nova geração

Dentre os sequenciadores considerados de nova geração, podemos destacar o Illumina, sendo ele o atual padrão ouro para muitas análises em NGS, sua base se dá por sequênciamento de síntese (SBS), onde existem três linhas de equipamentos que geram diferentes tamanhos de leitura em pares de base (BP), MiSeq (250x2), HiSeq (120x2) e NextSeq, esse sequenciador tem uma grande variedade de protocolos de sequenciamento e análise de dados já bem estabelecidas, o que acaba então barateando o custo (ILLUMINA, 2017).

Esses novos métodos e abordagens para sequenciamento geram uma enorme quantidade de dados por ano, visto que o custo médio para se sequenciar reduziu e grandes centros já conseguem gerar um nível superior de dados com seus próprios sequenciadores .

2.6 Bioinformática

O advento dos sequenciadores de nova geração trouxe uma nova perspectiva, pelo aumento expressivo de dados biológicos e também pela necessidade de processar esses dados, como as análises genômicas. Com o aumento do fluxo de dados biológicos houve a necessidade de armazenar esses dados de forma digital e tratar os mesmos, e assim surgiu os primeiros bancos de dados, utilizando a abordagem de informática.

Assim sendo, a bioinformática é responsável pela interpretação dos dados genômicos, com a criação de bancos de dados especializados e ferramentas que vão ajudar o processo de análise, aperfeiçoando cada vez mais as mesmas. Assim como os profissionais em bioinformática, os bioinformatas, que exercem papel fundamental

na evolução da pesquisa, pois as análises de dados geradas serão o provável próximo gargalo na pesquisa biológica (MOROZOVA; MARRA, 2008).

A bioinformática ganhou força com a chegada do NGS, diversas técnicas tiveram que ser desenvolvidas para trabalhar na complexidade e interpretadores para os dados computacionais, alguns desses fundamentos básicos de conceito e técnicas para a bioinformática que são utilizados até os dias de hoje (HANDELSMAN, 2004).

Apesar dos pioneiros dessa área de biologia computacional não usassem o termo “Bioinformática”, tinham a visão que a computação, a matemática e a biologia molecular podiam ser combinadas de maneira poderosa para responder questões fundamentais da ciência da vida (HAGEN, 2000).

Com o advento do NGS, que veio paralelamente ao projeto do genoma humano, muitos outros projetos começaram a ser feitos (HOWARD, 2000). O que trouxe uma nova abordagem que contribuiu para a redução do tempo de geração de arquivos de sequenciamento, Sequenciamento Shotgun (WGS, *Whole Genome Shotgun Sequencing)* (“Celera’s Approach”, 2000). Isso inovou a forma com que lê e organiza as sequências, comparado com a técnica já estabelecida e considerada padrão ouro da época, o sequenciamento via sistema de Sanger (SANGER; COULSON, 1975).

O método de sanger automatizado gera saídas longas, mas tinha pouca capacidade por sequenciador, logo, se torna um trabalho que exige muito mais tempo para realizar o sequenciamento e o montante de dados por rodada é bem menor (SANGER; COULSON, 1975).

A criação dos bancos públicos foi de grande avanço para essa área, visto que se criou uma forma de guardar todos esses dados em apenas um lugar de forma organizada. O GenBank é um dos maiores bancos de dados genômicos do mundo, mantido pelo NCBI, mas também existem outros bancos de dados bem conhecidos, como o GOLD (BENSON et al., 2007).

Desde então, novos sequenciadores têm sido desenvolvidos e melhorados e seu custo barateado, assim como o tratamento dos dados gerados, o desenvolvimento de técnicas para o processamento dessa enorme quantidade de informação, visto que sequenciar um genoma não é mais complexo como era nos primórdios do sequenciamento, fazendo então a era das ômicas possível (ES, 2011).

2.6.1 Montagem de genomas

Para cada sequenciador existe um padrão de leituras que são lidas e geradas, logo, algoritmos diferentes são necessários para cada tipo de arquivo de saída dos sequenciadores. Existem abordagens para a montagem de genomas, A técnica *De Novo* utiliza algoritmos para montar o genoma do zero, sem nenhum método de comparação e a montagem por referência, utilizará uma sequência já conhecida e disponível para realizar a montagem.

Um dos formatos mais conhecidos é o (.fasta) um formato baseado em texto para representar sequências nucleotídicas ou peptídicas, nas quais pares de bases ou aminoácidos são representados usando códigos de letra única, por meio de parâmetros já estabelecidos para valores. As suas informações são concatenadas e codificadas para facilitar a leitura dos arquivos (METZKER, 2010).

Um exemplo que podemos dar é o sequenciador Illumina MiSeq, que utiliza o formato de biblioteca "*paired-end*". Suas leituras são então processadas por montadores que suportem esse tipo de formato de leitura, montadores clássicos como o Velvet (ZERBINO; BIRNEY, 2008) e o A5 (COIL; JOSPIN; DARLING, 2014).

Essas ferramentas são comumente utilizadas para realizar a montagem desse tipo de dado. Geralmente para uma montagem de boa qualidade, são utilizados diversos montadores, visto que cada um tem seu algoritmo e possivelmente montará diferentes regiões de maneira melhor. Em vista disso, diversas técnicas de refinamento pós montagem, como a utilização de um integrador de montagens como o CISA (LIN; LIAO, 2013), são utilizadas para finalizar o genoma e aumentar sua cobertura.

2.6.2 Anotação e re-anotação

O processo de anotação dos genomas consiste em identificar os genes, regiões funcionais e regiões não codificantes, uma vez que com o genoma montado, temos apenas as sequências de nucleotídeos, mas não sabemos o que elas significam, a anotação nos dá as respostas.

Existem diversas "*pipelines*" para anotação, que são algoritmos e ferramentas automatizadas que reúnem diversas outras ferramentas e bancos de dados, dentre eles, destacam-se o BASys (VAN DOMSELAAR et al., 2005), BG7 (PAREJA-TOBES et al., 2012), PROKKA (SEEMANN, 2014), EuGene-PP (SALLET; GOUZY; SCHIEX,

2014), NCBI Prokaryotic Genome Annotation Pipeline (PGAP), (TATUSOVA et al., 2016), GENIX (KREMER et al., 2016).

Esta é uma das tarefas mais importantes em um processo de caracterização e identificação dos genes no genoma, nos *pipelines* estão incluídos algoritmos poderosos com abordagens pré-definidas para os dados inseridos. O processo de re-anotação é importante em genomas que foram anotados por uma abordagem diferente da que será utilizada como padrão, nesse caso, é interessante padronizar a anotação para um melhor aproveitamento dos dados, executando todas as anotações em um único e padronizado anotador.

2.6.3 Dados genômicos e o potencial biotecnológico

Diversos estudos vem buscando aplicações biotecnológicos em microrganismos, como no fungo *Aspergillus flavus* (DE VRIES et al., 2017). Nesse estudo, foi identificado que metabólitos secundários, genes e *clusters* de genes potencialmente envolvidos na formação de aflatoxina e outros metabólitos e que atuam na degradação de polímeros complexos como carboidratos. A análise de vias metabólicas secundárias facilita a descoberta de novos compostos com propriedades farmacêuticas, bem como novas enzimas para degradação da biomassa.

O sequenciamento é uma ferramenta muito importante para obtenção de uma quantidade muito rica de dados. Com a vinda do NGS, muitos dados são possíveis de serem gerados diariamente e com isso diversas tecnologias que podem vir a ser descobertas ou padrões novos identificados, como exemplo, os TALENS (*Transcription Activator-like Effectors*) e o sistema CRISPR/CAS (*Clustered Regularly Interspaced Short Palindromic Repeats/ CRISPR associated proteins)*.

Os TALENS são comuns em *Xanthomonas* sp. *e* são proteínas secretadas pela bactéria para ajudar na infecção de espécies vegetais (CLEVELAND et al., 2009). A aplicação do sistema TALEN na genômica se deu pela sua capacidade de forte reconhecimento de nucleotídeos específicos, levando então a aplicações como ativação, supressão, deleção e inserção de uma sequência de DNA (MOORE; CHANDRAHAS; BLERIS, 2014).

O sistema CRISPR funciona como uma espécie de sistema imune contra fagos (MAHFOUZ; LI, 2011), que foi identificado primeiramente em genomas de *Escherichia coli.* e posteriormente foi descoberta uma aplicação em edição genética, o que

revolucionou a área (ISHINO et al., 1987, BARRANGOU; HORVATH, 2012). O que reforça que tecnologias novas podem surgir por genes identificados e padrões novos.

2.7 Abordagem pangenômica

Estudos de genômica comparativa tem ganhado muito espaço, visto que existe o potencial em microrganismos, os estudos comparativos são aspectos-chave da biologia, incluindo estudos em metabolismo primário e secundário, resposta ao estresse, degradação da biomassa e transdução de sinal, revelaram a conservação e a diversidade entre as espécies e estes estudos podem ser usados para a busca de genes e proteínas com potencial biotecnológico (SANDER; JOUNG, 2014).

Genes homólogos são aqueles segmentos de DNA que possuem a mesma origem, mas que podem ter funções idênticas ou não. Os genes homólogos são divididos em ortólogos, parálogos e xenólogos. Os genes ortólogos são os genes que possuem a mesma função e uma origem em comum, tendo se separado no surgimento de uma nova espécie, ou seja, durante o processo de especiação. Assim, todo descendente terá sua própria cópia do gene. Os genes parálogos, ao contrário dos ortólogos, têm funções distintas e são gerados dentro de uma mesma espécie quando ocorre uma duplicação no genoma.

O pangenoma é a interação geral e comparação entre informações genéticas de organismos de um mesmo gênero ou espécie. São discutidas características da estrutura do Pangenoma e processos que controlam o que é preservado nos genes e a sua variabilidade genética, que são os aspectos evolutivos, adaptativos e as interações de patógeno-hospedeiro (GLASNER et al., 2008).

A utilização da abordagem pangenômica tem como objetivos o prognóstico da situação epidêmica, desenvolvimento de métodos de profilaxia e diagnósticos e também da avaliação de possíveis consequências da genética molecular e intervenções de engenharia genética (DE VRIES et al., 2017).

O pangenoma pode se separar em três conjuntos de dados: I) Genoma núcleo ou *core genome*, II) Genoma acessório e III) Genes espécie-específicos (Figura 4), (TETS, 2003).

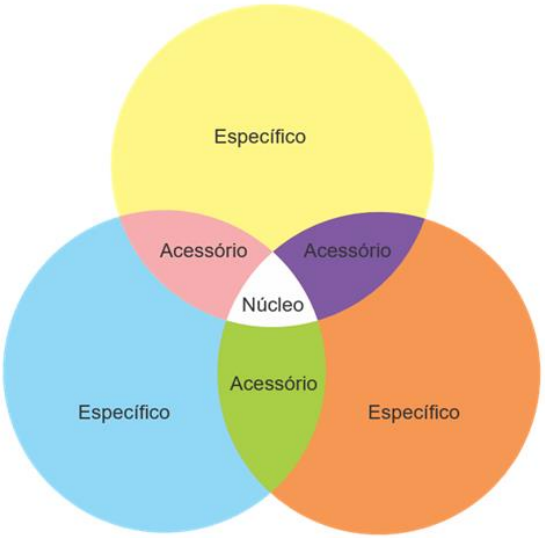

Figura 4. Ilustração de um Pangenoma por diagrama de Venn, dividindo as categorias de um pangenoma. Onde em branco temos o núcleo (core), que são todos os genes compartilhados por todos os genomas. As partes em rosa, verde e roxo são as respectivas partes do genoma acessório, onde fica a região dos genes que são compartilhados por mais de um genoma. Já as partes azul, laranja e amarelo demonstram a parte dos genes espécie-específicos, que são exclusivos de um isolado. Adaptado de MUZZI; DONATI, (2011ª).

O genoma núcleo é o conjunto de genes que estão presentes em todos os genomas analisados no pangenoma. Em geral esses genes são característicos de funcionamentos básicos de um organismo, como replicação, tradução e manutenção do equilíbrio celular. Assim sendo, quanto mais distante for um organismo do outro, menos genes no core teremos em comum, demonstrando que são filogeneticamente distantes. Genomas que tem grande proximidade filogenética irão apresentar um grande número de genes no núcleo (MUZZI; DONATI, 2011b).

O genoma acessório, são os genes que são compartilhados por alguns dos isolados, tendo mais de um genoma em comum. O que se tem conhecimento é que esses genes estão mais ligados com as funções relacionadas a sua virulência, sobrevivência a estresses e resistência a antibióticos. Ao que se imagina, grande parte desses genes são obtidos via transferência horizontal e de evolução parafilética (UDAONDO et al., 2016).

As regiões do pangenoma onde apenas um genoma tem o gene, são chamadas de espécie-específicos, que irão ter o papel de funções adaptativas. Muitos desses genes estão relacionados à patogenicidade e virulência de organismos patogênicos. Para organismos que não são patogênicos, tem se conhecimentos que

esses genes estão mais relacionados a rotas metabólicas (BROCKHURST et al., 2019).

Existem duas denominações para pangenomas, visto sua distribuição dos genes, que são os pangenomas abertos e pangenomas fechados. Aberto significa que há uma grande chance de novos genomas incluírem novos genes ainda não descritos para o gênero ou espécie é alto, o que é um caso comum para organismos que possuem um pangenoma grande e que novos genes irão ser identificados devido ao seu tamanho. O pangenoma fechado se aplica a organismos que vivem isolados e restritos a um nicho, logo essa situação dificulta a transferência horizontal de genes e consequentemente o pool de genes não está em expansão, como no aberto (MUZZI; DONATI, 2011c).

3. Objetivos

3.1 Objetivo geral

Esse estudo teve como principal objetivo a análise *in silico* do pangenoma de bactérias do gênero *Curtobacterium* sp.

3.2 Objetivos específicos

- Sequenciar, montar e anotar estruturalmente os genomas isolados de *Curtobacterium* sp.;
- Obter os genomas de *Curtobacterium* sp. no banco de genomas e re-anotar estruturalmente;
- Identificar grupos ortólogos;
- Mapear os genes de rotas metabólicas;
- Realizar a anotação funcional.

4. Material e Métodos

O resumo das abordagens utilizadas na montagem, anotação, seleção, filtragem e re-anotação estrutural dos genomas utilizados, assim como as abordagens

pangenomicas estruturais, análises metabólicas, dos grupos ortólogos, anotação funcional e de proteínas efetoras estão descritas na Figura 5.

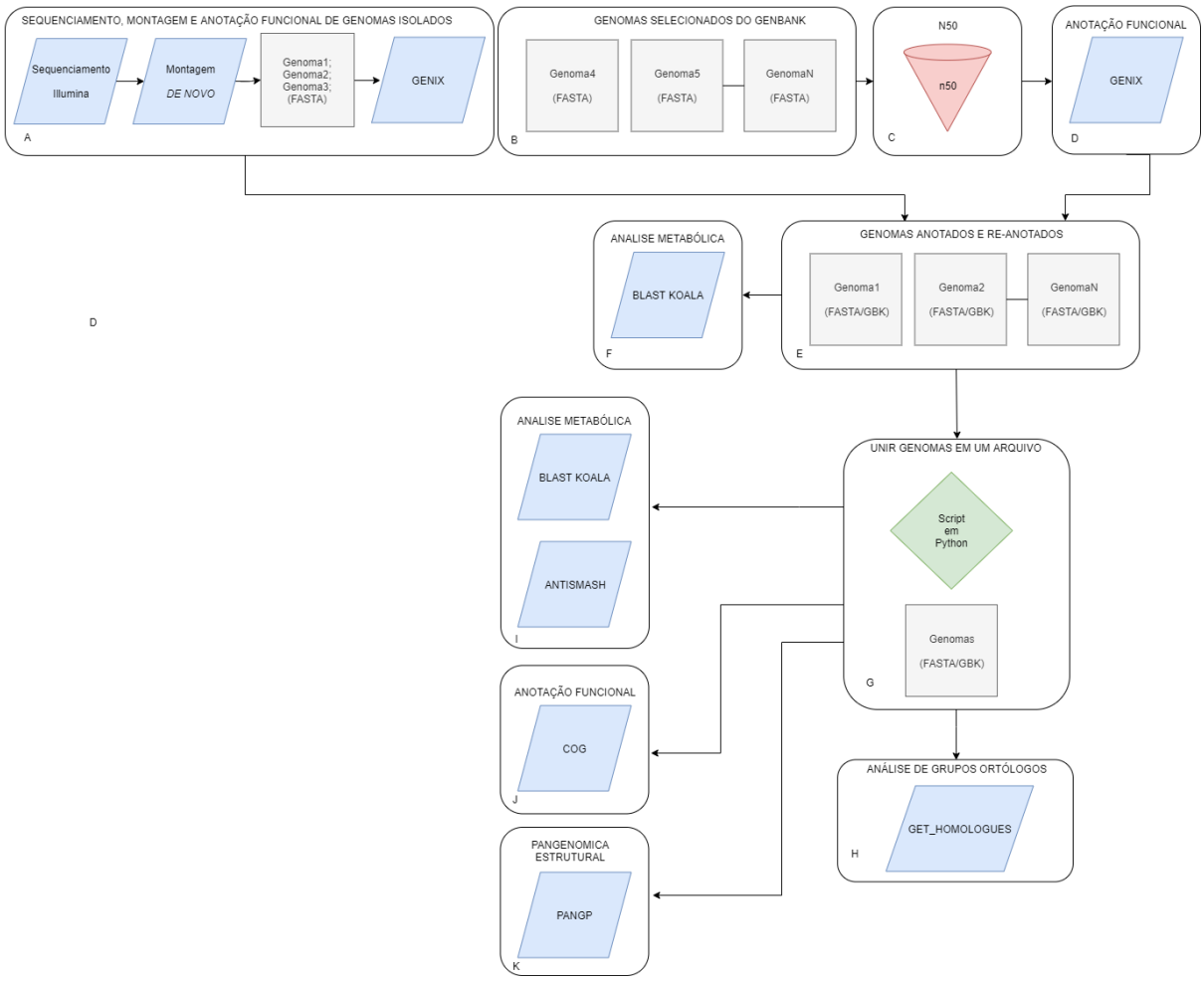

Figura 5. Resumo da metodologia utilizada para a análise do pangenoma de *Curtobacterium* sp. (A) Utilizando os arquivos recebidos das empresas, realizamos a montagem e anotação dos arquivos de genomas, totalizando em 3 genomas isolados de *Curtobacterium* sp. (B) Utilizando o banco de dados GenBank do NCBI foram selecionados 16 genomas de *Curtobacterium* sp., sendo 3 genomas completos e 13 rascunhos; (C) Utilizando a métrica N50, os 16 genomas foram filtrados em 3 genomas representativos; (D) Re-anotação estrutural dos genes dos 3 genomas obtidos do GenBank e a anotação estrutural dos 3 isolados sequenciados, utilizando a ferramenta GENIX; (E) Genomas re-anotados no formato (.fasta) e (.gbk); (F) Os arquivos (.fasta) dos genomas re-anotados foram submetidos na plataforma online do BlastKoala e da plataforma online antiSMASH. (G) Os 6 genomas reanotados foram unidos em um único arquivo por meio de um script; (H) Para cada genoma através do GET_HOMOLOGUES fez-se a análise de *clusters* dessas proteínas efetoras; (I) Análise Metabólica utilizando o BlastKoala e antiSMASH do core genoma gerado pela análise do GET_HOMOLOGUES; (J) Anotação Funcional dos genes do core genoma utilizando o COG; (K) Para determinar o tipo de pangenoma para *Curtobacterium sp.* utilizou-se a ferramenta PanGP.

4.1 Sequenciamento, montagem e anotação de isolados genoma de *Curtobacterium*.

Os DNA genômico foi isolado em colaboração com o Departamento de Fitossanidade da Universidade Federal de Pelotas, mantido pela Prof. Dra. Andréa Bittencourt Moura em conjunto com o Laboratório de Bioinformática e Proteômica, da Universidade Federal de Pelotas, onde o primeiro e o terceiro isolados foram enviados para a empresa NeoProspecta (https://neoprospecta.com/), e o segundo para a empresa GenOne (www.genone.com.br). Os genomas foram sequenciados utilizando a plataforma de sequenciamento Illumina MiSeq, com sua bliblioteca "*paired-end"*, que gerou os arquivos de saída no formato (.fastq).

Para a realização da montagem, a utilização do método *De novo* foi utilizada e os arquivos foram posteriormente processados para a montagem no método. Utilizando os montadores SPAdes (COMPUTATIONAL PAN-GENOMICS CONSORTIUM, 2018) , ABySS (BANKEVICH et al., 2012) , Ray (SIMPSON et al., 2009) , CLC Genomics WortkBench (https://www.qiagenbioinformatics.com/), A5 (BOISVERT; LAVIOLETTE; CORBEIL, 2010) e Velvet (COIL; JOSPIN; DARLING, 2014) , os arquivos de saída no formato (.fasta) dos montadores foi utilizado em uma montagem de melhor qualidade no integrador de montagens CISA (ZERBINO; BIRNEY, 2008) , que realizou a montagem consenso que foi utilizada para posterior anotação. Os genomas foram submetidos ao genbank e com seus respectivos BioProjects identificados, a denominação e seu respectivo código de acesso no GenBank são: Csp1 (PRJNA508935), Csp2 (PRJNA508936) e Csp3 (PRJNA548463).

A anotação foi realizada com o a ferramenta GENIX (KREMER et al., 2016), utilizando linha de comando no terminal. Esse programa utiliza como arquivo de entrada o formato (.fasta), sendo ele um *pipeline,* o mesmo está descrito no anexo A, para anotação, utiliza uma combinação de ferramentas de bioinformática como o Prodigal (HYATT et al., 2010), tRNAscan-SE (HYATT et al., 2010), RNAmmer(LOWE; EDDY, 1997), Aragorn (LAGESEN et al., 2007), HMMER (LASLETT; CANBACK, 2004), BLASTn, INFERNAL (EDDY, 1995), Rfam (NAWROCKI; KOLBE; EDDY, 2009), Antifam (GRIFFITHS-JONES et al., 2003) e o conjunto de dados não redundante gerado pelo CD-HIT (EBERHARDT et al., 2012). Os arquivos de saída dessa ferramenta são no formato GenBank (.gbk), (.gff) e formato Feature Table (.tbl),

FASTA com arquivos contendo sequências das proteínas (.faa), features não codificadas (.ffn) e sequências de DNA codificadas (.fna).

4.2 Obtenção dos genomas de *Curtobacterium* sp. no GenBank

A obtenção dos genomas foi realizada através do banco de dados GenBank, presente na plataforma do NCBI. Dezesseis genomas de *Curtobacterium* sp. Foram obtidos sendo que destes, 3 genomas foram selecionados e 13 genomas foram considerados inaptos para a análise, visto a qualidade dos rascunhos, desta forma, apenas os genomas com cobertura maior foram utilizados. Os genomas escolhidos são respectivamente: CP017580, CP018783, LT576451.

4.3 Re-anotação estrutural dos genomas do GenBank

A re-anotação estrutural dos 3 genomas selecionados do GenBank foi realizada utilizando a ferramenta GENIX (KREMER et al., 2016).

4.4 Análise dos grupos ortólogos

Com os genes re-anotados, foi feita a análise dos grupos ortólogos utilizando a ferramenta GET_HOMOLOGUES (CONTRERAS-MOREIRA; VINUESA, 2013). Este programa tem a função de identificar *clusters* de proteínas e sequências de nucleotídeos homólogos em sequências similares, identificar grupos ortólogos de genes flanqueados por fases de leitura aberta (ORF, *open reading frames*) ao longo do genoma, definindo o core genoma e o pan genoma. O algoritmo está descrito no anexo B.

O GET_HOMOLOGUES tem como entrada os formatos (.gbk) e (.faa), diversos parâmetros foram utilizados, entre as diferentes etapas realizadas no programa. Após instalar todos os módulos necessários para o funcionamento da ferramenta, foi selecionado quais parâmetros seriam usados, dentre os diversos disponíveis pelo programa. Aqueles selecionados foram: -d (arquivo de entrada e configurações padrão), -t=1 (selecionar grupos ortólogos que estão presentes em pelo menos um dos genomas e configurações padrão), -M (para gerar os arquivos de Matrix).

4.5 Anotação funcional dos genes

A anotação funcional dos genes foi realizada utilizando a ferramenta *Cluster of Ortholog Groups* (COG) (TATUSOV et al., 2003) foi executada via script.

4.6 Análise pan-genômica estrutural

Para saber se um pangenoma é aberto ou fechado, utilizou-se a ferramenta de bioinformática PanGP (TATUSOV et al., 2003). O arquivo de entrada desse programa é o (.fasta), com isso, utilizando o programa será gerado um gráfico que demonstra a classificação do pangenoma.

Para utilizar o PanGP, outras ferramentas são utilizadas como dependências para rodá-lo, como PGAP(ZHAO et al., 2014), OrthoMCL (ZHAO et al., 2012) e PanOCT (LI; STOECKERT; ROOS, 2003).

4.7 Análise metabólica

Duas análises metabólicas foram realizadas, uma com o *core* genoma, que é resultado da ferramenta (Analise metabólica para o core - pangenoma – metabolitos secundários compartilhados) GET_HOMOLOGUES e outra com cada um dos genomas, de forma individual (Analise metabólica para cada genoma de forma individual geral do genoma, comparação). A ferramenta utilizada em ambos os casos foi o BlastKoala (FOUTS et al., 2012). O *core* genoma foi gerado através de um script em Python, que foi avaliado com o CD-HIT (KANEHISA; SATO; MORISHIMA, 2016) e posteriormente montado utilizando um script que uniu todos os genes do *cor*e genoma em um único arquivo (.fasta).

O arquivo de saída gerado foi submetido ao BlastKoala, para realizar as análises, assim como o restante dos genomas que formam esse pangenoma. Também foi utilizado as saídas do GENIX no formato (.faa) para submeter ao BlastKoala.

Os parâmetros utilizados para a submissão dos arquivos foram a utilização de um “*Taxonomy ID”* que foi preenchido o campo com o código taxonômico da espécie *Curtobacterium s*p. (2034), no qual foi obtido do banco de dados do NCBI, e o banco

de dados de genes utilizado no BlastKoala foi o "*species_prokaryotes*". Com essas informações, cada genoma foi submetido e seus resultados gerados.

Para cada genoma foi realizada também análise de metabólicos secundários, que utilizou a ferramenta via web-server antiSMASH (LI; GODZIK, 2006). Cada genoma então foi enviado com os seguintes parâmetros *"Extra Features*": "*All On*". Isso possibilita que ele utilize todos os bancos disponíveis.

5. Resultados e Discussões

5.1 Análise dos grupos ortólogos

A ferramenta GET_HOMOLOGUES, demonstrou um número de 2.119 grupos ortólogos consensos dentre os diferentes algoritmos que ele integra. Foram encontrados 2.197 grupos ortólogos com a ferramenta COG. Através do BDBH foi obtido 2.205 e com o OrhoMCL foram identificados 2.227 (Figura 6).

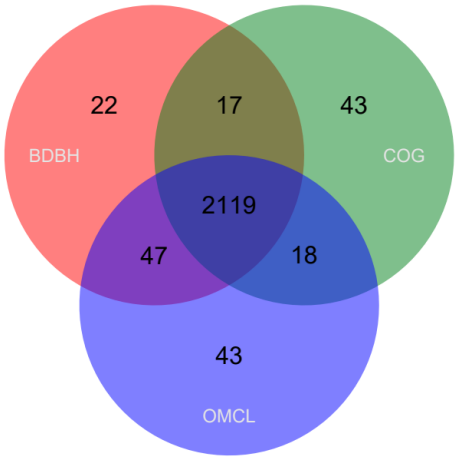

Figura 6. Diagrama dos grupos ortólogos do pangenoma de *Curtobacterium* sp. Em rosa a predição dos grupos ortólogos pelo algoritmo BDBH, em verde os ortólogos preditos pelo COG e em roxo os grupos ortólogos gerados pelo OrthoMCL.

Um dos resultados do GET_HOMOLOGUES é uma figura representativa do pangenoma, que é formado pelo núcleo (core), com 2.137 *clusters*, Cloud 2.499, *Soft*

core 2.413 e *Shell* 745. O total de *clusters* de genes é de 5.657 nos 6 genomas. (Figura 7).

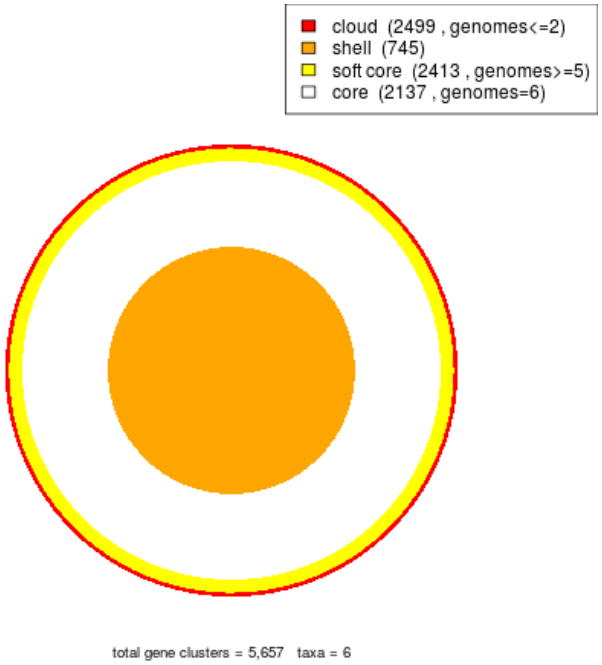

Figura 7. Pangenoma organizado de acordo com o *software* GET_HOMOLOGUES. O total de clusters de genes foi de 5,657. O programa tem maneiras diferentes de tratar o pangenoma, com nomenclaturas diferentes do usual, utilizando cloud para genoma acessório, shell para genoma espécie-específico e a categoria nova que é a soft core, que irá de genes compartilhados por quase todos os genomas, >=5, sendo que o total de genomas é 6. Então temos em vermelho são os genes que se encontram em 2 genomas no máximo, em laranja encontra-se o *shell*, que são genes que são únicos de cada genoma, em amarelo estão os *soft-core*, que são genes que são compartilhados por menos que 5 genomas e em branco, temos o *core* genoma, que são os genes compartilhados por todos os 6 genomas.

A partir da análise de grupos ortólogos, foi possível identificar um número esperado de grupos consensos, reforçando assim, a utilização de múltiplos algorítmicos para a uma análise mais acurada, com resultados do GET_HOMOLOGUES, com as ferramentas implementadas internamente como o COG, BDBH e OrthoMCL. Como relata o desenvolvedor do GET_HOMOLOGUES CONTRERAS-MOREIRA et al., (2017) esse modelo de análise, com a utilização de algoritmos internos ajuda na acurácia dos resultados. Foi realizada uma análise utilizando o CD-HIT, onde foram selecionados os genes representativos, filtrando de cada *cluster* os genes representativos dos *clusters*, obtendo então o core genoma filtrado somente com as sequencias representativas.

5.2 Análise pan-genômica estrutural

Utilizando a ferramenta PanGP, foi possível gerar resultados referentes aos seis genomas analisados. Foram obtidos aproximadamente 3400 a 5600 grupos ortólogos, o core genoma teve 1850 a 3900 grupos. Sobre as características desse pangenoma, podemos dizer que ele é fechado, devido a adição do pool de genes ser mínima em relação com todos, quando se comparam todos contra todos. (Figura 8).

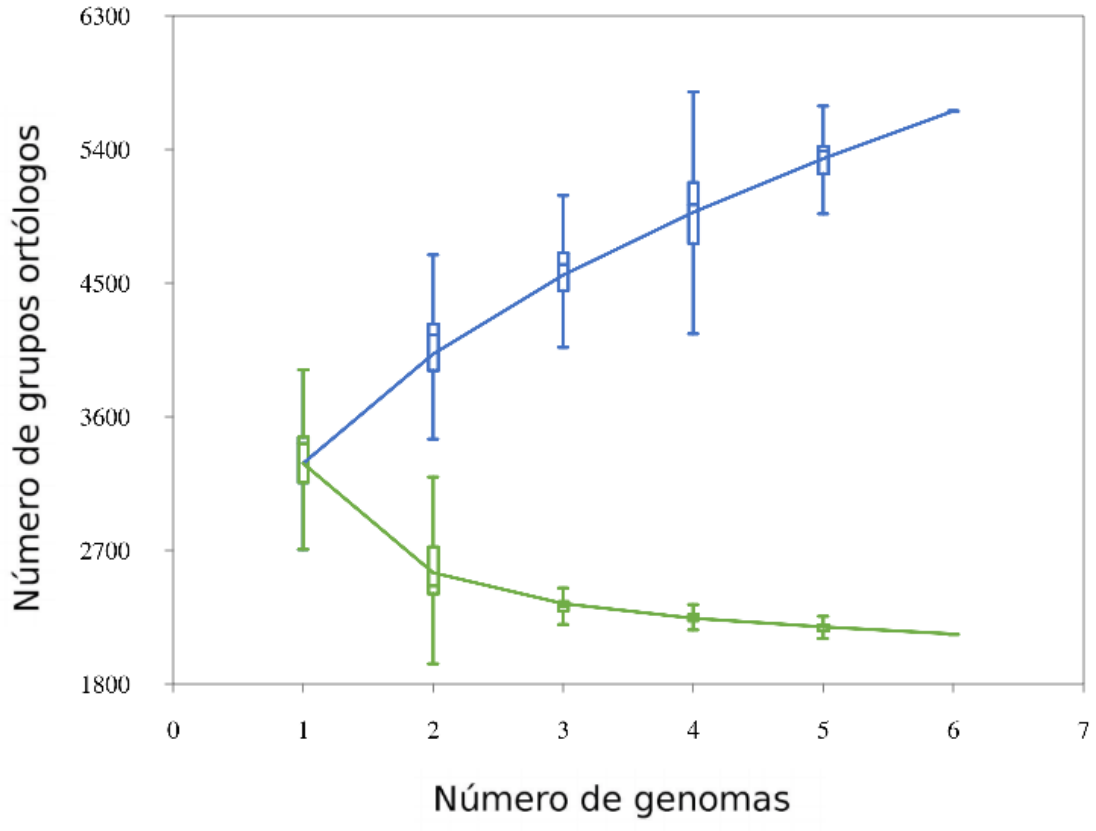

Figura 8. Gráfico descrevendo o *core* genoma e pangenoma de *Curtobacterium* sp. em relação os grupos ortólogos. O número de genomas utilizados encontra-se no eixo x, no eixo y o número de grupos ortólogos. Em azul representam pangenoma e os dados em verde representam o core genoma.

As comparações são feitas entre os genomas, onde na primeira marca, se compara o genoma um contra ele mesmo, o segundo genoma é comparado com o primeiro e também contra o segundo, assim, comparando uns contra os outros, até o último estágio, onde eles são comparados todos contra todos, essa variabilidade se

no gráfico se dá devido a comparação, por exemplo, de cinco genomas comparados contra 5, sendo então uma variabilidade alta.

A análise de características do pangenoma dos 6 genomas analisados revelou que é do tipo fechado, já que o gráfico gerado pelo *software* PanGP (Figura 8), que, quando mais genomas fossem adicionados, teria menos grupos ortólogos sendo identificados. Isso é causado por possivelmente o *core* genoma representar grupos ortólogos em comum entre todos os genomas em questão, como cada um apresenta uma diferença de especificidade aos hospedeiros, logo, possuem genes que irão interagir exclusivamente a cada um, e a adição de genomas não adicionar para a determinação do core genoma.

Um dado que confirma isso são os números de grupos ortólogos demonstrados, uma vez que os genomas comparados todos contra todos não apresentam nenhuma adição ao pool genético, o que demonstra que a adição de novos genomas não contribuirá para a determinação do *core* genoma de *Curtobacterium sp.*

5.3 Anotação funcional dos genes

A partir dos dados do *core* genoma, foi possível identificar 2.119 *clusters* preditos a partir do GET_HOMOLOGUES (Figura 6).

A anotação funcional realizada a partir do software COG, os resultados obtidos são que 1348 genes têm função desconhecida ou não encontrada no banco de dados, 702 genes têm função geral apenas de predição, ou seja, não identificado, mas que apresenta possivelmente uma função. Não identificada, 270 genes estão relacionados ao transporte e metabolismo de aminoácidos, 258 genes estão relacionados à produção e conversão de energia, 245 genes estão relacionados à tradução, estrutura ribossomal e biossíntese, 238 genes estão relacionados à replicação, recombinação e reparo, 231 genes estão relacionados a transcrição, 230 genes estão relacionados a transporte e metabolismo de carboidratos, 212 genes estão relacionados a transporte e metabolismo de íons inorgânicos, 203 genes estão relacionados à modificação pós-traducional, renovação de proteínas e chaperonas, 188 genes relacionados à parede celular, membrana e biogênese de envelopes, 179 genes relacionados a transporte e metabolismo de coenzimas, 158 genes relacionados ao tráfico intracelular, secreção e transporte vesicular, 152 genes relacionados a mecanismo de transdução de sinal, 96 genes relacionados com motilidade celular, 95

genes relacionados com transporte e metabolismo de nucleotídeos, 94 genes relacionados com metabolismo lipídico, 88 genes relacionados à biossíntese, transporte e catabolismo de metabólitos secundários, 72 genes relacionados ao controle do ciclo celular, divisão celular e particionamento cromossômico, 46 genes relacionados a mecanismos de defesa, 25 genes relacionados ao processamento e modificação de RNA, 19 genes relacionados à estrutura e dinâmica da cromatina, 12 genes relacionados ao citoesqueleto, 2 genes relacionados à estrutura nuclear e 1 gene relacionado a estruturas extranucleares. (Figura 9)

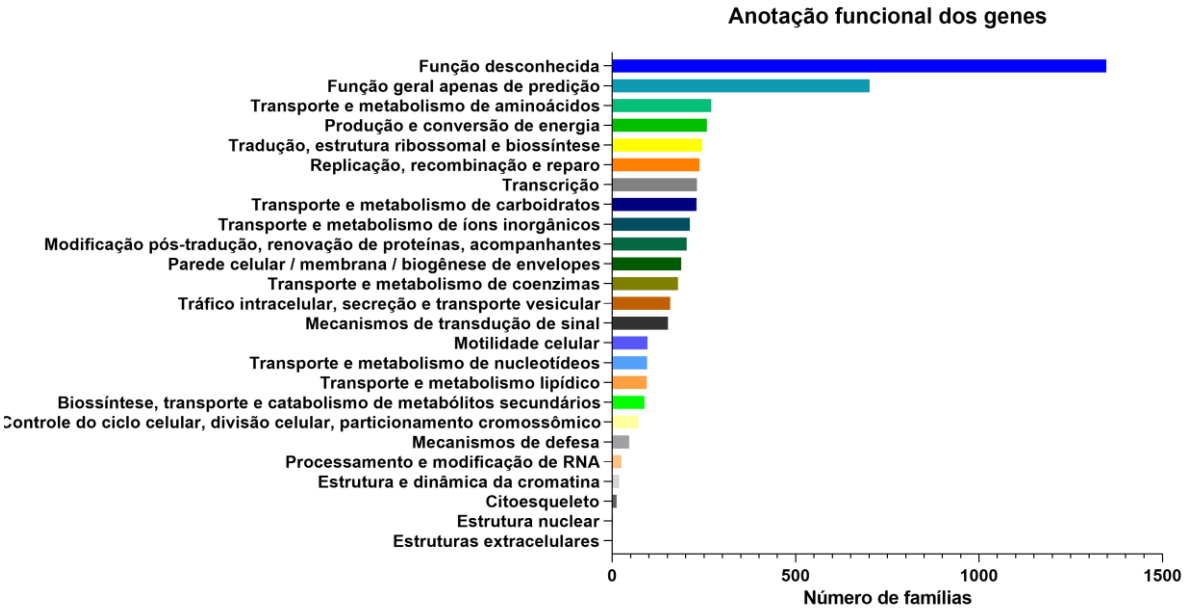

Figura 9. Funções preditas para o *Core* Genoma de *Curtobacterium* sp.

A anotação funcional dos grupos ortólogos do *core* genoma é de grande importância para determinar os processos celulares e metabólicos são compartilhados entre o gênero. Em sua maioria, os grupos sem função foram os mais encontrados, o que sugere que esses genes não possuam similaridade com o banco disponível no COG ou alguma função desconhecida ou não elucidada, sugerindo uma análise futuramente do genoma acessório, com a finalidade de determinar as funções específicas de cada genoma, pegando então suas características exclusivas para poder investigar alguma aplicação para os genes do genoma acessório.

5.4 Análise Metabólica

5.4.1 Rotas metabólicas

Foi realizada a análise metabólica do core genoma (Figura 10) e dos 6 genomas separadamente (Figura 11), no qual revelou rotas muito importantes, como de processos celulares, processamento de informações genéticas (tradução, transcrição e replicação), informações do ambiente (resposta a alterações ambientais), metabolismo (rotas de catabolismo e anabolismo), rotas de defesa (degradação de xenobióticos).

Em média o número de *hits* por genoma foi sempre abaixo da metade, um número considerável de proteínas que não tiveram nenhum *hit* e outras não possuíam nenhum processo metabólico vinculado. Já no core genoma, o número de *hits* superou mais da metade. Tendo então mais proteínas totais relacionadas ao metabolismo.

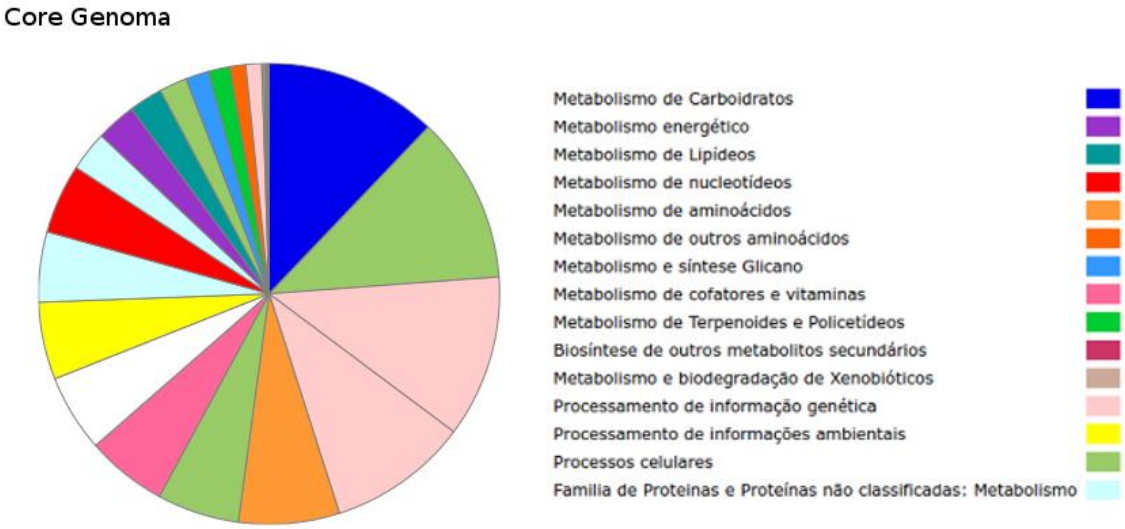

Figura 10. Análise do metabolismo do *core* genoma, a partir do *software* BlastKoala. Essas rotas foram preditas pelo *software*, onde as regiões relacionadas ao metabolismo de carboidratos, metabolismo energético, metabolismo de lipídeos, metabolismo de nucleotídeos, metabolismo de aminoácidos, metabolismo de outros aminoácidos, metabolismo de cofatores e vitaminas, metabolismo de terpenoides e policetídeos, biossíntese de outros metabolismos secundários, metabolismo e biodegradação de xenobióticos, processamento de informação genética, processamento de informações ambientais, processos celulares e por último proteínas não classificadas e famílias de proteínas relacionadas com metabolismo.

Figura 11. Análise do metabolismo preditos pelo *software* BlastKoala dos 6 genomas de *Curtobacterium sp.* A, B, C, D, E e F, sendo eles respectivamente A (CP017580), B (CP018783), C (LT576451), D (Csp1), E (Csp2) e F (Csp3). Essas rotas foram preditas pelo *software* BlastKoala, onde as regiões relacionadas ao metabolismo de carboidratos, metabolismo energético, metabolismo de lipídeos, metabolismo de nucleotídeos, metabolismo de aminoácidos, metabolismo de outros aminoácidos, metabolismo de cofatores e vitaminas, metabolismo de terpenoides e policetídeos, biossíntese de outros metabolismos secundários, metabolismo e biodegradação de xenobióticos, processamento de informação genética, processamento de informações ambientais, processos celulares e por último proteínas não classificadas e famílias de proteínas relacionadas com metabolismo.

A análise metabólica do *core* genoma demonstrou que em grande parte, os grupos encontrados são relacionados ao metabolismo, significando então que a maioria dos

processos estão relacionados ao catabolismo e ao anabolismo. Tendo em vista que uma grande quantidade não apresentou nenhum *hit,* um pouco menos da metade e em seguida a predição para nenhum processo metabólico também representa uma grande porção desse *core* genoma MENG et al., (2017), já relatou em seu trabalho com pangenoma de *Streptococcus mutans,* onde 35,6% dos grupos ortólogos eram ligados ao metabolismo.

Para cada genoma em relação as rotas metabólicas, houve uma regularidade quanto a presença de processos do metabolismo, fisiologia, processos celulares, processamento de informações ambientais, processamento de informações genéticas, que são basicamente processos comuns para o funcionamento básico desses isolados, porém, as predições para as proteínas sem função ou sem *hit,* demonstram que há diferenças entre os mesmos e isso provavelmente está associado aos genes exclusivos relacionados a especificidade a hospedeiros e patogênese, tornando-as importantes alvos de investigações futuras.

5.4.2 Análise de metabólitos secundários

Foi realizada uma análise de metabólitos secundários dos 6 genomas, utilizando a ferramenta antiSMASH, anexo C.

O genoma CP017580 teve como resultado 7 *clusters* com similaridade identificados, sendo 2 deles NRPS-*Like* (região 1 e 2), onde um deles é encontrado com uma similaridade de 13% com *Arginomycin*; na região 3, o T3PKS, com similaridade de 23% com o *cluster* Pentalenolactone; na região 4 o tipo Terpeno, que tem similaridade com Carotenoides de 50%; a região 5 o tipo Linaridin, foi encontrada com uma similaridade de 4% ao *cluster Foxicins* A-D; na região 6 a *Siderophore* com 60% de similaridade ao *Desferrioxidamine* B; e na região 7 o tipo *Betalactone* com 11% de similaridade ao *cluster Gobichelin*.

O genoma CP018783 teve 7 regiões encontradas, onde as 2 primeiras regiões são NRPS-*Like,* com uma similaridade de 13% ao *cluster* de *Arginomycin;* a região 3 é do tipo *Siderophore*, que tem similaridade de 60% com o *cluster* de *Desferrioximine* B; a região 4 tem o tipo *Bacteriocin*, mas não tem similaridade com nenhum *cluster*, a região 5 é do tipo Terpeno, com uma similaridade de 50% com o *cluster* Carotenoides; a região 6 é do tipo T3PKS, onde a sua similaridade é de 23% com *Pentanolactone; e* a região 7 é do tipo *Betalactone* similar em 7% com *Microansamycin.*

O genoma LT576451 teve 7 regiões identificadas. Na região 1, do tipo Siderophore, com similaridade de 75% em *Desferrioxamine;* na região 2 o tipo Terpeno, com similaridade em 50% com Carotenoides; a região 3 é do tipo T3PKS, com similaridade de 6% ao *Bottromycin* A2; a região 4 é do tipo *Bacteriocin*, sem cluster similar; a região 5 é do tipo *Betalactone*, com similaridade de 7% ao *Microansamycin*, as regiões 6 e 7 são do tipo NRPS-*Like,* tendo a região 6 similar em 13% ao *cluster* da *Arginomycin*, enquanto a região 7 não é similar a nenhum.

O genoma Csp1 teve 9 regiões identificadas. A região 1 é do tipo NRPS-*like,* similar em 3% ao *cluster* A400926; a região 2 é do tipo *Siderophore*, com similaridade de 75% com *Desferrioxamine;* na região 3 o tipo identificado é o NRPS-*like*, com similaridade de 13% a *Arginomycin*, a região 4 é do tipo T3PKS, com similaridade de 6% em *Hedamycin;* a região 5 e a região 9 são do tipo Terpeno, com similaridade de 50% em Carotenoide; as regiões 6 e 7 são do tipo *Bacteriocin*, sem *cluster* similar; a região 8 é do tipo *Betalactone*, com similaridade de 7% ao *cluster* Microansamycin.

O genoma Csp2 teve 9 regiões encontradas, sendo a região 1 e 3 do tipo NRPS-like, sendo a região 3 similar em 13% com o *cluster Arginomycin;* a região 2 é do tipo *Siderophore*, com similaridade de 75% com *Desferrioxamine;* a região 4 é do tipo *Betalactone*, com similaridade de 7% com *Microansamycin;* a região 5 é do tipo T3PKS, com similaridade de 6% com *Hedamycin;* a região 6 é do tipo Terpeno, com similaridade de 66% ao *cluster* dos carotenoides, as regiões 7, 8 e 9 são do tipo *Bacteriocin,* com somente a região 9 similar a um *cluster*, sendo 6% parecida com *Oxazolomycin.*

O genoma Csp3 teve 8 regiões identificadas, sendo as regiões 1 e 6 do tipo NRPS-*like*, apenas a região do tipo 1 tem similaridade, sendo ela de 13% com o *cluster Arginomycin;* a região 2 é do tipo *Siderophore*, similar em 75% ao *cluster Desferrioxamine;* a região do tipo 3 é do tipo T3PKS, com similaridade de 4% ao *Phthoxazolin;* a região 4 é do tipo Terpeno, com similaridade de 50% em Carotenoide, a região 5 e 7 são regiões de tipo *Bacteriocin*, sem similaridade a nenhum *cluster* e a região 8 é do tipo *Betalactone*, com similaridade em 7% com *Microansamycin*.

A análise de metabolitos secundários revelou uma grande presença de fatores bem interessantes, como uma das características, como a revelada para a variação de cores em colônias (Figura 3), onde temos uma variação que vai do amarelo ao vermelho (AGARKOVA et al., 2012), o que pode estar relacionada a produção de carotenoides da bactéria, que foram identificados nas análises do antiSMASH,

causada via transferência horizontal. Carotenoides também são grandes antioxidantes e são utilizados na indústria alimentícia por não causarem reações alérgicas.

Foi demonstrado que a presença desta bactéria no solo ao redor de algumas plantas causa a redução de alguns sintomas de certas doenças. Isso pode estar relacionado com a bactéria apresentar vias metabólicas secundárias de produção de Bacteriocinas, foi confirmada também no antiSMASH, como um metabolito secundário. Essas proteínas são toxinas que inibem o crescimento de outras bactérias. São semelhantes ao fator killer de levedura. Um fator interessante é que esses peptídeos estão sendo estudadas para serem um possível novo antibiótico de espectro estreito (MIETHKE, M.; MARAHIEL, M. A., 2007). Além, também são utilizadas na indústria dos alimentos, tendo relação com o tempo de prateleira (COTTER; ROSS; HILL, 2013).

Um dos metabolitos secundários que se destacou também foi o Sideróforo, "*Siderophore"*, que em grego significa "Carregador de ferro", esse metabólito serve para ajudar no transporte de ferro entre as membranas das bactérias (NEILANDS, 1995). Os Sideróforos estão relacionados a um mecanismo adaptativo, onde o solo é pobre em ferro e se torna resposta um recurso de captação, nas bactérias gram-positivas ricas em GC (por exemplo o mesmo filo da *Curtobacterium*, *Actinobacteria*) o DtxR (repressor da toxina da difteria), a *Corynebacterium diphtheriae* conhecido como a produção da toxina da difteria, que também é regulada por este sistema como relata MIETHKE; MARAHIEL, (2007). Sendo então um mecanismo de virulência presente nesta bactéria, já que possibilita a sobrevivência e captação de ferro do ambiente. Não foram identificados com confiabilidade nenhum outro metabólito com potencial de utilização biotecnológica.

A relação dessa bactéria com o feijão e a causa da murcha ainda não está bem elucidada, mas o que podemos afirmar, é que esses metabolitos secundários ajudam na resistência da bactéria com a planta e com os estresses para manutenção, como já abordado em BRAS; BIOL, (2010)

7. Conclusão

A partir da abordagem utilizada foi possível sequenciar, montar e anotar os genomas de *Curtobacterium* sp., foi demonstrado que o pangenoma de *Curtobacterium* sp. é fechado, foi possível identificar rotas metabólicas para os

isolados de *Curtobacterium sp.*, assim como para o genoma núcleo. A busca por metabótitos secundários dessa bactéria, nos possibilitou entender um pouco mais sobre a patogenicidade e a sua variação entre os isolados, visto que em seu metabolismo secundário apresenta fatores de virulência que ajudam na patogenicidade. Como perspectivas futuras, análises do pangenoma acessório são de grande importância para caracterização de possíveis genes de patogenicidade, também fazer uma análise mais profunda sobre o metabolismo secundário buscando entender sobre os mecanismos e desenvolver um produto biotecnológico.

8. Referências

AGARKOVA, I. V et al. Genetic diversity among Curtobacterium flaccumfaciens pv. flaccumfaciens populations in the American high plains. **Canadian journal of microbiology**, v. 58, n. 6, p. 788–801, jun. 2012.

ALLEN, E. E.; BANFIELD, J. F. **Community genomics in microbial ecology and evolutionNature Reviews Microbiology**, jun. 2005.

BANKEVICH, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of Computational Biology**, v. 19, n. 5, p. 455–477, 2012.

BARRANGOU, R.; HORVATH, P. CRISPR: New Horizons in Phage Resistance and Strain Identification. **Annual Review of Food Science and Technology**, v. 3, n. 1, p. 143–162, 2012.

BENSON, D. A. et al. GenBank. **Nucleic Acids Research**, v. 36, n. Database, p. D25–D30, 23 dez. 2007.

BOISVERT, S.; LAVIOLETTE, F.; CORBEIL, J. Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. **Journal of Computational Biology**, v. 17, n. 11, p. 1401–1415, 1 nov. 2010.

BRAS, U. D. E.; BIOL, N. Etiologia , Epidemiologia e Fisiologia da Murcha de Curtobacterium Etiologia , Epidemiologia e Fisiologia da Murcha de Curtobacterium. 2010.

BROCKHURST, M. A. et al. The Ecology and Evolution of Pangenomes. **Current biology : CB**, v. 29, n. 20, p. R1094–R1103, 21 out. 2019.

BULGARI, D. et al. Restructuring of endophytic bacterial communities in grapevine yellows-diseased and recovered Vitis vinifera L. plants. **Applied and environmental microbiology**, v. 77, n. 14, p. 5018–22, 15 jul. 2011.

Celera's Approach. Disponível em: <https://www.ocf.berkeley.edu/~edy/genome/celera.html>. Acesso em: 24 out. 2019.

CHASE, A. B. et al. Evidence for Ecological Flexibility in the Cosmopolitan Genus Curtobacterium. **Frontiers in microbiology**, v. 7, p. 1874, 2016.

CLEVELAND, T. E. et al. Potential of Aspergillus flavus genomics for applications in biotechnology. **Trends in Biotechnology**, v. 27, n. 3, p. 151–157, 2009.

COIL, D.; JOSPIN, G.; DARLING, A. E. A5-miseq : an updated pipeline to assemble microbial genomes from Illumina MiSeq data Motivation : Results : Availability : Contact : **Bioinformatics**, p. 5–8, 2014.

COLLINS, M. D.; JONES, D.; SCHOFIELD, G. M. Reclassification of “Corynebacterium haemolyticum” (MacLean, Liebow & Rosenberg) in the Genus Arcanobacterium gen.nov. as Arcanobacterium haemolyticum nom.rev., comb.nov. **Microbiology**, v. 128, n. 6, p. 1279–1281, 1 jun. 1982.

COMPUTATIONAL PAN-GENOMICS CONSORTIUM. Computational pan-genomics: status, promises and challenges. **Briefings in bioinformatics**, v. 19, n. 1, p. 118–135, 2018.

CONTRERAS-MOREIRA, B. et al. Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. **Frontiers in Plant Science**, v. 8, 14 fev. 2017.

CONTRERAS-MOREIRA, B.; VINUESA, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. **Applied and Environmental Microbiology**, v. 79, n. 24, p. 7696–7701, dez. 2013.

COTTER, P. D.; ROSS, R. P.; HILL, C. **Bacteriocins-a viable alternative to antibiotics?Nature Reviews Microbiology**, fev. 2013.

Curtobacterium flaccumfaciens (Hedges, 1922) Collins & Jones, 1984. Disponível em: <https://www.gbif.org/pt/species/3225327/metrics>. Acesso em: 20 nov. 2019.

DE BOER, S. H.; SADDLER, G. E.; STEAD, D. E. **Names of Plant Pathogenic Bacteria, 1864-2003**. [s.l: s.n.]. Disponível em: <https://www.isppweb.org/Transfer_Files/Names of Plant Pathogenic Bacteria 2005.pdf>. Acesso em: 8 out. 2019.

DE VRIES, R. P. et al. **Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus**. [s.l: s.n.]. v. 18

DEES, M. W.; BRURBERG, M. B.; LYSOE, E. Complete genome sequence of the biofilm-forming Curtobacterium sp. strain BH-2-1-1, isolated from lettuce (Lactuca sativa) originating from a conventional field in Norway. **Genomics data**, v. 10, p. 135–136, dez. 2016.

DEES, M. W.; BRURBERG, M. B.; LYSØE, E. Complete genome sequence of the biofilm-forming Curtobacterium sp. strain BH-2-1-1, isolated from lettuce (Lactuca sativa) originating from a conventional field in Norway. **Genomics Data**, v. 10, p. 135–136, dez. 2016.

EBERHARDT, R. Y. et al. AntiFam: a tool to help identify spurious ORFs in protein annotation. **Database : the journal of biological databases and curation**, v. 2012, p. bas003, 2012.

EDDY, S. R. **Multiple alignment using hidden Markov models**. [s.l: s.n.]. Disponível em: <www.aaai.org>. Acesso em: 24 out. 2019.

ES, L. Initial impact of the sequencing of the human genome. **Nature**, v. 470, n. 7333, p. 187–197, 2011.

FIELDS, S. **Site-seeing by sequencingScience**, 8 jun. 2007.

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

FUNKE, G.; ARAVENA-ROMAN, M.; FRODL, R. First description of Curtobacterium spp. isolated from human clinical specimens. **Journal of clinical microbiology**, v. 43, n. 3, p. 1032–1036, mar. 2005.

GLASNER, J. D. et al. Niche-specificity and the variable fraction of the Pectobacterium pan-genome. **Molecular plant-microbe interactions : MPMI**, v. 21, n. 12, p. 1549–60, dez. 2008.

GOODFELLOW, M.; WILLIAMS, S. T. Ecology of Actinomycetes. **Annual Review of Microbiology**, v. 37, n. 1, p. 189–216, out. 1983.

GRIFFITHS-JONES, S. et al. **Rfam: An RNA family databaseNucleic Acids Research**, 1 jan. 2003.

HAGEN, J. B. The origins of bioinformatics. **Nature Reviews Genetics**, v. 1, n. 3, p. 231–236, dez. 2000.

HANDELSMAN, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. **Microbiology and Molecular Biology Reviews**, v. 68, n. 4, p. 669–685, 1 dez. 2004.

HOWARD, K. The Bioinformatics Gold Rush. **Scientific American**, v. 283, n. 1, p. 58–63, jul. 2000.

HYATT, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC Bioinformatics**, v. 11, p. 119, 8 mar. 2010.

ILLUMINA, Illumina sequencing introduction. **Illumina sequencing introduction**, n. October, p. 1–8, 2017.

ISHINO, Y. et al. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isoenzyme conversion in Escherichia coli, and identification of the gene product. **Journal of Bacteriology**, v. 169, n. 12, p. 5429–5433, 1987.

KANEHISA, M.; SATO, Y.; MORISHIMA, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. **Journal of Molecular Biology**, v. 428, n. 4, p. 726–731, 22 fev. 2016.

KIM, M. K. et al. Curtobacterium ginsengisoli sp. nov., isolated from soil of a ginseng field. **International journal of systematic and evolutionary microbiology**, v. 58, n. Pt 10, p. 2393–2397, out. 2008.

KREMER, F. S. et al. Genix: a new online automated pipeline for bacterial genome annotation. **FEMS Microbiology Letters**, v. 363, n. 23, p. fnw263, dez. 2016.

LACAVA, P. T. et al. The endophyte Curtobacterium flaccumfaciens reduces symptoms caused by Xylella fastidiosa in Catharanthus roseus. **Journal of microbiology (Seoul, Korea)**, v. 45, n. 5, p. 388–393, out. 2007.

LAGESEN, K. et al. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. **Nucleic Acids Research**, v. 35, n. 9, p. 3100–3108, maio 2007.

LASLETT, D.; CANBACK, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. **Nucleic acids research**, v. 32, n. 1, p. 11–6, 2004.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. **Genome Research**, v. 13, n. 9, p. 2178–2189, 1 set. 2003.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, v. 22, n. 13, p. 1658–1659, 1 jul. 2006.

LIN, S.-H.; LIAO, Y.-C. CISA: contig integrator for sequence assembly of bacterial genomes. **PloS one**, v. 8, n. 3, p. e60843, 2013.

LOCKHART, D. J.; WINZELER, E. A. Genomics, gene expression and DNA arrays [In Process Citation]. **Nature**, v. 405, n. 0028-0836 SB-M SB-X, p. 827–836, 2000.

LOWE, T. M.; EDDY, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. **Nucleic Acids Research**, v. 25, n. 5, p. 955–964, 1 mar. 1997.

MAHFOUZ, M. M.; LI, L. TALE nucleases and next generation GM crops. **GM crops**, v. 2, n. 2, p. 99–103, 2011.

MCINERNEY, M. J. et al. Physiology, Ecology, Phylogeny, and Genomics of Microorganisms Capable of Syntrophic Metabolism. **Annals of the New York Academy of Sciences**, v. 1125, n. 1, p. 58–72, 26 mar. 2008.

MEDEMA, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. **Nucleic Acids Research**, v. 39, n. suppl_2, p. W339–W346, 1 jul. 2011.

MENG, P. et al. Exploring the Genomic Diversity and Cariogenic Differences of Streptococcus mutans Strains Through Pan-Genome and Comparative Genome Analysis. **Current Microbiology**, v. 74, n. 10, p. 1200–1209, 1 out. 2017.

METZKER, M. L. Sequencing technologies — the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31–46, 8 jan. 2010.

MIETHKE, M.; MARAHIEL, M. A. Siderophore-Based Iron Acquisition and Pathogen Control. **MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS**, v. 71, n. 3, p. 413–451, 2007.

MOORE, R.; CHANDRAHAS, A.; BLERIS, L. Transcription activator-like effectors: A toolkit for synthetic biology. **ACS Synthetic Biology**, v. 3, n. 10, p. 708–716, 2014.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. **Genomics**, v. 92, n. 5, p. 255–64, nov. 2008.

MUZZI, A.; DONATI, C. **Population genetics and evolution of the pan-genome of Streptococcus pneumoniaeInternational Journal of Medical Microbiology**, dez. 2011a.

MUZZI, A.; DONATI, C. Population genetics and evolution of the pan-genome of Streptococcus pneumoniae. **International Journal of Medical Microbiology**, v. 301, n. 8, p. 619–622, dez. 2011b.

MUZZI, A.; DONATI, C. Population genetics and evolution of the pan-genome of Streptococcus pneumoniae. **International Journal of Medical Microbiology**, v. 301, n. 8, p. 619–622, dez. 2011c.

NAWROCKI, E. P.; KOLBE, D. L.; EDDY, S. R. Infernal 1.0: inference of RNA alignments. **Bioinformatics (Oxford, England)**, v. 25, n. 10, p. 1335–7, 15 maio 2009.

NEILANDS, J. B. **Siderophores: Structure and function of microbial iron transport compoundsJournal of Biological Chemistry**American Society for Biochemistry and Molecular Biology Inc., , 10 nov. 1995.

OSDAGHI, E. et al. Bacterial wilt of common bean (Phaseolus vulgaris) caused by Curtobacterium flaccumfaciens pv. flaccumfaciens in Iran. **Australasian Plant Disease Notes**, v. 10, n. 1, 22 dez. 2015.

PAREJA-TOBES, P. et al. BG7: A New Approach for Bacterial Genome Annotation Designed for Next Generation Sequencing Data. **PLoS ONE**, v. 7, n. 11, p. e49239, 21 nov. 2012.

RAUPACH, G. S.; KLOEPPER, J. W. Mixtures of plant growth-promoting rhizobacteria enhance biological control of multiple cucumber pathogens. **Phytopathology**, v. 88, n. 11, p. 1158–1164, 1998.

RAUPACH, G. S.; KLOEPPER, J. W. Biocontrol of cucumber diseases in the field by plant growth-promoting rhizobacteria with and without methyl bromide fumigation. **Plant Disease**, v. 84, n. 10, p. 1073–1075, 2000.

SALLET, E.; GOUZY, J.; SCHIEX, T. EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. **Bioinformatics (Oxford, England)**, v. 30, n. 18, p. 2659–61, 15 set. 2014.

SANDER, J. D.; JOUNG, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. **Nature Biotechnology**, v. 32, n. 4, p. 347–350, 2014.

SANGER, F.; COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **Journal of Molecular Biology**, v. 94, n. 3, 25 maio 1975.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, v. 30, n. 14, p. 2068–2069, 15 jul. 2014.

SHERF, A. F.; MACNAB, A. A. **Vegetable Diseases and Their Control, second edition**. A Willey-i ed. [s.l.] John Wiley & Sons, 1986.

SIMPSON, J. T. et al. ABySS: A parallel assembler for short read sequence data. **Genome Research**, v. 19, n. 6, p. 1117–1123, jun. 2009.

SOARES, R. M. et al. First report of Curtobacterium flaccumfaciens pv. flaccumfaciens on soybean in Brazil. **Tropical Plant Pathology**, v. 38, n. 5, p. 452–454, set. 2013.

Soja - Portal Embrapa. Disponível em: <https://www.embrapa.br/soja/cultivos/soja1>. Acesso em: 19 nov. 2019.

STAMATOYANNOPOULOS, J. A. The genomics of gene expression. **Genomics**, v. 84, n. 3, p. 449–57, set. 2004.

TATUSOV, R. L. et al. The COG database: An updated vesion includes eukaryotes. **BMC Bioinformatics**, v. 4, 11 set. 2003.

TATUSOVA, T. et al. NCBI prokaryotic genome annotation pipeline. **Nucleic Acids Research**, v. 44, n. 14, p. 6614–6624, 19 ago. 2016.

TETS, V. V. [Pangenome]. **Tsitologiia**, v. 45, n. 5, p. 526–31, 2003.

UDAONDO, Z. et al. Analysis of the core genome and pangenome of Pseudomonas putida. **Environmental microbiology**, v. 18, n. 10, p. 3268–3283, 2016.

VAN DIJK, E. L. et al. Ten years of next-generation sequencing technology. **Trends in genetics : TIG**, v. 30, n. 9, p. 418–26, 1 set. 2014.

VAN DOMSELAAR, G. H. et al. BASys: a web server for automated bacterial genome annotation. **Nucleic Acids Research**, v. 33, n. Web Server, p. W455–W459, 1 jul. 2005.

ZERBINO, D. R.; BIRNEY, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**, v. 18, n. 5, p. 821–829, maio 2008.

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

ZHAO, Y. et al. PanGP: A tool for quickly analyzing bacterial pan-genome profile. **Bioinformatics**, v. 30, n. 9, p. 1297–1299, 2014.

9. Anexos

Anexo A

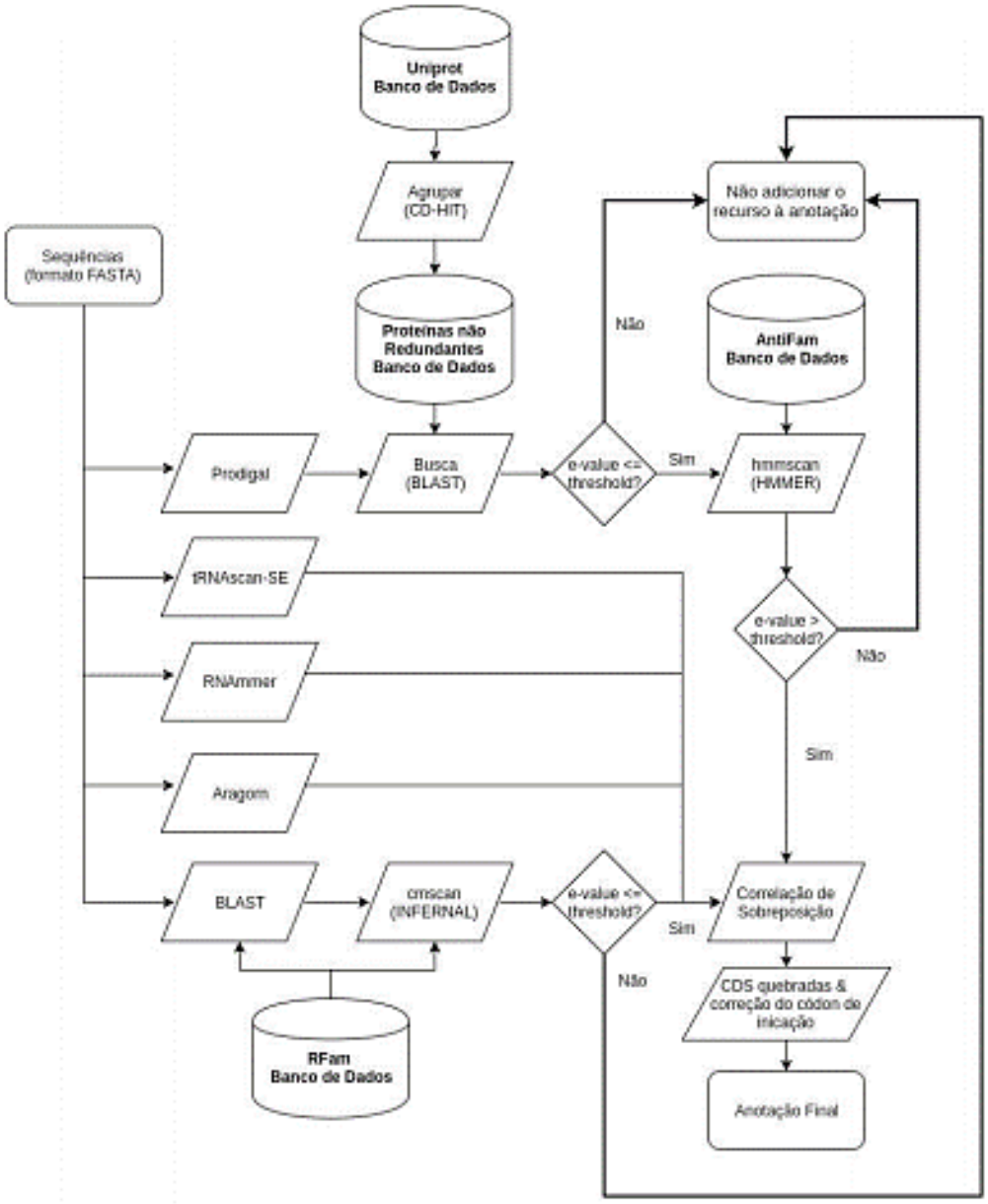

Figura 12. Pipeline do software GENIX. Adaptado de KREMER et al., 2016.

Anexo B

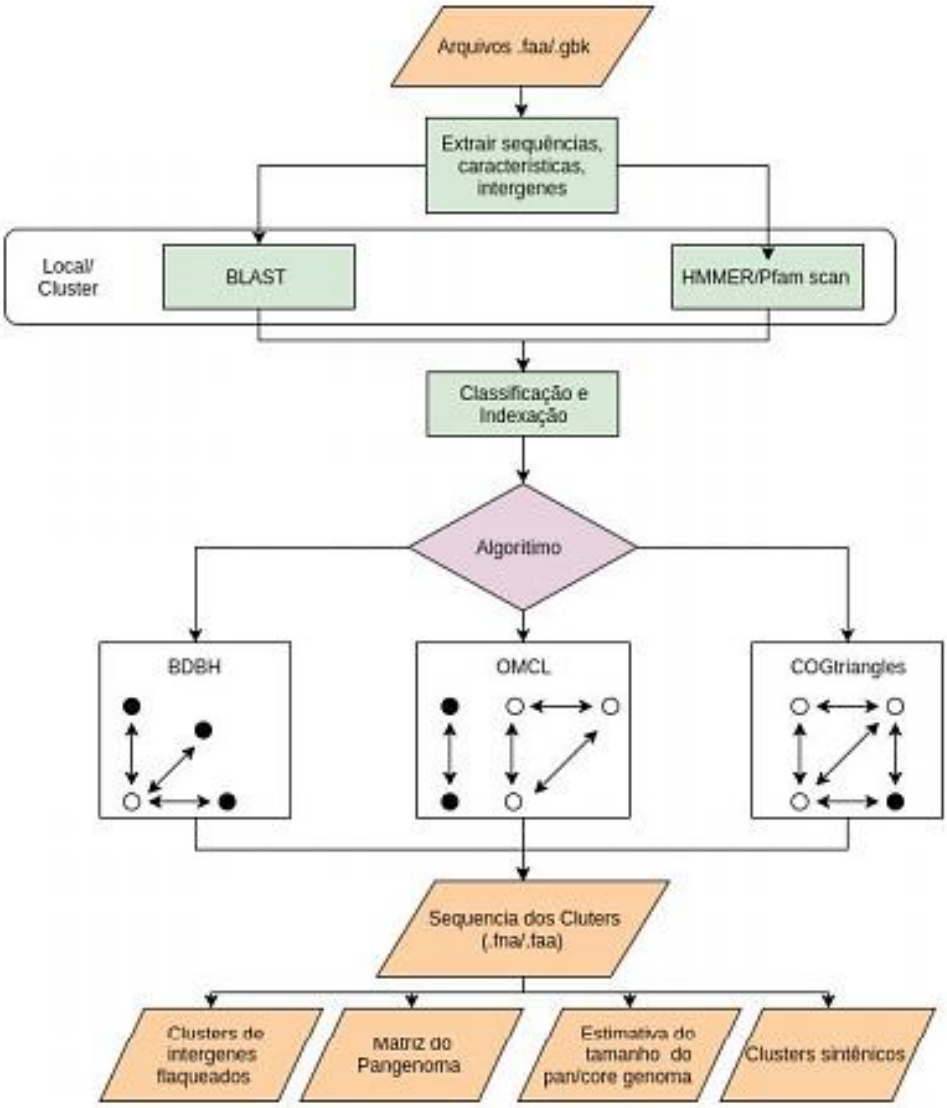

Figura 12. Visão geral do algoritimo do *software* GET_HOMOLOGUES. Adaptado de (CONTRERAS-MOREIRA; VINUESA, 2013).

ANEXO C

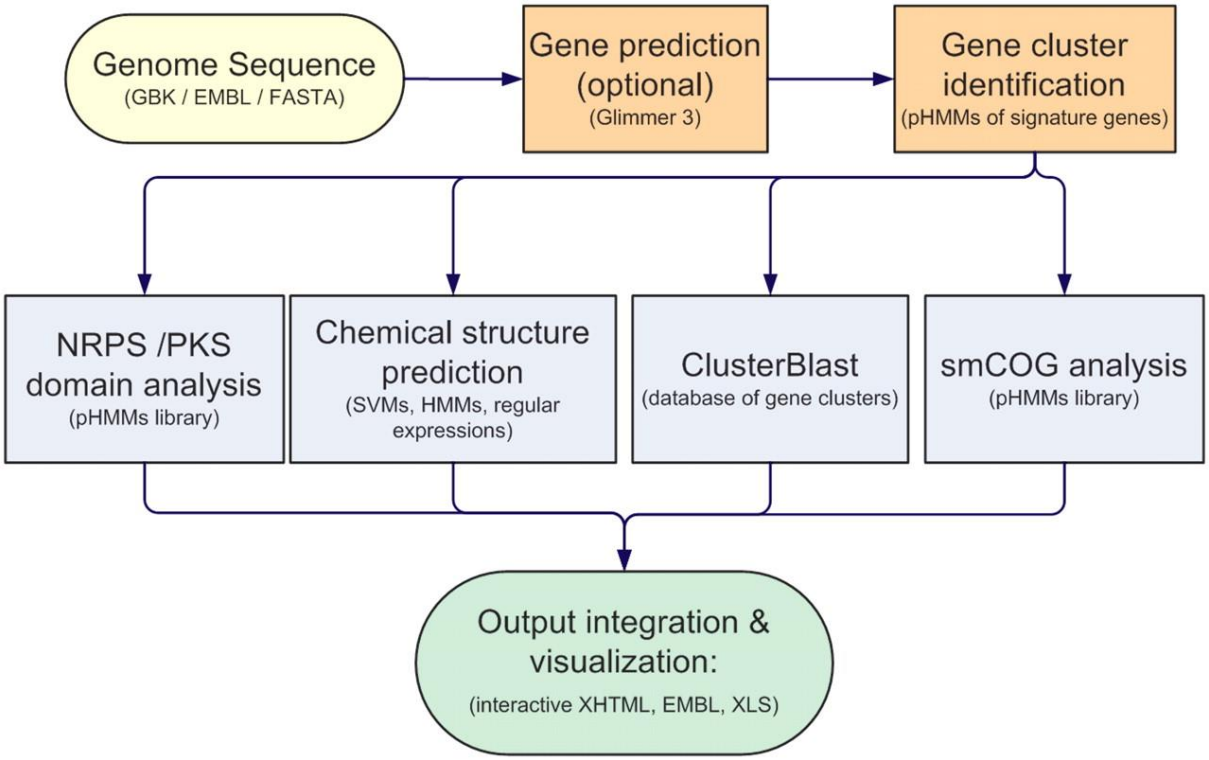

Figura 13. AntiSMASH. Esboço do pipeline para análise genômica de metabólitos secundários. Obtido de Kai Blin et al., (2013)