Universidade Federal de Uberlândia - UFU

Instituto de Biologia - InBio

Hugo Martins Correia

Prospecção e classificação de profagos em genomas de bactérias do gênero *Proteus*

Uberlândia, MG

2021

Hugo Martins Correia

Prospecção e classificação de profagos em genomas de bactérias do gênero *Proteus*

Trabalho de Conclusão de Curso apresentado ao Instituto de Biologia Ciências Biológicas como requisito para a obtenção do título de Licenciado em Ciências Biológicas, da Universidade Federal de Uberlândia.

Orientador: Prof. Dr. Flávio Tetsuo Sassaki

Coorientador: Ms. Richard Costa Polveiro

Uberlândia

2021

Prospecção e classificação de profagos em genomas de bactérias do gênero *Proteus*

Trabalho de Conclusão de Curso aprovado para a obtenção do título de Licenciado em Ciências Biológicas da Universidade Federal de Uberlândia pela banca examinadora formada por:

Uberlândia, 22 de outubro de 2021.

_____

Prof. Dr. Flávio Tetsuo Sassaki

_____

Profa. Dra. Maria Aparecida Scatamburlo Moreira

_____

Prof. Dr. Pedro Marcus Pereira Vidigal

Dedicatória

Dedico este trabalho a todas as crianças que um dia sonharam em ser cientistas e a todos os cientistas que as apoiam com seu trabalho. Nos apoiamos sobre o ombro de gigantes para vislumbrar e tentar construir um futuro melhor.

Agradecimentos

Agradeço primeiramente à minha mãe, Elielma, à minha avó, Ordalia, e à minha irmã, Brenda por todo o apoio que me deram ao longo da minha vida e por me motivarem a fazer o meu melhor. Sem vocês eu não estaria aqui!

Agradeço à minha madrinha, Daiane, ao meu padrinho, Wesley, à minha tia, Joice, e ao meu tio, Fábio, e toda a minha família.

Agradeço à Samantha, à Safira e ao Tico pela companhia e carinho.

Agradeço ao meu orientador, o professor Dr. Flávio Tetsuo Sassaki e ao meu coorientador, Ms. Richard Costa Polveiro, por me orientarem neste projeto e por todo o auxílio que me deram ao longo do desenvolvimento dele.

Agradeço também à professora Dra. Maria Aparecida Scatamburlo Moreira e ao professor Dr. Pedro Marcus Pereira Vidigal pela gentileza e disponibilidade de compor minha banca.

Agradeço a todos aqueles que direta ou indiretamente contribuíram para que esse trabalho fosse realizado.

Resumo

Bacteriófagos são vírus que infectam bactérias. Durante seu processo de replicação os fagos podem se integrar aos genomas bacterianos e com isso influenciarem o processo evolutivo das bactérias. Fagos estão presentes em diversos ambientes e desempenham várias funções relevantes nos diferentes ecossistemas, além de também estarem envolvidos em alguns processos de interesse da indústria médica como o uso de fagos como agentes antibacterianos. Uma das abordagens mais utilizadas para o estudo de bacteriófagos é através da sua prospecção em genomas bacterianos por meio de ferramentas computacionais de bioinformática. Entretanto, essa abordagem ainda tem limitações em relação a classificação taxonômica e de completude dos genomas virais encontrados. Bactérias do gênero *Proteus* são patógenos oportunistas que infectam o trato urinário e cateteres, podendo formar biofilme e entupir os equipamentos médicos. Apesar de já existirem tratamentos que usam bacteriófagos para mitigar a ocorrência de infecções e formação de biofilme, os genomas das bactérias do gênero *Proteus* carecem de análises sistemáticas e massivas em busca de profagos. Nesse trabalho, buscamos explorar genomas de bactérias do gênero *Proteus* em busca de profagos e classificá-los em relação taxonomia e completude, de forma sistemática *in silico*, automatizada e reprodutível. Para isso, analisamos genomas disponíveis no GenBank com o programa VirSorter2 e classificamos os elementos virais em relação à sua completude utilizando o programa CheckV. Buscamos classificar taxonomicamente com outros programas, os elementos virais que tiveram mais de 90% de completude utilizando abordagens baseadas no conteúdo proteico, com o vContact2 e o VipTree, baseadas na presença de genes no pangenoma, com o Roary, e na similaridade genômica, com o VIRIDIC. Foram encontrados 3907 elementos virais dos quais, após um processamento 1106 foram classificados como genomas completos ou de alta completude. Nosso resultado propõe a criação de duas novas subfamílias, *Bcepmuvirinae* e *Menelausvirinae*, 20 novos gêneros e 47 novas espécies virais dentro da família *Myoviridae*. A metodologia aqui empregada permite que em trabalhos futuros haja uma maior sistematização da verificação de completude e permite compararmos quantitativamente futuras metodologias.

Palavras-chave: Bacteriófagos. Bioinformática. Taxonomia. Genética de microorganismos.

Abstract

Bacteriophages are viruses that infect bacteria. During their replication process phages can be integrated with bacterial genomes and thus influence the bacterial evolutionary process. Phages are present in different environments and play several relevant roles in different ecosystems, in addition to being involved in some processes of interest to the medical industry, such as the use of phages as antibacterial agents. One of the most used approaches for the study of bacteriophages is through its prospection in bacterial genomes, through computational bioinformatics tools. However, this approach still has limitations regarding a taxonomic and complete classification of the genomes found. *Proteus* bacteria are opportunistic pathogens that infect the urinary tract and catheters, and can even form biofilm and clog medical equipment. Although treatments that use bacteriophages already exist to mitigate the occurrence of proliferation and biofilm formation, the genomes of bacteria of the the genomes of bacteria of the genus Proteus have great need of systematically and massively search of prophages. In this work, we seek to explore genomes of bacteria of the *Proteus* genus in search of prophages and classify them in relation to taxonomy and completeness, in a systematic, automated and reproducible way. For this, we analyzed the genomes available in GenBank with VirSorter2, we classified the viral elements in relation to their completion using CheckV. We sought to taxonomically classify viral elements that have more than 90% exploited completeness based on protein content, with vContact2 and VipTree; based on the presence of genes in the pangenome, with Roary; and on genomic similarity, with VIRIDIC. 3907 viral elements were found, of which, after a trimming process, 1106 were classified as complete or high-quality genomes. Our result proposes the creation of two new subfamilies, *Bcepmuvirinae* and *Menelausvirinae*, 20 new genera and 47 new viral species within the *Myoviridae* family. The methodology used allows for a greater systematization of the completeness check in future works and allows us to quantitatively compare future methodologies.

Keywords: Bacteriophages. Bioinformatics. Taxonomy. Microorganism genetics.

Sumário

Introdução

Os Vírus são o grupo taxonômico com maior diversidade do planeta (BREITBART; ROHWER, 2005). De fato, é estimado que existem 10^{31} partículas virais no planeta Terra (Bar-On et al., 2018; MUSHEGIAN, 2020). Apesar da sua grande diversidade, nós ainda conhecemos muito pouco sobre essas entidades, tanto que grande parte das sequências que estão disponíveis em bancos de dados ainda não foram caracterizadas (BRISTER et al., 2014).

Bacteriófagos são vírus que infectam bactérias e, como todos os outros vírus, eles precisam da maquinaria celular de um hospedeiro para que possam se replicar (OFIR; SOREK, 2018). A existência dos fagos foi descoberta, de forma independente, por Frederick W. Twort em 1915, enquanto tentava cultivar o *Vaccinia virus* em meio de ágar na ausência de células vivas, e Félix d'Herelle em 1917, ao investigar um surto de disenteria bacilar. Em meio às suas pesquisas, ambos observaram a ocorrência de lise em colônias bacterianas de onde supuseram a existência dos bacteriófagos (KUTTER; SULAKVELIDZE, 2004; VANDAMME; MORTELMANS, 2018).

A replicação dos fagos se dá após uma colisão ao acaso entre o bacteriófago e a bactéria, assim a partícula viral se liga a um receptor celular localizado na membrana celular da bactéria, um processo denominado adsorção. Após a adsorção, o fago insere seu material genético na bactéria e passa a utilizar o maquinário celular bacteriano para produção de mais material genético e proteínas do fago (TORTORA; FUNKE; CASE, 2012).

Há três tipos de ciclos reprodutivos: o ciclo lítico, o lisogênico e o crônico. No ciclo lítico os ácidos nucleicos se mantêm separados do DNA bacteriano, como moléculas flutuantes livres, e ao serem transcritos passam a sintetizar novos componentes virais e DNA. O material replicado se organiza espontaneamente no processo de maturação formando novos vírions. Ao final do ciclo as novas partículas são liberadas da célula causando a lise bacteriana (TORTORA; FUNKE; CASE, 2012). No ciclo lisogênico, ocorre a inserção e integração do material genético do vírus ao cromossomo bacteriano, e então o DNA do fago passa a ser chamado de profago. Durante a reprodução bacteriana o profago também será replicado e transmitido à nova geração onde permanecerá latente, pela ação de proteínas repressoras codificadas pelo genoma do fago. Porém, caso ocorra a indução do genoma devido à ação de substâncias químicas, radiação ultravioleta ou mesmo eventos espontâneos raros, o profago poderá entrar em ciclo lítico, formando novas partículas virais e causando a lise bacteriana

(TORTORA; FUNKE; CASE, 2012). No ciclo crônico há uma produção de partículas virais as quais são liberadas durante longos intervalos sem causar o rompimento significativo das células hospedeiras. Dessa forma, nesse ciclo os fagos são produzidos e liberados, sem causar a lise bacteriana (HOBBS; ABEDON, 2016). Uma espécie de fago pode apresentar apenas um ou até todos esses ciclos (HOBBS; ABEDON, 2016).

Várias análises de profagos mostraram que eles são encontrados predominantemente em loci mais próximos de genes não codificantes (por exemplo, genes de tRNA), de alguns genes funcionais e de regiões intergênicas. Essas regiões são chamadas de *hotspots* de integração. Nas regiões dos *hotspots* de integração se assume que a chance de ocorrer uma recombinação deletéria é menor, portanto, os eventos de integração que ali ocorrem possivelmente são mais toleráveis ou vantajosos para o bacteriófago (RAMISETTY; SUDHAKARI, 2019).

Enquanto estão integrados ao genoma bacteriano os profagos podem contribuir positivamente ou negativamente para o *fitness* bacteriano, como ao introduzir fatores de virulência (VF) ou inutilizar genes durante o processo de integração, promovendo dessa forma uma conversão lisogênica (BOYD et al., 2002). Também é possível que fagos temperados evitem uma superinfecção por fagos filogeneticamente relacionados (defesa homotípica) (PETROVA; BROUSSARD; HATFULL, 2015) ou, por meio de proteínas dedicadas, por fagos não filogeneticamente relacionados (defesa heterotípica) (DEDRICK et al., 2017; KOONIN et al., 2020).

Fagos impõem uma grande pressão evolutiva sobre as bactérias, ao mesmo tempo em que os vírus são selecionados na forma que infectam e se replicam no hospedeiro. Diante disso, o hospedeiro é selecionado pela eficiência de seus mecanismos em combatê-los, de forma que ocorre uma "corrida armamentista" entre ambos, o que culmina no desenvolvimento de mecanismos como o sistema de defesa CRISPR, o sistema de restrição modificação e o processo de infecção abortiva (DY et al., 2014; STERN; SOREK, 2011). Essa interação ainda possui outras implicações ambientais e médicas sobre a ciclagem global de nutrientes (SUTTLE, 2007), o clima global (FUHRMAN, 1999; SUTTLE, 2007), a evolução da biosfera (COMEAU; KRISCH, 2005) e sobre a virulência de microrganismos patogênicos (BRÜSSOW; CANCHAYA; HARDT, 2004).

Há ainda aplicações industriais para os fagos, como as técnicas desenvolvidas para a prevenção e remediação de colonizações bacterianas, que consistem na inserção e

revestimento prévio de superfícies com fagos (MELO et al., 2016; NZAKIZWANAYO et al., 2015). Essas estratégias incluem desde abordagens experimentais até outras já estabelecidas, oferecidas comercialmente e utilizadas em hospitais, como para evitar a ocorrência de bactérias do gênero *Proteus* em cateteres hospitalares (BERNASCONI et al., 2017; VIERTEL; RITTER; HORZ, 2014), devido a capacidade desse gênero de se estabelecer em cateteres e de causar o entupimento do equipamento. Este entupimento é decorrente da sua capacidade de formar biofilme tanto em superfícies artificiais, como no trato urinário e em outros tecidos (ARMBRUSTER; MOBLEY, 2012).

Proteus é um gênero de Proteobactérias Gram-negativas da família *Morganellaceae* (ADEOLU et al., 2016) e contém as espécies: *P. alimentorum*, *P. cibarius* (HYUN et al., 2016), *P. cibi*, *P. columbae*, *P. faecis*, *P. hauserii*, *P. mirabilis*, *P. penneri*, *P. terrae* e *P. vulgaris* (DAI et al., 2018a, 2018b, 2019; HYUN et al., 2016; O'HARA et al., 2000; SNEATH; MCGOWAN; SKERMAN, 1980). As espécies desse gênero podem ser encontradas em vários ambientes desempenhando diversos papéis ecológicos (DRZEWIECKA, 2016). Quatro delas podem provocar doenças, são elas: *P. mirabilis*, *P. vulgaris*, *P. penneri* e *P. hauserii*. Dentre elas, *P. mirabilis* é a terceira causa mais comum de infecções complicadas do trato urinário (12%) e a segunda causa mais comum de bacteriúria associada a um cateter em pacientes cateterizados por longo prazo (JACOBSEN et al., 2008), podendo também causar intoxicação alimentar, infecções respiratórias e de feridas, bacteremia e outras infecções. Doenças associadas à infecção por *P. mirabilis* também foram relatadas em aves com falha reprodutiva, macacos e furões com diarreia e cães com otite crônica (SUN et al., 2020). *P. hauserii* também pode provocar a morte de peixes ornamentais (KUMAR et al., 2015). Assim, *Proteus* spp. são consideradas patógenos oportunísticos, sendo relatadas como causadoras de infecções de trato urinário, bacteremia, meningoencefalite neonatal, empiema e osteomielite (O'HARA; BRENNER; MILLER, 2000).

Uma das abordagens possíveis para estudarmos bacteriófagos é através da prospecção de seus genomas nos genomas bacterianos já sequenciados (HATFULL; HENDRIX, 2011). Apesar dessa abordagem ser bastante eficiente, ela possui algumas limitações, uma delas é a dificuldade de definir se um genoma detectado é um genoma completo ou apenas um fragmento do genoma de um fago. Buscando mitigar este problema, diversos autores realizam curadorias manuais em seus genomas a fim de verificar se eles possuem todas as proteínas

essenciais para a formação de uma partícula viral, porém nem sempre são descritos os critérios utilizados neste processo (ZRELOVS; DISLERS; KAZAKS, 2020).

Nosso conhecimento sobre a taxonomia e sobre a forma com que os fagos são classificados continua aumentando, evidenciando sua grande importância. A forma com que classificamos passou de uma classificação baseada na estrutura do capsídeo viral por meio de eletromicroscopia (ACKERMANN; PRANGISHVILI, 2012), para uma que também engloba análises de proteínas e do genoma (ROUX et al., 2018). Ainda existem muitos desafios que enfrentamos para termos uma classificação filogenética dos vírus em grupos monofiléticos, alguns motivos para isso são devido à natureza mosaica dos genomas virais (ACKERMANN, 2011).

Nos genomas virais ocorrem interações e alterações no material genético constantemente, e esse processo permite a sobrevivência de linhagens, aumenta a sua diversidade e faz com que as pressões seletivas passem a atuar em nível de módulos no genoma (BOTSTEIN, 1980). No entanto, a intensa troca de informações genéticas dificulta a definição de ancestralidade, uma vez que vírus não relacionados filogeneticamente podem apresentar genomas semelhantes (ACKERMANN, 2011).

Por mais que existam aplicações para alguns dos fagos conhecidos que infectam as bactérias do gênero Proteus, ainda existe a necessidade de explorações mais profundas, sistemáticas e massivas nos genomas delas com técnicas atuais, que identificam profagos mais precisamente (BERNASCONI et al., 2017; VIERTEL; RITTER; HORZ, 2014). Neste trabalho, buscamos explorar os genomas em busca de profagos e classificá-los em relação a sua completude e taxonomia, de forma sistemática, automatizada e reproduzível.

Metodologia

O tratamento e análise dos dados fundamentaram-se nos pressupostos de pesquisa *in silico* e qualitativa de cunho exploratório experimental. Os dados utilizados para a condução desse projeto foram compostos por genomas completos de espécies do gênero *Proteus* provenientes do banco de dados GenBank do NCBI (National Center for Biotechnology Information) disponibilizados até o dia 5 de abril de 2021. Diante disso, 756 genomas de

espécies do gênero *Proteus* foram explorados, com uma quantidade diferente de genomas para cada espécie, de acordo com a tabela abaixo:

Tabela 1 - Número de genomas bacterianos por espécie

Espécie	Nº de genomas
P. alimentorum	3
P. cibi	1
P. columbae	2
P. faecis	2
P. genomosp.	3
P. hauseri	5
P. mirabilis	628
P. myxofaciens	1
P. penneri	8
P. sp.	59
P. terrae	20
P. vulgaris	24
Total	756

Busca de bacteriófagos integrados aos genomas

Para analisarmos a existência de profagos nos genomas bacterianos utilizamos o programa VirSorter2 (GUO et al., 2021) na versão 2.2.1 (Fig. 1). Os parâmetros utilizados exigiram que as sequências possuíssem ao menos um gene viral e no mínimo 10 mil pares de bases, por conta da dificuldade de diferenciar profagos com menos de 10 mil pares de base de outros elementos integrativos dos genomas bacterianos (BOBAY; ROCHA; TOUCHON, 2013; CANCHAYA et al., 2003; CASJENS, 2003). As sequências com menos de dois genes foram excluídas, exceto quando se tratava de genes característicos de vírus (hallmark genes), e foram mantidas as que atingiram escores de alta confiança, escore maior ou igual a 90% ou maior ou igual a 70% com ao menos um gene característico de vírus.

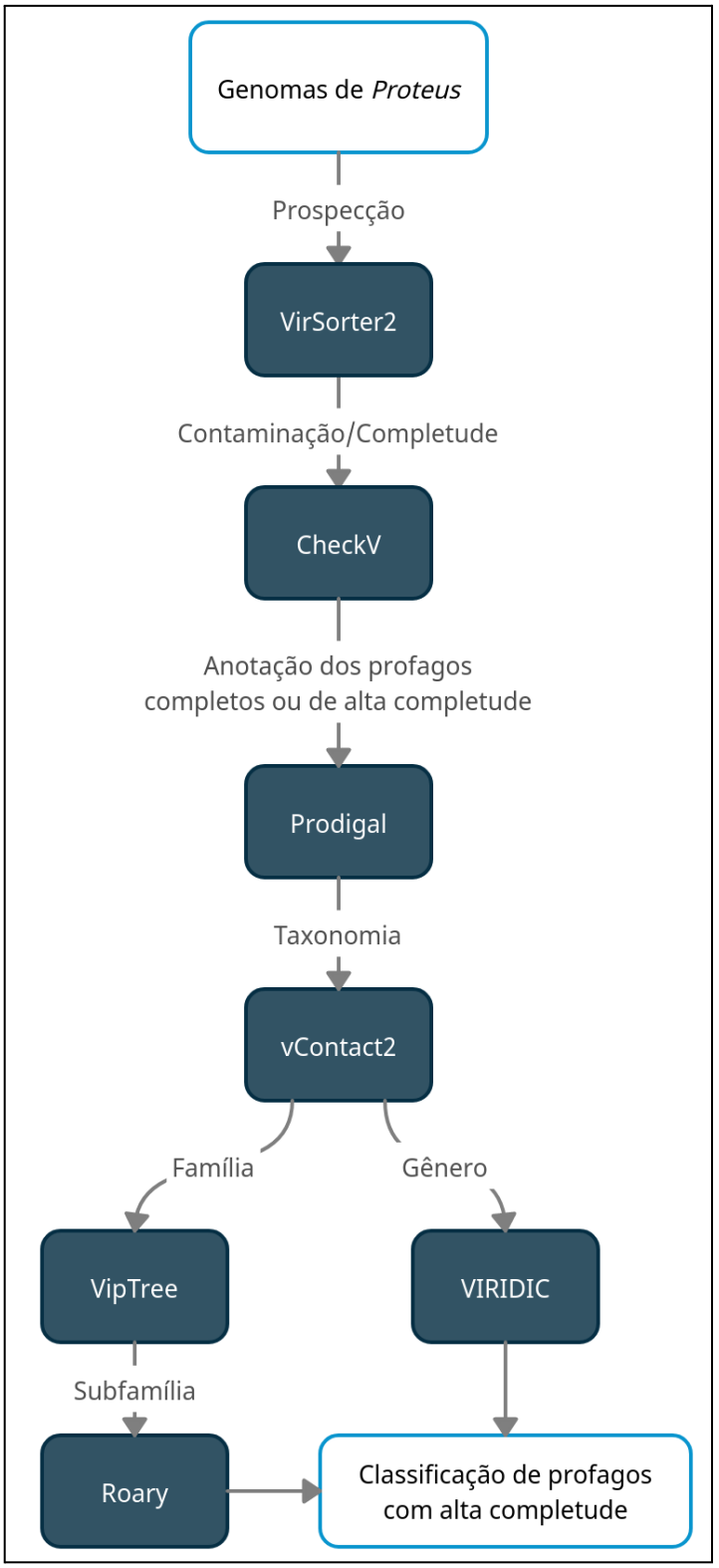

Figura 1 - Fluxograma dos programas utilizados ao longo deste trabalho.

Análise de contaminação e completude dos genomas virais

Ao prospectarmos profagos a partir de genomas bacterianos é possível que partes do genoma da bactéria sejam extraídas com os profagos. Para evitar que isso afete os resultados da classificação taxonômica e de completude e dos profagos, nós avaliamos a presença dessa contaminação nos genomas utilizando o programa CheckV (NAYFACH et al., 2020) na versão 0.7.0, com os parâmetros padrões (Fig. 1).

Para o CheckV avaliar a contaminação nos profagos, os genes são primeiro anotados como virais ou microbianos com base na comparação com um banco de dados personalizado de modelos ocultos de Markov, do inglês *Hidden Markov Models* (HMMs). Em seguida, é realizada uma varredura sobre a sequência comparando anotações de genes e conteúdo de GC entre um par de genes adjacentes. Esta informação é usada para calcular uma pontuação para cada posição intergênica e identificar os pontos que separam o que pertence ao vírus ou ao hospedeiro. Esses pontos são definidos por (NAYFACH et al., 2020):

- Uma pontuação maior que 1,2.
- Um mínimo de 2 genes específicos do hospedeiro na região putativa do hospedeiro, para sequências com 10 genes ou mais.
- Um mínimo de 2 genes virais na região viral putativa, sequências com 10 genes ou mais.
- Um mínimo de 30% dos genes anotados como microbianos na região putativa do hospedeiro.

Dessa forma, foram removidas das sequências as regiões de origem bacteriana e mantidas as regiões de origem viral em um novo arquivo e usando o CheckV, os novos arquivos foram analisados no aspecto da completude genômica (LI et al., 2021, ROUX et al., 2021), a qual geralmente é avaliada manualmente pela verificação da existência de genes estruturais fundamentais para a formação de uma partícula viral, como os genes de capsídeo (ZRELOVS; DISLERS; KAZAKS, 2020). Após o tratamento, mantivemos somente os genomas classificados como completos ou de alta completude, com completude estimada em mais de 90%, garantindo uma classificação mais precisa e um maior poder de inferência taxonômica ao nível de espécie (ROUX et al., 2018).

Atribuição de classificação taxonômica aos genomas virais

Buscando atribuir uma classificação taxonômica aos genomas virais, seguimos a metodologia proposta por Turner, Kropinski e Adriaenssens (2021). Para isso anotamos os genomas utilizando o programa Prodigal (HYATT et al., 2010) e agrupamos os genomas com o programa vContact2 (BOLDUC et al., 2017; JANG et al., 2019) (Fig. 1). Analisando as proteínas dos genomas e verificando quão similares elas são em relação às proteínas dos vírus da sua database e dos outros vírus que estão sendo analisados, o vContact2 agrupa e classifica os genomas montando os grupos virais. Portanto, a completude dos genomas e a qualidade da base de dados que o programa utiliza são elementos fundamentais para a obtenção de uma boa classificação.

Com o intuito de obtermos uma classificação mais precisa e atual, atualizamos a base de dados do programa vContact2 e incluímos nela somente genomas de espécies de bacteriófagos aceitas pelo Comitê Internacional de Taxonomia de Vírus (ICTV). Além disso, foram adicionados apenas bacteriófagos que possuem genomas completos e não fragmentados, disponibilizados no GenBank. Para isso foi feitas *pipelines* em R que utilizaram os pacotes stringr (WICKHAM, 2010), seqinr (CHARIF; LOBRY, 2007), Tidyverse (WICKHAM,2014; WICKHAM et al., 2019), reshape2 (WICKHAM, 2012), rjson (COUTURE-BEIL, 2018) e ggplot2 (WICKHAM, 2011). Essa *pipeline* nos possibilitou a construção de uma tabela contendo a classificação viral nos níveis de ordem, família, subfamília e gênero de acordo com a *Master Species List*, liberada em 18 de Maio de 2021 pelo ICTV, e agrupar as anotações de cada vírus disponíveis no Genbank na nova base de dados, garantindo assim sua replicabilidade, ela será disponibilizada na plataforma do GitHub no endereço https://github.com/martinshugoc/vContact2_taxonomy.

Assumimos que a classificação dos genomas que se agruparam com referências no vContact2 é equivalente à classificação do táxon comum a todas as referências com as quais agruparam (BOLDUC et al., 2017; JANG et al., 2019). Para sabermos onde os genomas se agrupavam dentro do táxon no qual foram classificados, utilizamos o programa VipTree com seus parâmetros padrões (NISHIMURA et al., 2017) (Fig. 1).

O ViPTree produz uma "árvore proteômica" (ROHWER; EDWARDS, 2002) baseadas em similaridades de sequência dos o genomas virais que são calculadas pelo tBLASTx. A “árvore

proteômica" é um dendrograma que representa as relações de similaridade genética entre os vírus. Foi demonstrado que os grupos virais identificados em uma "árvore proteômica" correspondem corretamente às taxonomias virais estabelecidas (NISHIMURA et al., 2017). Assim, para melhor compreensão taxonômica dos dados, os genomas foram analisados juntamente com todos os outros genomas completos não fragmentados de espécies aceitas pelo ICTV que pertencem ao mesmo táxon. Dessa forma obtivemos uma "árvore proteômica" montada de acordo com análises de semelhança das proteínas traduzidas.

Para definir as subfamílias dos bacteriófagos utilizamos o programa Roary (PAGE et al., 2015) para buscar *core genes* no pangenoma dos fagos (Fig. 1). O Roary classifica os genes em quatro classes de acordo com sua presença nos genomas analisados. *Core genes* são genes encontrados ao menos 99% dos genomas, os *soft-core genes* são genes encontrados entre 95 e 99% dos genomas, os *shell genes* são encontrados entre 15% e 95% dos genomas, enquanto os *cloud genes* são os genes que estão presentes em menos de 15% dos genomas (LIVINGSTONE; MORPHEW; WHITWORTH, 2018).

Para analisarmos o pangenoma dos fagos traduzimos as ORFs (*open read frames*) dos genomas utilizando o programa Prokka (SEEMANN, 2014), configurado para vírus, e obtivemos os arquivos no formato GFF3 com os quais analisamos os genes virais usando o programa Roary (PAGE et al., 2015). Originalmente o programa Roary foi desenhado para analisar o pangenoma de procariotos. No entanto, para a aplicabilidade que necessitamos e levando em consideração a maior variação genômica presente em genomas virais em relação aos genomas de procariotos, nós reduzimos o parâmetro mínimo de identidade de 95% para 40% (WANG; DUNBRACK, 2003).

Para identificarmos o gênero e a espécie dos bacteriófagos, nós realizamos uma análise de semelhança genômica com toda a família na qual eles foram inseridos com o programa VIRIDIC (MORARU; VARSANI; KROPINSKI, 2020) (Fig. 1). Taxonomicamente, para a entidade viral ser considerado membro de um gênero já existente, a sequência precisa compartilhar mais de 70% de semelhança com todas as sequências de determinado gênero da mesma família. Além disso, para ser considerada da mesma espécie, é necessário que as sequências compartilhem mais de 95% de semelhança com alguma espécie da mesma família (MORARU; VARSANI; KROPINSKI, 2020).

Tabelas e representações gráficas

As representações gráficas são montadas pelos próprios programas que realizam as análises ou foram montadas utilizando código escrito em R (Team R, 2013) para este fim, usando os pacotes stringr (WICKHAM, 2010), ggplot (WICKHAM, 2011) e gridExtra (AUGUIE; ANTONOV; AUGUIE, 2017). As tabelas foram montadas utilizando o LibreOffice ou o próprio R.

Resultados

Todos os 756 genomas completos de bactérias do gênero *Proteus* foram analisados utilizando o programa VirSorter2 e desses, somente a sequência ASM102812v1 apresentou inconsistências nas análises devido a sua alta fragmentação. Diante disso, observamos que a maioria das sequências de ASM102812v1 não atingiram o tamanho mínimo de 10 mil pares de base para serem analisadas pelo programa. Os resultados obtidos indicaram um total de 3907 sequências de provável origem viral, sendo todas classificadas como DNA de fita dupla (dsDNA). A espécie que apresentou mais profagos foi *P. columbae*, com 8,5 sequências por genoma (Fig. 2), já o grupo que apresentou menos profagos por genoma foi o das genomospécies de *Proteus*.

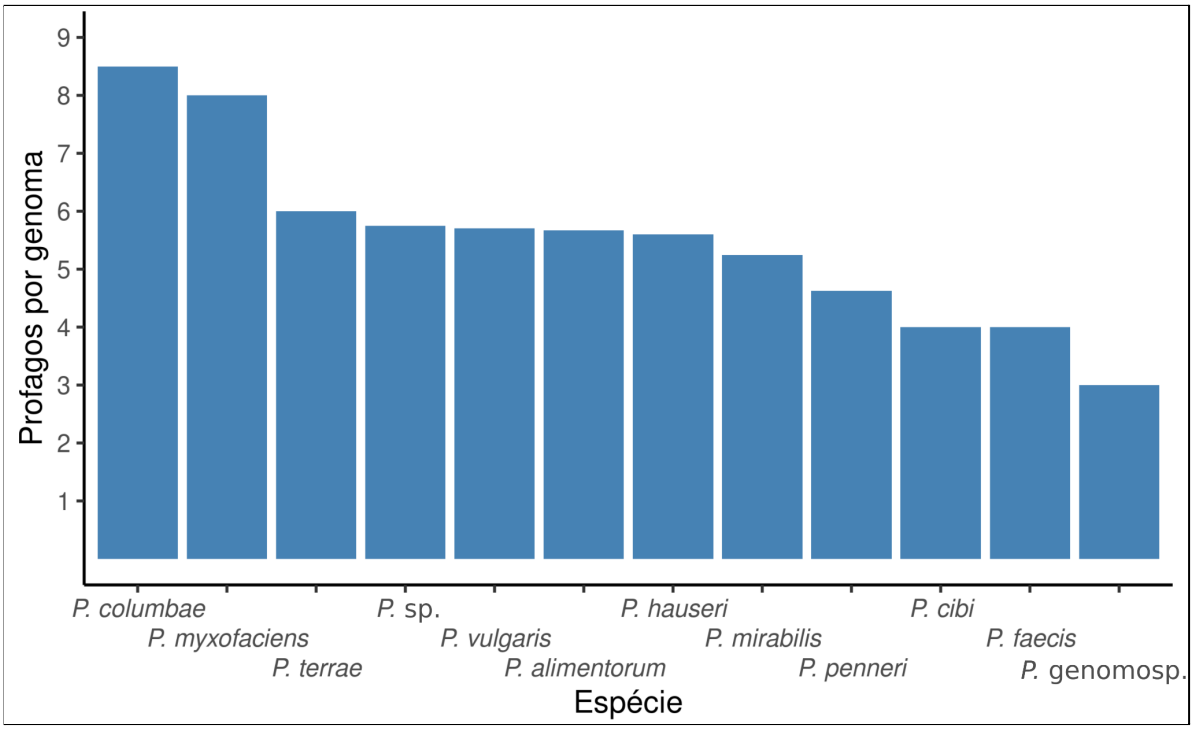

Figura 2 - Média de profagos por genoma das espécies de *Proteus*. *P. sp.* se refere a aqueles genomas que não foram identificados a nível de espécie.

O programa CheckV analisou os profagos extraídos pelo VirSorter2 antes da remoção dos fragmentos bacterianos e as classificou de acordo com sua completude. O resultado, representado na figura 3, indicou que 14,03% (n=548) das sequências foram classificadas como completas, 15,43% (n=604) como sequências de alta completude, 25,95% (n=1014) como sequências de média completude e 44,56% (n=1741) como sequências de baixa completude. A maior parte das sequências são de baixa completude, se tratando de fragmentos de fagos integrados aos genomas bacterianos que podem não possuir mais a capacidade de produzir partículas virais. Todavia, ainda há uma quantidade considerável de genomas completos ou de alta completude nestes genomas, que podem ser capazes de produzir vírus com capacidade de se replicarem.

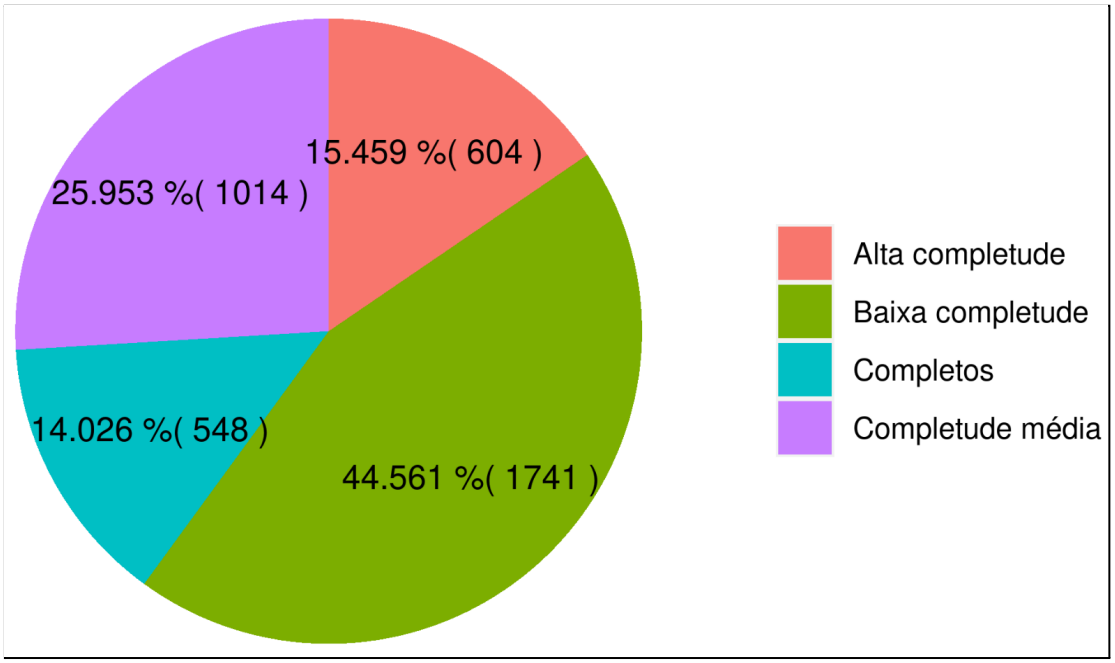

Figura 3 - Classificação da completude de profagos pelo programa CheckV aproximada em porcentagem. Entre parênteses está representado o número de sequências.

Após o processamento realizado pelo programa CheckV, foram gerados um total de 4015 profagos. As sequências geradas no processo foram reclassificadas em relação à sua completude (Fig. 4) e obtivemos que: 0,12% (n=5) são completas, 27,42% (n=1101) de alta completude, 28,39% (n=1140) de completude média e 44,06% (n=1769) de baixa completude. Isso indica que muitos profagos previamente classificados como completos, continham fragmentos de material genético proveniente do hospedeiro (bactérias) e após o processamento passaram a ter outras classificações. Possivelmente, o CheckV removeu fragmentos pequenos de profagos no processo ou, anteriormente a filtragem, genes do hospedeiro estavam interferindo na avaliação da completude viral.

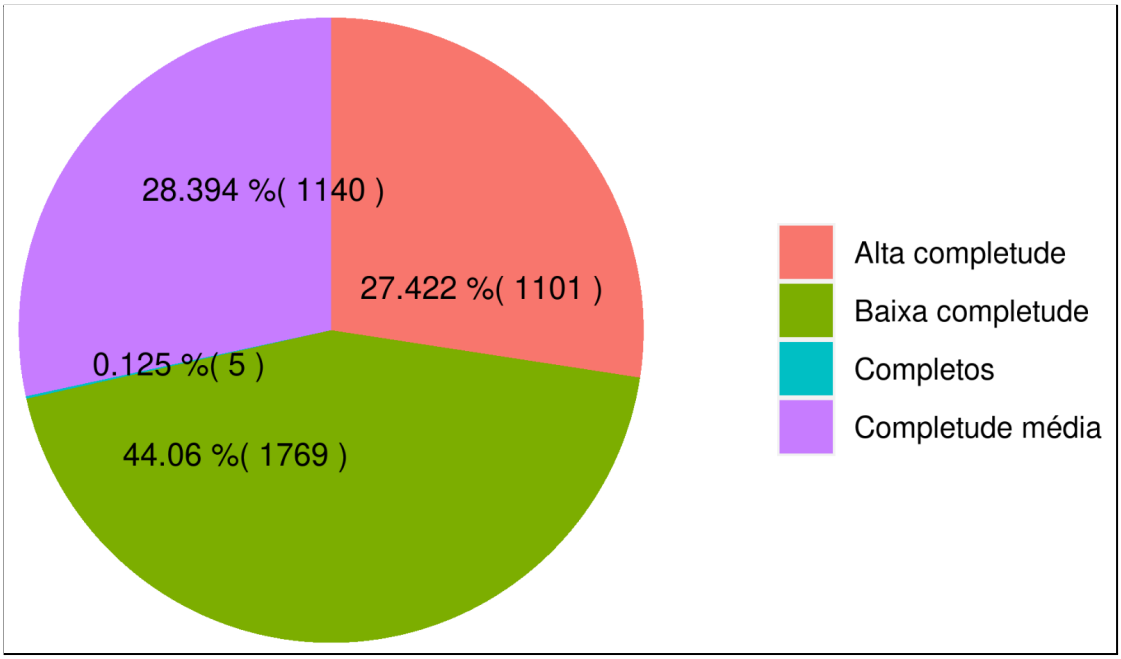

Figura 4 - Classificação da completude aproximada de profagos pelo programa CheckV, após o processamento, em porcentagem. Entre parênteses está representado o número de sequências.

O processo de remoção da contaminação foi feito sobre todos os profagos que foram encontrados neste estudo. Com isso alguns profagos de baixa completude passaram a ter uma completude maior, nas espécies de bactérias *P. columbae* e *P. myxofaciens* (Fig. 5 e 6), revelando que o processo também pode filtrar as sequências fazendo com que ocorra a melhora da classificação de alguns profagos. Tendo em vista que houve uma diminuição do número de genomas virais completos encontrados na primeira análise, vários destes genomas apresentavam partes do genoma do hospedeiro em si. Como essa diminuição ocorreu em todas as espécies bacterianas, existe a possibilidade de que não há relação entre a existência de contaminação com a espécie do hospedeiro. Por exemplo, aqui nós analisamos um genoma de *P. cibi* (Tabela 1). Antes da remoção dos fragmentos bacterianos havia um genoma viral completo, um de alta completude, um de completude média e um de baixa completude (Fig. 4 e Tabela Suplementar 1). Após o processamento, encontramos três genomas virais de alta completude e apenas um de baixa completude (Fig. 5 e Tabela Suplementar 1). O genoma viral que era completo possuía contaminação de sequências de nucleotídeos de origem bacteriana, que ao ser removida fez com que sua completude fosse reduzida, passando a ser classificado como de alta completude.

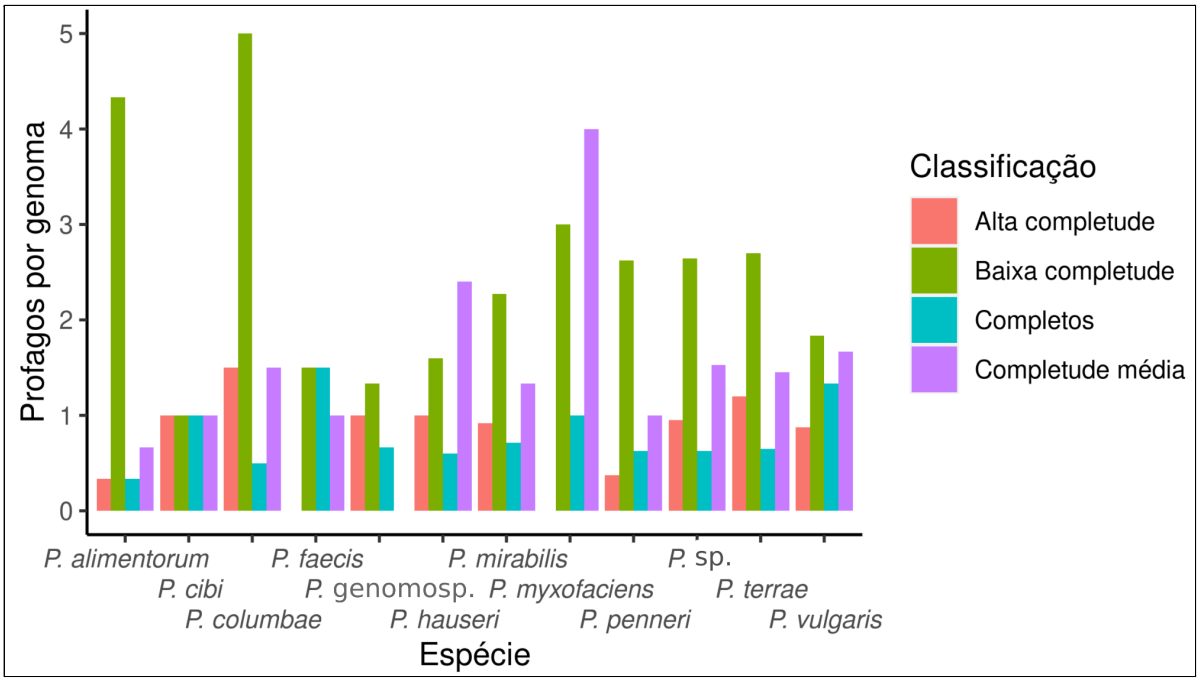

Figura 5 - Classificação da completude de profagos pelo programa CheckV por genoma das espécies de *Proteus* pré-processamento.

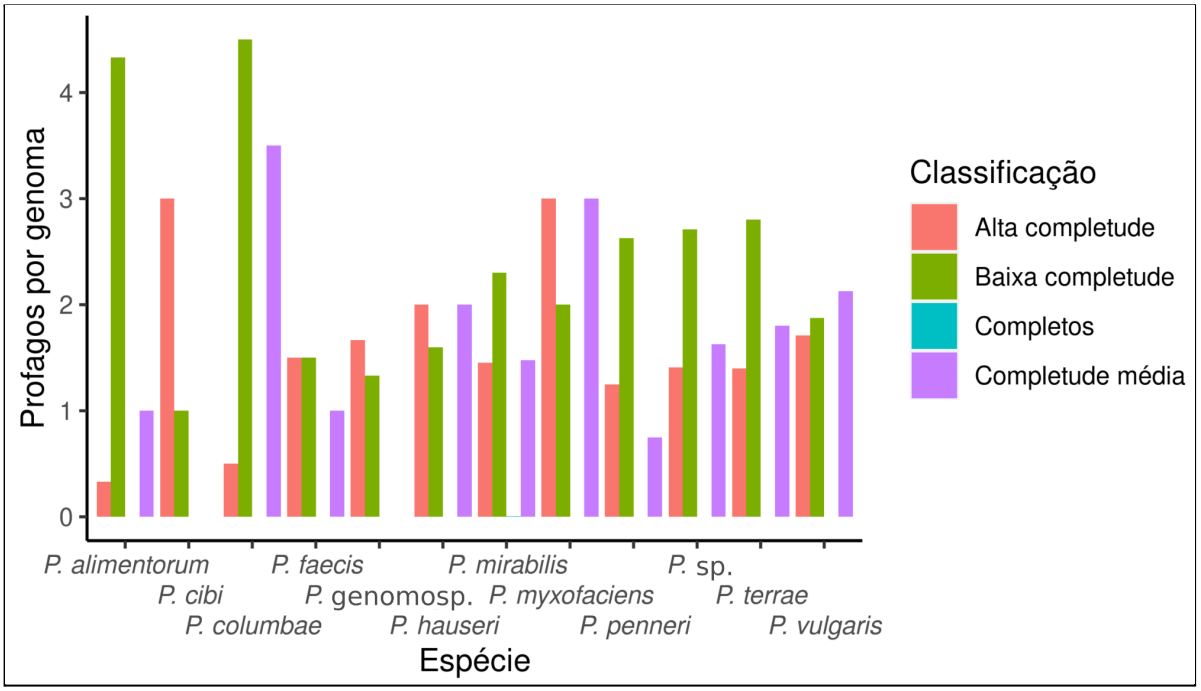

Figura 6 - Classificação da completude de profagos pelo programa CheckV por genoma das espécies de *Proteus* pós-processamento.

Na figura abaixo (Fig. 7), vemos que as espécies que mais apresentaram profagos de alta completude por genoma foram *P. cibi* e *P. myxofaciens*, ambas com três sequências. Por outro lado, a espécie que obteve mais profagos com média e baixa completude pelo número

de genomas foi *P. columbae*. Além disso, a espécie que deteve mais sequências de alta completude e completas (n=918), e de completude média e baixa (n=2373) foi a *P. mirabilis*, devido ao alto número de genomas (n=658) utilizados para a prospecção.

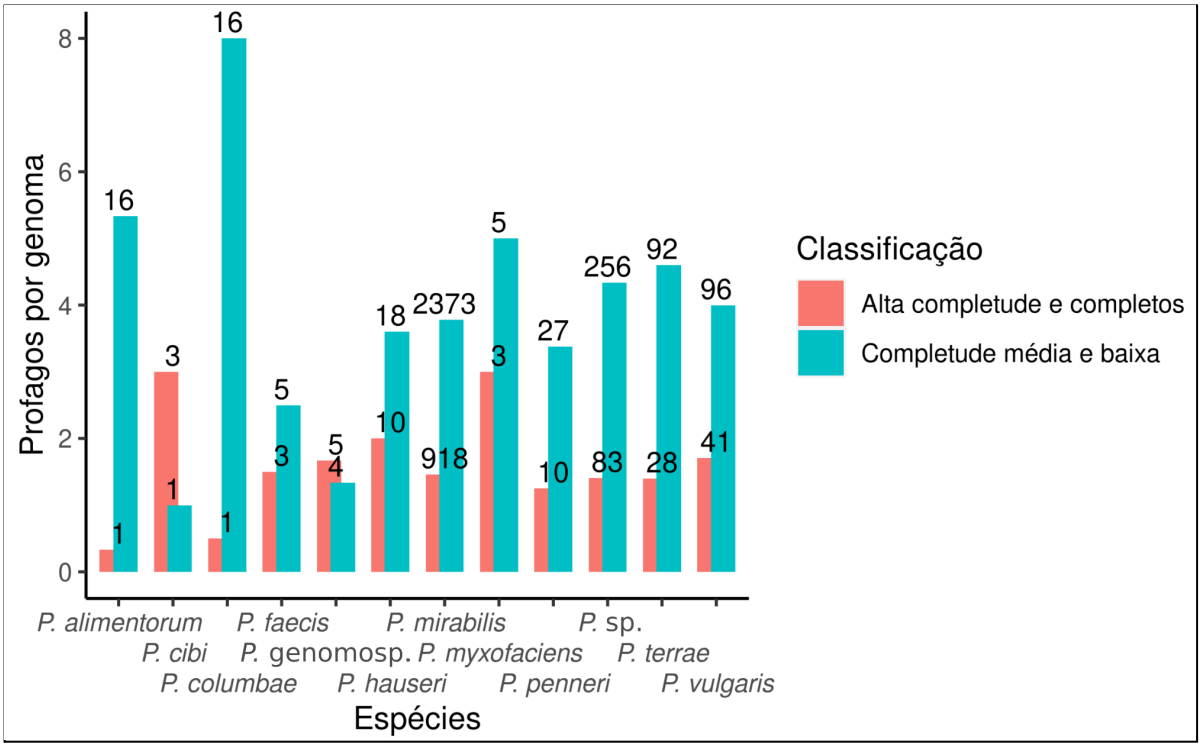

Figura 7 - Classificação da completude de profagos pelo programa CheckV por genoma das espécies de *Proteus* pós-processamento somadas. As barras representam o número de profagos por genomas bacterianos que foram encontrados em cada espécie e o número acima da barra representa o número absoluto de profagos encontrados. Por exemplo, a espécie *P. columbae* teve dois genomas incluídos neste estudo e nela foram encontrados 17 profagos, dos quais 16 tiveram qualidade média ou baixa e um é completo ou de alta qualidade. A barra tem tamanho 8, pois os 16 profagos são divididos entre os dois genomas bacterianos.

As sequências com ao menos 20 mil pares de base classificadas como completas ou de alta completude (n=1105), tiveram suas ORFs traduzidas, e foram agrupadas com sequências de espécies de bacteriófagos aceitas pelo ICTV (Fig. 8). Dessas sequências: 768 agruparam com outra sequência (*Clustered*) e delas 69 agruparam com alguma referência, 79 não agruparam com outras sequências formando grupos próprios *(Clustered/Singleton*), 26 não agruparam (*Outlier*) e 232 se agruparam em mais de um grupo (*Overlap*). O táxon superior comum a todas as sequências de referência foi a família *Myoviridae,* e a partir de então, tomamos que esta é a família à qual pertencem os genomas virais que agruparam as referências.

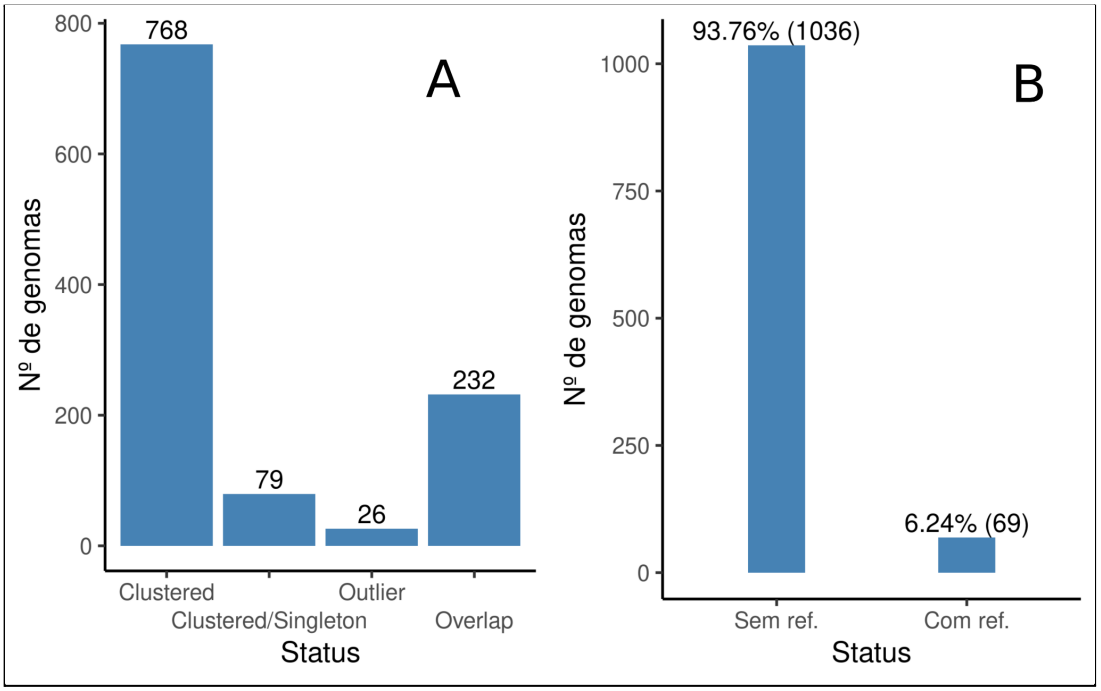

Figura 8 - Número de genomas por agrupamento do programa vContact2 (A) e bacteriófagos agrupados sem e com ref. (referência) (B). Entre parênteses está representada a quantidade de sequências. ref é abreviação hugo, então você tem que discriminar isso aqui na fig 8a.

Criando uma "árvore proteômica" dos genomas encontrados neste estudo e dos genomas de vírus já classificados em *Myoviridae*, obtivemos a distribuição destes genomas dentro da família. Os resultados obtidos indicaram que os novos bacteriófagos se dividem em dois grupos (Fig. 9, 10 e 11), um relacionado com o gênero *Bcepmuvirus* e outro próximo a subfamília *Peduovirinae*.

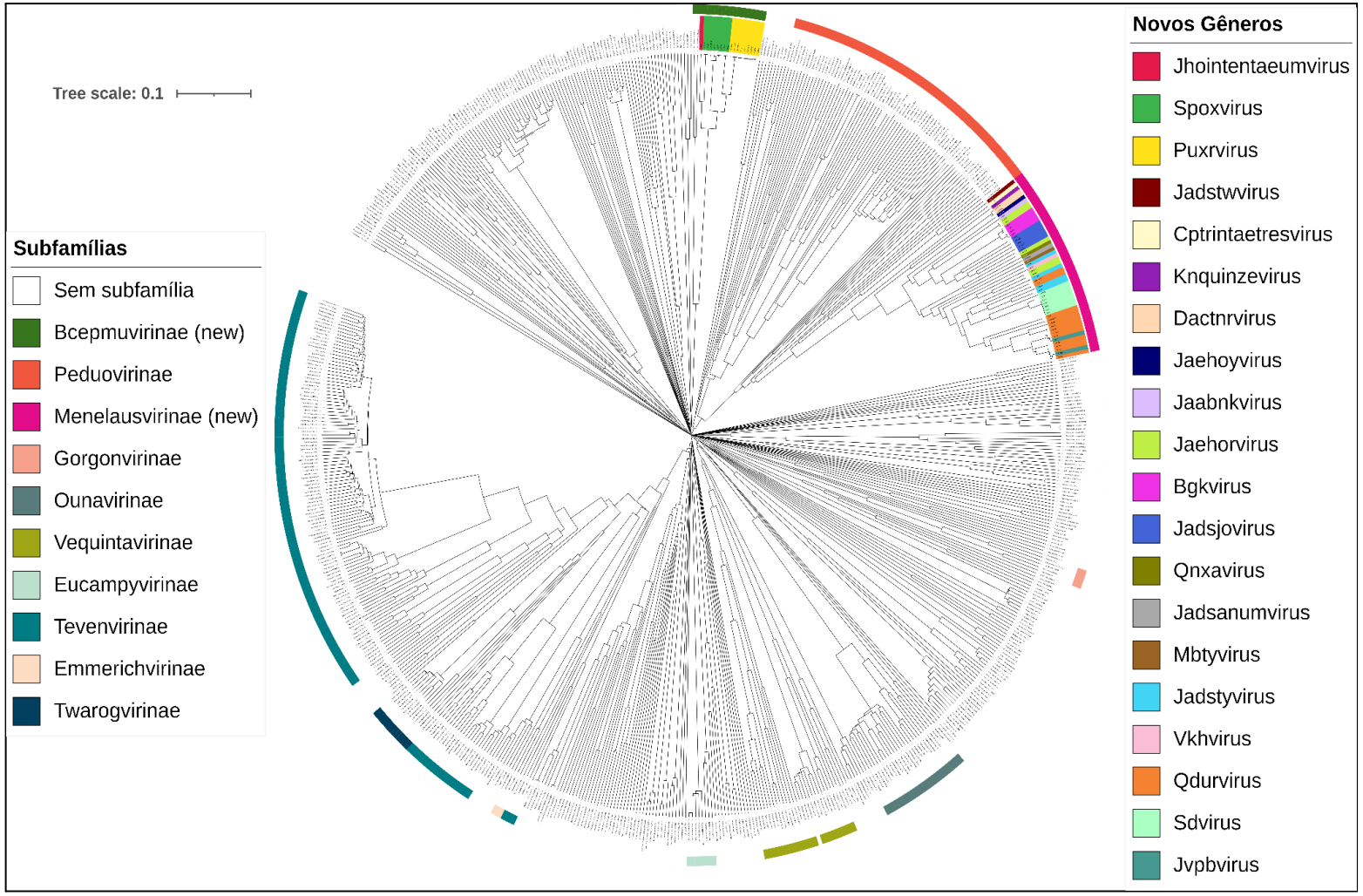

Figura 9 - "Árvore proteômica" de *Myoviridae* com os novos táxons representados. O arco superior e a legenda da esquerda correspondem às subfamílias. As cores abaixo do arco e a legenda da direita correspondem aos novos gêneros propostos.

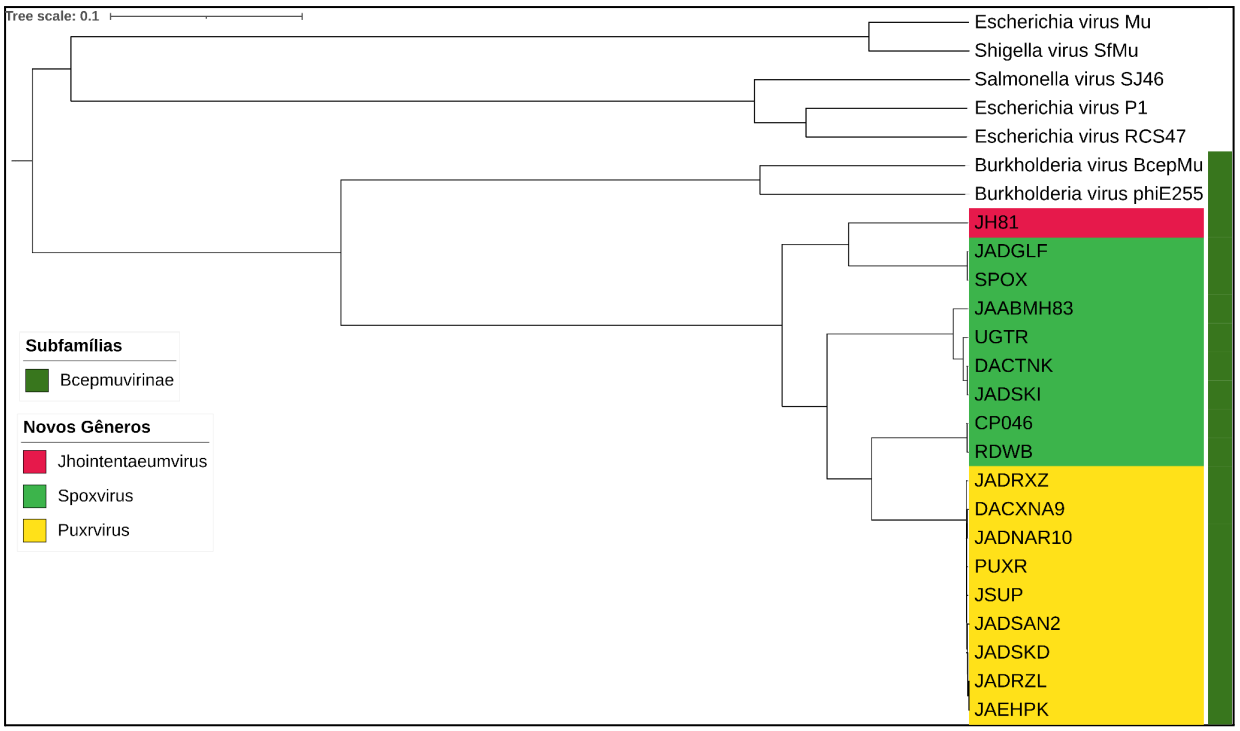

Figura 10 - Ramo destacado da "árvore proteômica" de *Myoviridae* onde está situada a subfamília *Bcepmuvirinae*. A marca a direita representa a extensão da nova subfamília e as cores correspondem aos novos gêneros propostos.

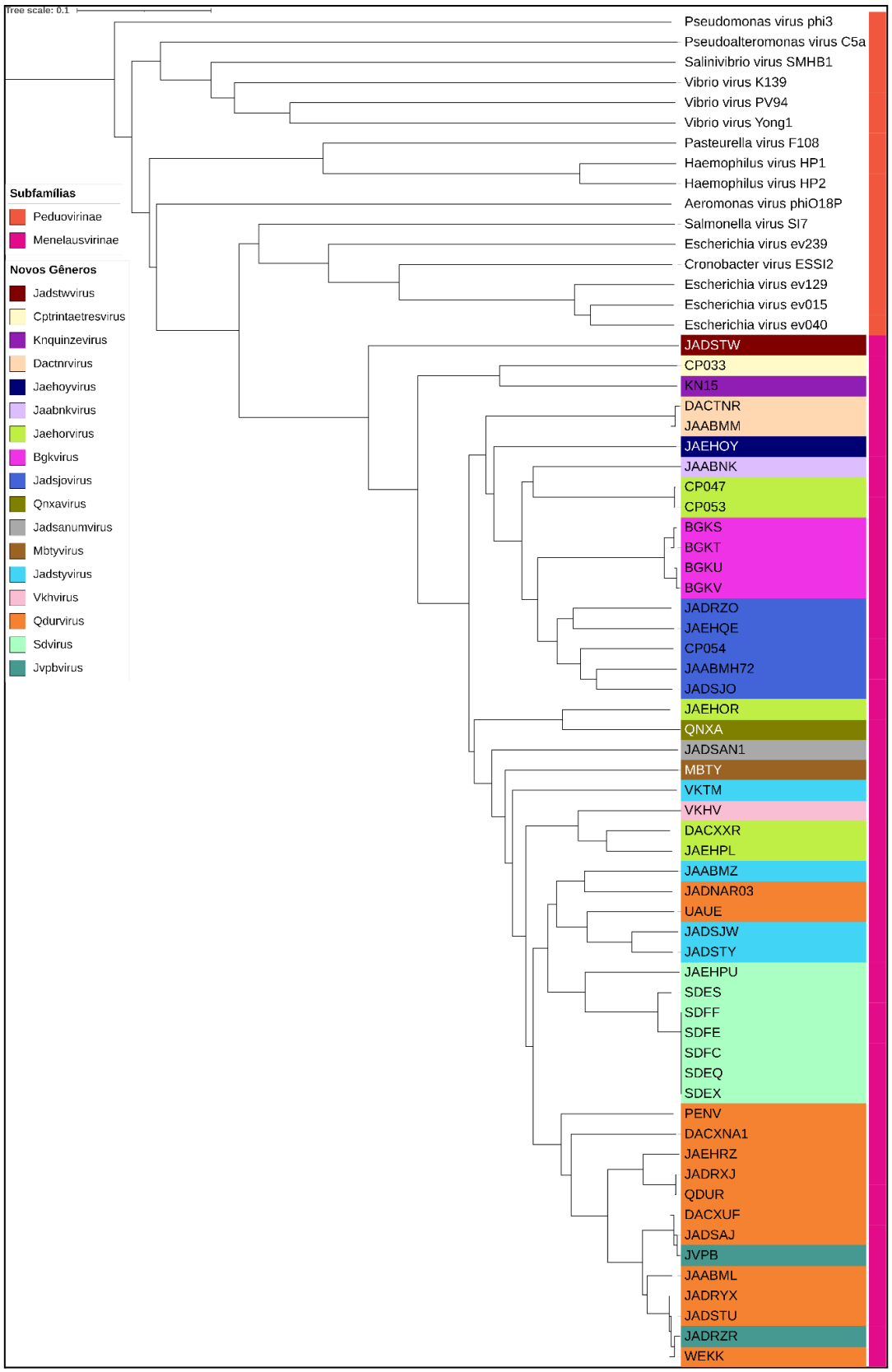

Figura 11 - Ramo destacado da "árvore proteômica" de *Myoviridae* onde está situada a subfamília *Menelausvirinae*. A marca a direita representa a extensão da nova subfamília e as cores correspondem aos novos gêneros propostos.

Através da análise do pangenoma viral de ambos os grupos, pudemos observar em quais subfamílias estes vírus estariam classificados. A formação de uma subfamília foi constatada quando ocorreu a existência de *core genes* no pangenoma viral.

Para o grupo relacionado ao gênero *Bcepmuvirus* houve a detecção de 31 *core genes* (Fig. 12) no pangenoma, sendo esses relacionados somente com os vírus do gênero *Bcepmuvirus*. Ao analisarmos juntamente outros clados próximos, não foi constatado a detecção de *core genes*, indicando que os vírus do grupo formam uma subfamília que foi nomeada por esse estudo de *Bcepmuvirinae*, devido ao nome do táxon mais antigo que ela contém, o gênero *Bcepmuvirus*.

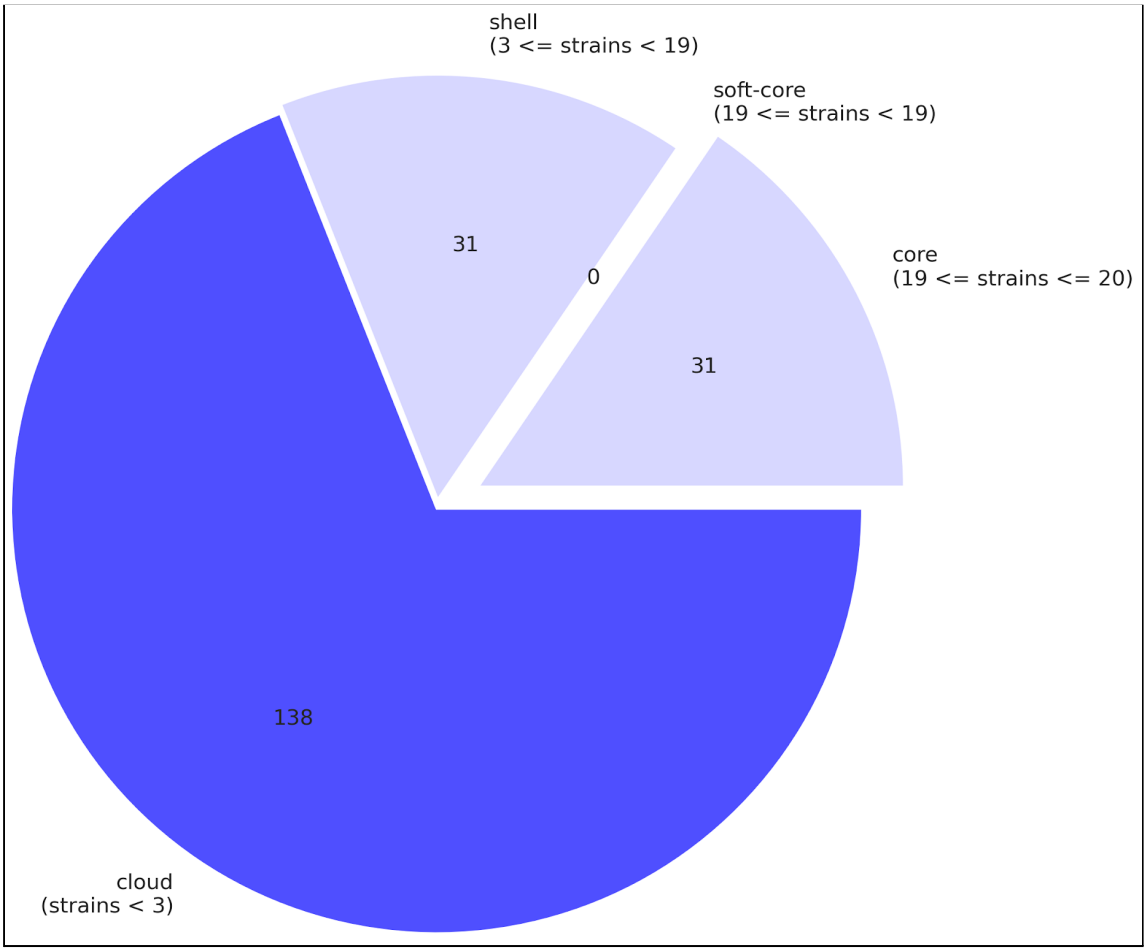

Figura 12 - Classificação pangenômica de *Bcepmuvirinae*. Entre parênteses está representada a quantidade de sequências em que um gene deve estar presente para pertencer a determinada categoria. Para ser considerado *core genes* os genes devem ser encontrados em ao menos 99% dos genomas, os *soft-core genes* são genes encontrados entre 95 e 99% dos genomas, os *shell genes* são encontrados entre 15% e 95% dos genomas, enquanto os *cloud genes* são os genes que estão presentes em menos de 15% dos genomas. Dentro do gráfico estão representados o número de genes de cada categoria.

Buscando identificar a subfamília do segundo grupo formado pelos novos bacteriófagos, nós realizamos uma análise de similaridade com o clado mais próximo, a subfamília *Peduovirinae*, porém não foram detectados *core genes* correspondentes. A detecção de 15 *core genes* (Fig. 13) ocorreu apenas entre o próprio grupo, o que pode sinalizar uma nova subfamília. Diante disso, nós sugerimos nomear a nova subfamília de *Menelausvirinae* em referência à passagem da Odisseia de Homero, onde é narrado o mito que o rei Menelau da Lacedemônia foi capaz de prender a deidade marinha Proteu e com isso conseguiu que este lhe contasse a verdade. O nome do gênero de bactérias *Proteus*, dos quais os genomas de vírus foram prospectados, faz referência ao mesmo mito devido à sua capacidade de apresentar várias formas, assim como o gênero dessa bactéria.

Ao catalogarmos a taxonomia de identidade dos gêneros e espécies dos novos bacteríofagos encontrados, analisamos os genomas avaliando a similaridade com os genomas de toda a família *Myoviridae* e também para cada uma das novas subfamílias. Dessa forma, analisamos se os vírus se tratavam de espécies já descritas e se pertenciam a algum gênero já existente.

Os resultados indicaram que as espécies não estavam inseridas em nenhum gênero da família *Myoviridae*, e assim formavam seus próprios gêneros, os quais nomeamos de acordo com o nome de uma das sequências dentro do gênero e o sufixo -*virus* (Fig. 9, 10 e 11). Para as espécies usamos o nome do gênero da bactéria hospedeira e a palavra vírus seguida de um código simplificado, que remete ao nome do arquivo da sequência.

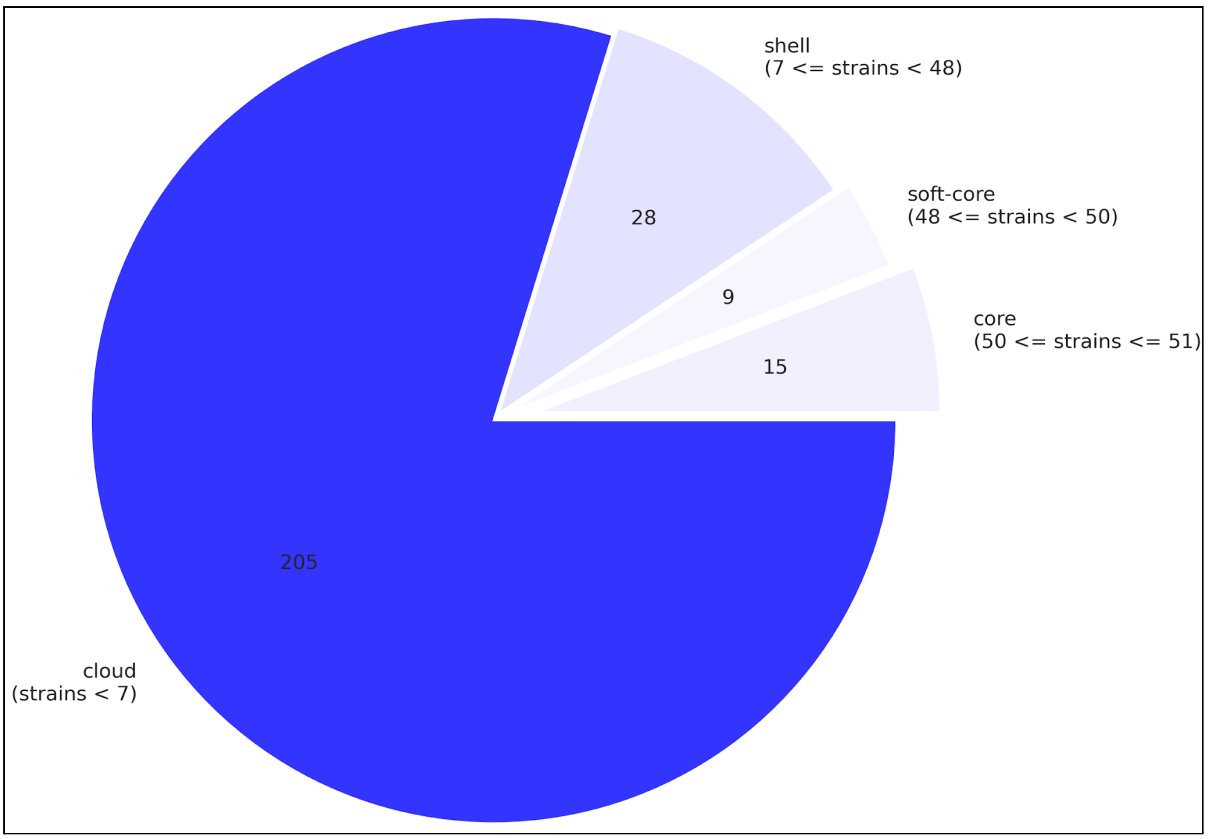

Figura 13 - Classificação pangenômica de *Menelausvirinae*. Entre parênteses está representada a quantidade de sequências em que um gene deve estar presente para pertencer a determinada categoria. Para ser considerado *core genes* os genes devem ser encontrados em ao menos 99% dos genomas, os *soft-core genes* são genes encontrados entre 95 e 99% dos genomas, os *shell genes* são encontrados entre 15% e 95% dos genomas, enquanto os *cloud genes* são os genes que estão presentes em menos de 15% dos genomas. Dentro do gráfico estão representados o número de genes de cada categoria.

Na subfamília *Bcepmuvirinae* (Fig. 14) identificamos que, além do gênero *Bcepmuvirus*, foram formados 3 novos gêneros:

- O gênero *Puxrvirus* contém 5 espécies: *Proteus virus DACXNA9* com uma sequência; *Proteus virus JADSAN2* com uma sequência; *Proteus virus JAEHPK* com duas sequências (JADRZL e JAEHPK); *Proteus virus JSUP* com uma sequência e *Proteus virus PUXR* com quatro sequências (PUXR, JADRXZ, JADNAR10 e JADSKD).
- O gênero *Spoxvirus* contém três espécies: *Proteus virus CP046* com duas sequências (CP046 e RDWB); *Proteus virus JADSKI* com quatro sequências (JADSKI, JAABMH83, DACTNK e UGTR) e *Proteus virus SPOX* com duas sequências (JADGLF e SPOX);

- O gênero *Jhointentaeumvirus* formou-se com uma espécie da sequência JH81. Devido à alta presença de "Ns" no genoma dessa sequência, é provável que se trate de um artefato que o VIRIDIC não pôde analisar devidamente, já que o programa aponta 91,3% de semelhança com ela mesma e o maior comprimento de seu genoma em relação ao genoma das outras sequências da mesma subfamília.

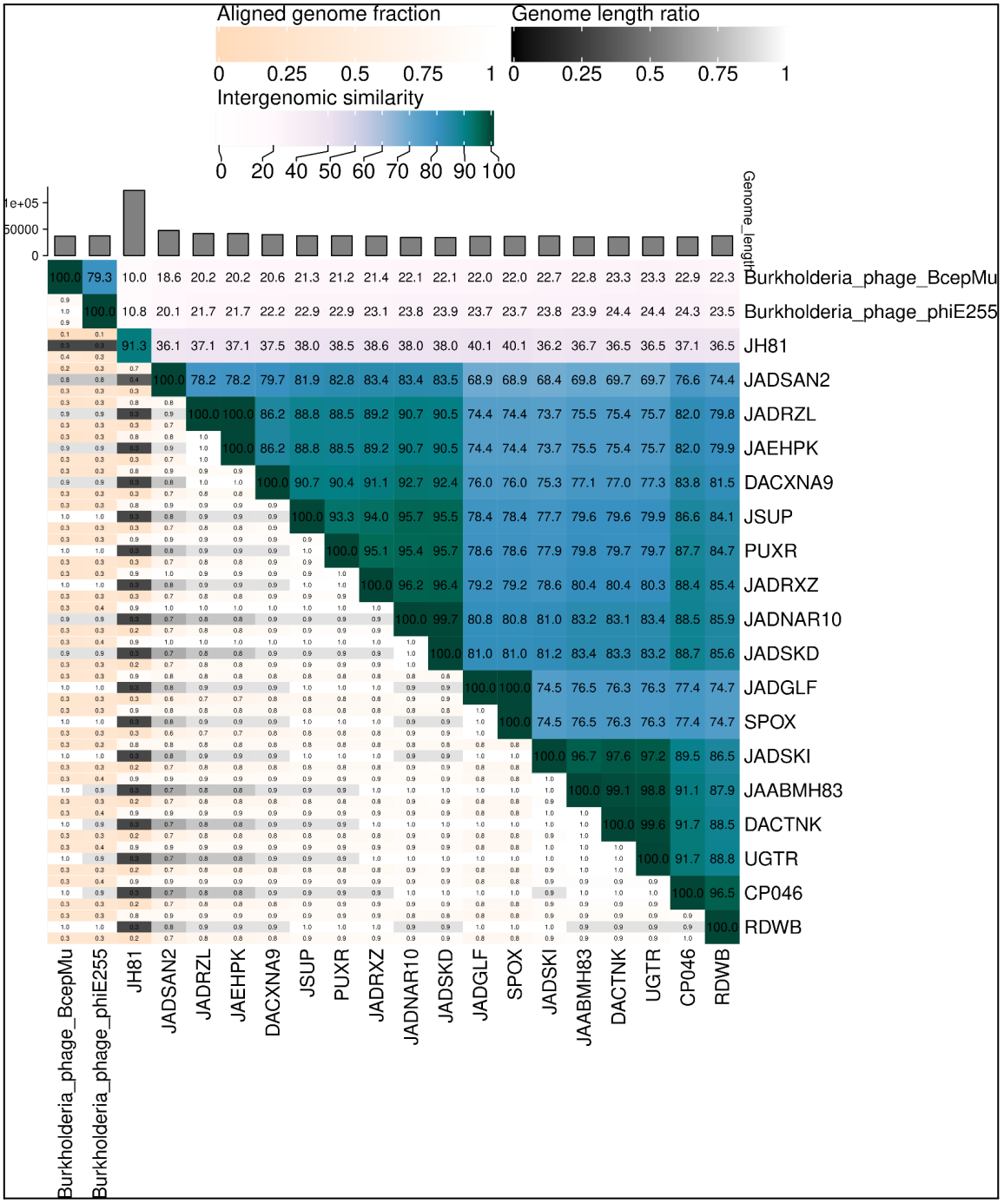

Figura 14 - Heatmap de similaridade da subfamília *Bcepmuviridae*. Na metade superior direita, o código de cores indica o agrupamento dos genomas com base na similaridade

intergenômica. Quanto mais similares os genomas, mais escura é a cor do quadrado em que está a porcentagem de similaridade dos genomas. Os números representam os valores de similaridade arredondados para a primeira casa após a vírgula. Na metade esquerda, há três valores, na ordem de cima para baixo: a fração alinhada do genoma daquela linha em relação ao genoma pareado com ele na coluna, arredondado para a primeira casa decimal. Abaixo, a proporção do comprimento do genoma (para os dois genomas neste par). Por último a fração alinhada do genoma da coluna sobre o genoma da linha. As cores representam os valores encontrados na linha, sendo que quanto mais escura a cor, menor é o valor encontrado. Espera-se que vírus mais próximos tenham comprimentos semelhantes.

Já na subfamília *Menelausvirinae*, os resultados indicam a formação de 17 novos gêneros (Fig. 15), sendo que 10 deles são gêneros com somente uma espécie. Os quais são: *Cptrintaetresvirus* com a espécie *Proteus virus CP033*; *Jaabnkvirus* com a espécie *Proteus virus JAABNK*; *Jadsanumvirus* com a espécie *Proteus virus JADSAN1*; *Jadstwvirus* com a espécie *Proteus virus JADSTW*; *Jaehoyvirus* com a espécie *Proteus virus JAEHOY*; *Knquinzevirus* com a espécie *Proteus virus KN15*, *Mbtyvirus* com a espécie *Proteus virus MTBY*; *Qnxavirus* com a espécie *Proteus virus QNXA* e *Vkhvirus* com a espécie *Proteus virus VKHV*, sendo todos com somente uma sequência, exceto *Jvpbvirus* com a espécie *Proteus virus JVPB* que possui duas sequências (JADRZR e JVPB).

Já o gênero *Bgkvirus* engloba duas espécies, *Proteus virus BGKT* com três sequências (BGKT,BGKS,BGKU) e *Proteus virus BGKV* com uma. O gênero *Dactnrvirus* também engloba duas espécies, *Proteus virus DACTNR* e *Proteus virus JAABMM*, com uma sequência cada. *Sdvirus* engloba três espécies, sendo elas *Proteus virus SDEX* com cinco sequências, *Proteus virus SDES* e *Proteus virus JAEHPU*, ambos com uma sequência.

Jadstyvirus e *Jaehorvirus* possuem quatro espécies cada. *Jadstyvirus* engloba *Proteus virus JAABMZ*, *Proteus virus JADSJW*, *Proteus virus JADSTY* e *Proteus virus VKTM*, todas com uma sequência. Além disso, *Jaehorvirus* engloba *Proteus virus DACXXR*, *Proteus virus JAEHOR*, *Proteus virus JAEHPL*, todas com uma sequência e *Proteus virus CP047* com duas (CP047 e CP053). O gênero *Jadsjovirus* engloba *Proteus virus JADSJO*, *Proteus virus JAEHQE*, *Proteus virus JADRZO*, *Proteus virus JAABMH72* e *Proteus virus CP054*, todas com uma sequência.

Qdurvirus é o gênero com mais espécies, sendo oito, e em número de sequências, treze. As espécies que possuem apenas uma sequência são: *Proteus virus DACXNA1*, *Proteus virus JADNAR03*, *Proteus virus JAEHRZ*, *Proteus virus PENV* e *Proteus virus UAUE*. Por outro lado, aquelas com duas são: *Proteus virus JADSAJ* (DACXUF,JADSAJ) e *Proteus*

virus QDUR (JADRXJ e QDUR). *Proteus virus WEKK* possui quatro sequências (JAABML, WEKK, JADRYX e JADSTU).

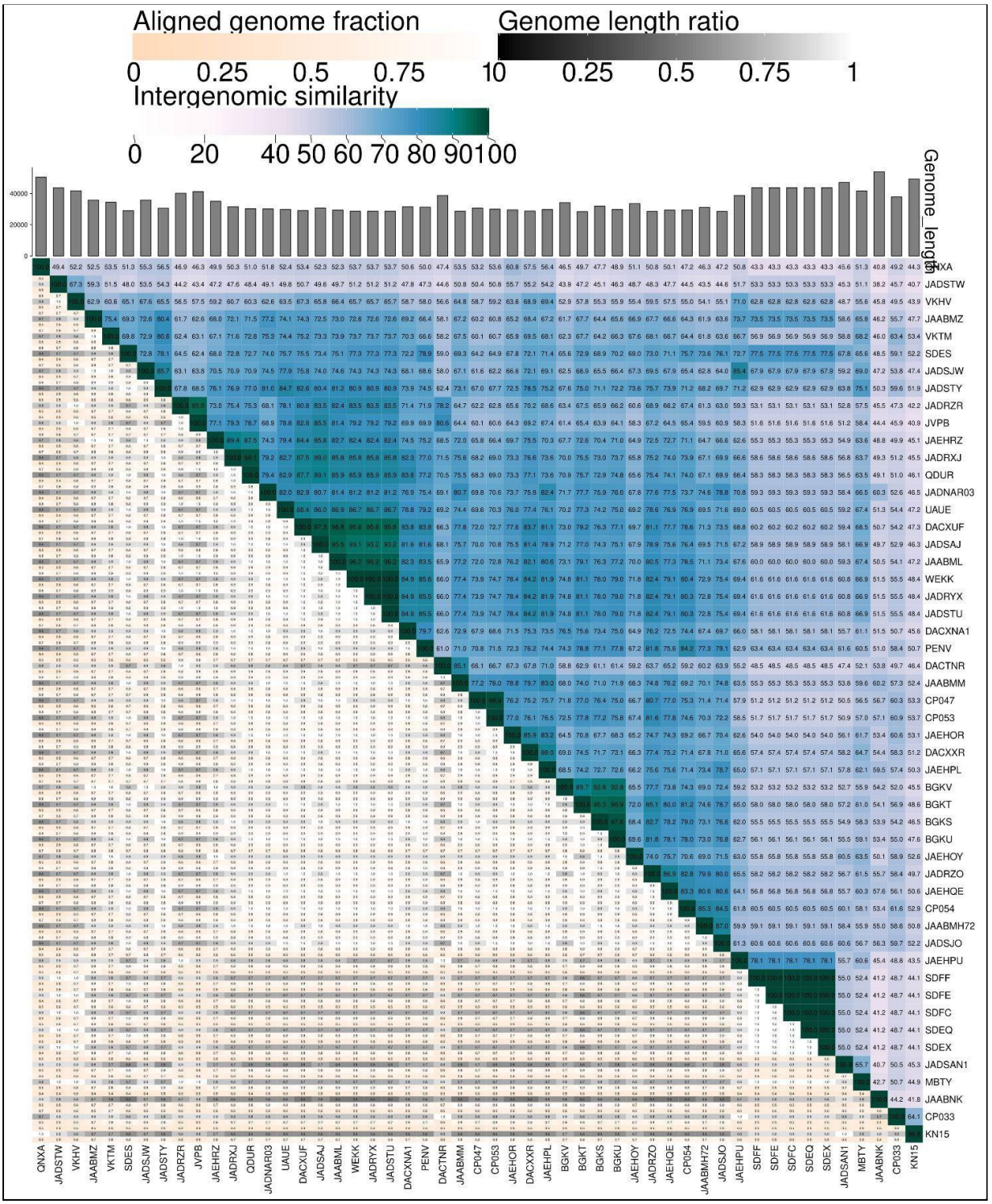

Figura 15 - Heatmap de similaridade da subfamília *Menelausviridae*. Na metade superior direita, o código de cores indica o agrupamento dos genomas com base na similaridade inter-genômica. Quanto mais similares os genomas, mais escura a cor dos genomas. Os

números representam os valores de similaridade arredondados para a primeira casa após a vírgula. Na metade esquerda, há três valores, na ordem de cima para baixo: a fração alinhada do genoma daquela linha em relação ao genoma pareado com ele na coluna, arredondado para a primeira casa decimal. Abaixo, a proporção do comprimento do genoma (para os dois genomas neste par). Por último a fração alinhada do genoma da coluna sobre o genoma da linha. As cores representam os valores encontrados na linha, sendo quanto mais escura a cor menor é o valor encontrado. Espera-se que vírus mais próximos tenham comprimentos semelhantes.

Discussão

Bactérias apresentam grandes variedades de sequências de profagos em seus genomas. Muitas dessas sequências conferem benefícios adaptativos ao hospedeiro (BOBAY; TOUCHON; ROCHA, 2014) e tornam-se verdadeiras reservas de diversidade genômica, para bactérias e outros bacteriófagos (NADEEM; WAHL, 2017). A ocorrência de profagos em genomas de *Proteus* spp. pode indicar que a diversidade genômica neste grupo pode ser amplamente afetada pelo fluxo gênico e as deleções que os bacteriófagos podem promover, como ocorre em outras bactérias da ordem *Enterobacteriales* (BAWN et al. 2020).

Ainda que tecnologias de sequenciamento genômico se tornaram cada vez mais baratas e a cada dia mais ferramentas de bioinformática estejam sendo desenvolvidas, muitos dos gêneros bacterianos que conhecemos atualmente não tiveram os seus bacteriófagos descritos (ZRELOVS; DISLERS; KAZAKS, 2020). Em verdade, muitos genomas bacterianos não foram explorados e analisados de forma profunda, sendo que em alguns casos até procedimentos mais básicos, como a anotação, não foram feitos da maneira correta (WOOD et al., 2012). Para que nós possamos aprimorar a análise dos dados já existentes, é necessário que implementemos soluções cada vez melhores para os problemas que enfrentamos nos experimentos e nas análises dos dados gerados.

A contaminação de sequências com fragmentos dos genomas de outros organismos é um problema na análise genética de microrganismos, porque pode levar à uma cascata de erros que afetam diversos trabalhos que utilizam os dados disponibilizados. Dessa forma, é muito recomendado que sejam utilizadas abordagens que mitiguem ou indiquem a ocorrência desses erros (MERCHANT; WOOD; SALZBERG, 2014; KOREN et al., 2014). O uso do CheckV (NAYFACH et al., 2020) nos traz a possibilidade prevenir a ocorrência de contaminação expressiva de fragmentos bacterianos nos genomas de vírus que venham a ser propostos como novas espécies. Além disso, o processo de descontaminação permitiu reduzir

a influência de genes advindos do hospedeiro na classificação dos genomas virais, aumentando a confiabilidade dos resultados obtidos.

Ainda assim, há a possibilidade de o CheckV descartar fagos completos e funcionais que naturalmente possuem uma quantidade significativa de genoma bacteriano em suas sequências, como ocorreu após remoção de contaminação. Dito isso, salientamos que os genomas considerados completos antes da descontaminação podem conter dados importantes para a compreensão da evolução bacteriana e da evolução destes fagos (WAHL; BATTESTI; ANSALDI, 2018). Ainda, os genomas incompletos também podem indicar que os fagos estão sob intensa seleção nos genomas bacterianos e ao se integrarem podem perder sua capacidade de formação de partículas infectantes. Além disso, mesmo ao perderem essa capacidade, ainda podem ter contribuído com fatores de virulência ou outro aparato genético interessante para a bactéria ou outros fagos (HARRISON; BROCKHURST, 2017).

A dificuldade de aplicar metodologias manuais que indiquem a completude dos genomas em grandes conjuntos de dados torna esse tipo de estudo extenuante. Nesse âmbito, o uso do CheckV (NAYFACH et at., 2020) facilitou a verificação da completude e contaminação dos genomas virais, o que possibilitou a melhora da completude dos profagos aqui encontrados. Mesmo assim, o CheckV ainda não descarta totalmente o uso da curadoria manual, pois a ferramenta ainda pode ter erros ou apresentar resultados imprecisos a depender do fago analisado. Por exemplo, se um fago for muito distinto daqueles que conhecemos e estão integrados à base de dados do programa, é possível que ele não consiga classificar este fagos como completos, apesar de ser um fago totalmente funcional. Outras técnicas e processos de análises podem ainda ser usados para melhorar a classificação de completude dos genomas de profagos, como a integração de métodos para a predição de genes estruturais (CANTU et al., 2020) ou de outros genes essenciais para a formação de um vírion com capacidade de se replicar. Aqui tratamos de um passo importante para permitir que essas melhorias ocorram e que a curadoria manual possa ser utilizada em conjunto de dados menores, ou como uma etapa de validação da completude genômica.

A segunda aplicação do CheckV (NAYFACH et al., 2020) nos permitiu inferir a completude dos profagos estudados de forma simples e sistemática. Isso torna o todo o processo de verificação, mais fácil de ser reproduzido. Nós ainda pudemos focar nos profagos que pudessem ser classificados ao nível de espécie e descrever novas espécies de acordo com os critérios propostos pelo *Minimum Information about an Uncultivated Virus Genome*

(MIUViG) (ROUX et al., 2018). Com isso permitimos que estudos de diversidade e ecologia viral possam ser conduzidos utilizando esses dados, pois diminui a chance de esses estudos estarem lidando com fragmentos virais ao invés de vírus completos.

A avaliação do CheckV não é um fim em si, com essa avaliação podemos também estabelecer comparações em relação a outros programas e de fato mensurar quando e o quanto determinado programa auxilia na definição de um fago completo. Devido à própria natureza da avaliação manual, essas comparações não eram tão simples de serem realizadas, pois envolviam métodos que nem sempre eram descritos (ZRELOVS; DISLERS; KAZAKS, 2020) e que utilizavam de particularidades da observação de cada observador, que nem sempre eram comparáveis.

Os genomas que foram classificados com o vContact2 (BOLDUC et al., 2017; JANG et al., 2019) apresentam diversas proteínas semelhantes com as referências com as quais eles se agruparam. Isso indica que, apesar de não ter classificado grande parte dos genomas, aqueles que foram classificados, possivelmente, possuem alguma relação filogenética com as referências do mesmo grupo. Os fagos que não agruparam com referências podem tratar se de novos gêneros virais que não se encaixam em nenhuma família descrita (BOLDUC et al., 2017; JANG et al., 2019). Para classificarmos esses genomas precisaremos de abordagens diferentes e que consigam classificá-los em níveis taxonômicos superiores, por exemplo análises filogenéticas de genes marcadores, como as proteínas do capsídeo (ADRIAENSSENS, 2021).

A “árvore proteômica” plotada com o VipTree (NISHIMURA et al., 2017) reforçou o resultado obtido no vContact2, pois espécies dos ramos próximos dos dois grupos formados pelos novos genomas também estavam presentes nos grupos de vírus em que as sequências se agruparam. Com a árvore também pudemos observar que a família Myoviridae é claramente separada em dois ramos que não compartilham semelhanças em relação ao seu conteúdo proteico, o que reforça os achados de trabalhos já publicados (TURNER; KROPINSKI; ADRIAENSSENS, 2021) que indicam a polifilia dessa família.

Com a análise de pangenoma, fomos capazes de definir de quais subfamílias os bacteriófagos possivelmente são membros e quais são os *core genes* que as definem. Dessa forma propomos duas novas subfamílias que se somam às outras oito subfamílias já aceitas de *Myoviridae* (AHMAD; ADDY; HUANG, 2021).

Utilizando os critérios do VIRIDIC de 70% de semelhança para definir um gênero e 95% de semelhança definindo uma espécie (MORARU; VARSANI; KROPINSKI, 2020), nós propomos 20 novos gêneros e 47 novas espécies. Somando aos gêneros e espécies já existentes em *Myoviridae* totalizam 237 gêneros e 672 espécies (AHMAD; ADDY; HUANG, 2021). Adicionalmente, todas as caracterizações virais necessitam futuramente de comprovações experimentais para validar as predições descritas neste trabalho.

Conclusão

Neste trabalho atingimos o nosso objetivo de explorar os profagos nos genomas de bactérias do gênero *Proteus*. Encontramos 3907 sequências de provável origem viral das quais, após serem processadas, 1106 foram classificadas como profagos completos ou de alta completude. Classificando esses profagos, propomos duas novas subfamílias, 20 novos gêneros e 47 novas espécies de vírus para integrarem a família *Myoviridae*.

Perspectivas futuras

Futuramente ainda pretendemos explorar diversas características dos profagos e dos novos táxons propostos, tais como a funcionalidade dos *core genes* que definem as duas novas subfamílias, a caracterização desses genomas e das suas proteínas. Ainda poderemos avaliar a funcionalidade ecológica desses fagos ao estudarmos quais deles tratam-se de fagos de ciclo lítico ou lisogênico, e o possível uso desses fagos na área médica ou de controle bacteriano baseado na presença de certos genes e características de seus genomas.

Da parte das bactérias poderemos estudar a contribuição dos profagos para a diversidade e patogenicidade das espécies do gênero *Proteus*, seja por carregamento de genes de resistência ou outros fatores de virulência importantes, e a predição da localização deles dentro do genoma bacteriano, se estão presentes em ilhas de patogenicidade ou em outras regiões importantes do genoma.

Referências

ACKERMANN, H.-W.; PRANGISHVILI, D.. Prokaryote viruses studied by electron microscopy. **Archives Of Virology**, [S.L.], v. 157, n. 10, p. 1843-1849, 3 jul. 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00705-012-1383-y.

ACKERMANN, Hans W. Bacteriophage taxonomy. **Microbiology Australia**, [S.L.], v. 32, n. 2, p. 90, 2011. CSIRO Publishing. http://dx.doi.org/10.1071/ma11090.

ADRIAENSSENS, Evelien M.. Phage Diversity in the Human Gut Microbiome: a taxonomist's perspective. **Msystems**, [S.L.], v. 6, n. 4, p. 1-5, 31 ago. 2021. American Society for Microbiology. http://dx.doi.org/10.1128/msystems.00799-21.

AHMAD, Abdelmonim Ali; ADDY, Hardian Susilo; HUANG, Qi. Biological and Molecular Characterization of a Jumbo Bacteriophage Infecting Plant Pathogenic Ralstonia solanacearum Species Complex Strains. **Frontiers In Microbiology**, [S.L.], v. 12, p. 1-15, 27 set. 2021. Frontiers Media SA. http://dx.doi.org/10.3389/fmicb.2021.741600.
Alegre: Artmed, 2012.

ARMBRUSTER, C. E.; MOBLEY, H. L. T. Merging mythology and morphology: the multifaceted lifestyle of Proteus mirabilis. **Nature reviews. Microbiology**, v. 10, n. 11, p. 743–54, 8 nov. 2012.

AUGUIE, Baptiste; ANTONOV, Anton; AUGUIE, Maintainer Baptiste. Package 'gridExtra'. **Miscellaneous Functions for "Grid" Graphics**, 2017.

BAR-ON, Yinon M.; PHILLIPS, Rob; MILO, Ron. The biomass distribution on Earth. **Proceedings Of The National Academy Of Sciences**, [S.L.], v. 115, n. 25, p. 6506-6511, 21 maio 2018. Proceedings of the National Academy of Sciences. http://dx.doi.org/10.1073/pnas.1711842115.

BAWN, Matt; ALIKHAN, Nabil-Fareed; THILLIEZ, Gaëtan; KIRKWOOD, Mark; WHEELER, Nicole E.; PETROVSKA, Liljana; DALLMAN, Timothy J.; ADRIAENSSENS, Evelien M.; HALL, Neil; KINGSLEY, Robert A.. Evolution of Salmonella enterica serotype Typhimurium driven by anthropogenic selection and niche adaptation. **Plos Genetics**, [S.L.], v. 16, n. 6, p. 1-29, 8 jun. 2020. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pgen.1008850.

BERNASCONI, O. J. et al. In vitro activity of three commercial bacteriophage cocktails against multidrug-resistant Escherichia coli and Proteus spp. strains of human and non-human origin. **Journal of global antimicrobial resistance**, v. 8, p. 179–185, mar. 2017.

BOBAY, L.-M.; ROCHA, E. P. C.; TOUCHON, M. The adaptation of temperate bacteriophages to their host genomes. **Molecular biology and evolution**, v. 30, n. 4, p. 737–51, abr. 2013.

BOBAY, L.-M.; TOUCHON, M.; ROCHA, E. P. C.. Pervasive domestication of defective prophages by bacteria. **Proceedings Of The National Academy Of Sciences**, [S.L.], v. 111, n. 33, p. 12127-12132, 4 ago. 2014. Proceedings of the National Academy of Sciences. http://dx.doi.org/10.1073/pnas.1405336111.

BOLDUC, Benjamin; JANG, Ho Bin; DOULCIER, Guilhem; YOU, Zhi-Qiang; ROUX, Simon; SULLIVAN, Matthew B.. VConTACT: an ivirus tool to classify double-stranded dna viruses that infect archaea and bacteria. **Peerj**, [S.L.], v. 5, p. 1-26, 3 maio 2017. PeerJ. http://dx.doi.org/10.7717/peerj.3243.

BORTOLAIA, Valeria; KAAS, Rolf s; RUPPE, Etienne; ROBERTS, Marilyn C; SCHWARZ, Stefan; CATTOIR, Vincent; PHILIPPON, Alain; ALLESOE, Rosa L; REBELO, Ana Rita; FLORENSA, Alfred Ferrer. ResFinder 4.0 for predictions of phenotypes from genotypes. **Journal Of Antimicrobial Chemotherapy**, [S.L.], v. 75, n. 12, p. 3491-3500, 11 ago. 2020. Oxford University Press (OUP). http://dx.doi.org/10.1093/jac/dkaa345.

BOTSTEIN, David. A THEORY OF MODULAR EVOLUTION FOR BACTERIOPHAGES. Annals Of The New York Academy Of Sciences, [S.L.], v. 354, n. 1, p. 484-491, nov. 1980. Wiley. http://dx.doi.org/10.1111/j.1749-6632.1980.tb27987.x.

BOYD, E. F.; BRÜSSOW, H. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. **Trends in microbiology**, v. 10, n. 11, p. 521–9, nov. 2002.

BREITBART, Mya; ROHWER, Forest. Here a virus, there a virus, everywhere the same virus? **Trends In Microbiology**, [S.L.], v. 13, n. 6, p. 278-284, jun. 2005. Elsevier BV. http://dx.doi.org/10.1016/j.tim.2005.04.003.

BRISTER, J. Rodney; AKO-ADJEI, Danso; BAO, Yiming; BLINKOVA, Olga. NCBI Viral Genomes Resource. **Nucleic Acids Research**, [S.L.], v. 43, n. 1, p. 571-577, 26 nov. 2014. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gku1207.

BRÜSSOW, H.; CANCHAYA, C.; HARDT, W.-D. Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. **Microbiology and Molecular Biology Reviews**, v. 68, n. 3, p. 560–602, 1 set. 2004.

CANCHAYA, C. et al. Prophage Genomics. **Microbiology and Molecular Biology Reviews**, v. 67, n. 3, p. 473–473, set. 2003.

CANTU, Vito Adrian; SALAMON, Peter; SEGURITAN, Victor; REDFIELD, Jackson; SALAMON, David; EDWARDS, Robert A.; SEGALL, Anca M.. PhANNs, a fast and accurate tool and web server to classify phage structural proteins. **Plos Computational Biology**, [S.L.], v. 16, n. 11, p. 1-18, 2 nov. 2020. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pcbi.1007845.

CASJENS, S. Prophages and bacterial genomics: what have we learned so far? **Molecular microbiology**, v. 49, n. 2, p. 277–300, 17 jul. 2003.

CHARIF, Delphine; LOBRY, Jean R.. SeqinR 1.0-2: a contributed package to the r project for statistical computing devoted to biological sequences retrieval and analysis. **Structural Approaches To Sequence Evolution**, [S.L.], p. 207-232, 2007. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-35306-5_10.

COMEAU, A. M.; KRISCH, H. M. War is peace — dispatches from the bacterial and phage killing fields. **Current Opinion in Microbiology**, v. 8, n. 4, p. 488–494, ago. 2005.

COUTURE-BEIL, Alex; COUTURE-BEIL, Maintainer Alex. Package 'rjson'. **URL: https://cran. r-project. org/web/packages/rjson/rjson. pdf**, 2018.

DAI, H. et al. Proteus alimentorum sp. nov., isolated from pork and lobster in Ma'anshan city, China. **International journal of systematic and evolutionary microbiology**, v. 68, n. 4, p. 1390–1395, 1 abr. 2018b.

DAI, H. et al. Proteus columbae sp. nov., isolated from a pigeon in Ma'anshan, China. **International journal of systematic and evolutionary microbiology**, v. 68, n. 2, p. 552–557, 1 fev. 2018a.

DAI, H. et al. Proteus faecis sp. nov., and Proteus cibi sp. nov., two new species isolated from food and clinical samples in China. **International journal of systematic and evolutionary microbiology**, v. 69, n. 3, p. 852–858, 1 mar. 2019.

DEDRICK, R. M. et al. Prophage-mediated defence against viral attack and viral counter-defence. **Nature microbiology**, v. 2, n. 3, p. 16251, 9 jan. 2017.

DRZEWIECKA, D. Significance and Roles of Proteus spp. Bacteria in Natural Environments. **Microbial ecology**, v. 72, n. 4, p. 741–758, 9 nov. 2016.

DY, R. L. et al. Remarkable Mechanisms in Microbes to Resist Phage Infections. **Annual Review of Virology**, v. 1, n. 1, p. 307–331, 3 nov. 2014.

FERNÁNDEZ, Lucía; GUTIÉRREZ, Diana; GARCÍA, Pilar; RODRÍGUEZ, Ana. The Perfect Bacteriophage for Therapeutic Applications—A Quick Guide. **Antibiotics**, [S.L.], v. 8, n. 3, p. 1-16, 23 ago. 2019. MDPI AG. http://dx.doi.org/10.3390/antibiotics8030126.

FUHRMAN, J. A. Marine viruses and their biogeochemical and ecological effects. **Nature**, v. 399, p. 541–548, 1999.

GUO, Jiarong; BOLDUC, Ben; ZAYED, Ahmed A.; VARSANI, Arvind; DOMINGUEZ-HUERTA, Guillermo; DELMONT, Tom O.; PRATAMA, Akbar Adjie; GAZITÚA, M. Consuelo; VIK, Dean; SULLIVAN, Matthew B.. VirSorter2: a multi-classifier, expert-guided approach to detect diverse dna and rna viruses. **Microbiome**, [S.L.], v. 9, n. 1, p. 1-13, 1 fev. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s40168-020-00990-y.

HARRISON, Ellie; BROCKHURST, Michael A.. Ecological and Evolutionary Benefits of Temperate Phage: what does or doesn't kill you makes you stronger. **Bioessays**, [S.L.], v. 39, n. 12, p. 1700112, 6 out. 2017. Wiley. http://dx.doi.org/10.1002/bies.201700112.

HATFULL, Graham F; HENDRIX, Roger W. Bacteriophages and their genomes. **Current Opinion In Virology**, [S.L.], v. 1, n. 4, p. 298-303, out. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.coviro.2011.06.009.

HOBBS, Zack; ABEDON, Stephen T.. Diversity of phage infection types and associated terminology: the problem with ⁞lytic or lysogenic⁞. **Fems Microbiology Letters**, [S.L.], v. 363, n. 7, p. 1-8, 29 fev. 2016. Oxford University Press (OUP). http://dx.doi.org/10.1093/femsle/fnw047.

HOCKENBERRY, Adam J.; WILKE, Claus O.. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. **Peerj**, [S.L.], v. 9, n. 1, p. 1-16, 6 maio 2021. PeerJ. http://dx.doi.org/10.7717/peerj.11396.

HYATT, Doug; CHEN, Gwo-Liang; LOCASCIO, Philip F; LAND, Miriam L; LARIMER, Frank W; HAUSER, Loren J. Prodigal: prokaryotic gene recognition and translation initiation site identification. **Bmc Bioinformatics**, [S.L.], v. 11, n. 1, p. 1-11, 8 mar. 2010. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2105-11-119.

HYUN, D.-W. et al. Proteus cibarius sp. nov., a swarming bacterium from Jeotgal, a traditional Korean fermented seafood, and emended description of the genus Proteus. **International Journal of Systematic and Evolutionary Microbiology**, v. 66, n. 6, p. 2158–2164, 10 jun. 2016.

JACOBSEN, S. M. et al. Complicated catheter-associated urinary tract infections due to Escherichia coli and Proteus mirabilis. **Clinical microbiology reviews**, v. 21, n. 1, p. 26–59, jan. 2008.

JANG, Ho Bin; BOLDUC, Benjamin; ZABLOCKI, Olivier; KUHN, Jens H.; ROUX, Simon; ADRIAENSSENS, Evelien M.; BRISTER, J. Rodney; KROPINSKI, Andrew M; KRUPOVIC, Mart; LAVIGNE, Rob. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. **Nature Biotechnology**, [S.L.], v. 37, n. 6, p. 632-639, 6 maio 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41587-019-0100-8.

JOENSEN, Katrine Grimstrup; SCHEUTZ, Flemming; LUND, Ole; HASMAN, Henrik; KAAS, Rolf S.; NIELSEN, Eva M.; AARESTRUP, Frank M.. Real-Time Whole-Genome Sequencing for Routine Typing, Surveillance, and Outbreak Detection of Verotoxigenic Escherichia coli. **Journal Of Clinical Microbiology**, [S.L.], v. 52, n. 5, p. 1501-1510, maio 2014. American Society for Microbiology. http://dx.doi.org/10.1128/jcm.03617-13.

KIEFT, Kristopher; ZHOU, Zhichao; ANANTHARAMAN, Karthik. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community

function from genomic sequences. **Microbiome**, [S.L.], v. 8, n. 1, p. 1-23, 10 jun. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s40168-020-00867-0.

KOREN, Sergey; TREANGEN, Todd J; HILL, Christopher M; POP, Mihai; PHILLIPPY, Adam M. Automated ensemble assembly and validation of microbial genomes. **Bmc Bioinformatics**, [S.L.], v. 15, n. 1, p. 1-9, 3 maio 2014. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2105-15-126.

KUMAR, Raj; SWAMINATHAN, T. Raja; KUMAR, Rahul G.; DHARMARATNAM, Arathi; BASHEER, V.s.; JENA, J.K.. Mass mortality in ornamental fish, Cyprinus carpio koi caused by a bacterial pathogen, Proteus hauseri. **Acta Tropica**, [S.L.], v. 149, p. 128-134, set. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.actatropica.2015.05.022.

KUTTER, Elizabeth; SULAKVELIDZE, Alexander (Ed.). **Bacteriophages: biology and applications**. Crc press, 2004.

LI, Zexin; PAN, Donald; WEI, Guangshan; PI, Weiling; ZHANG, Chuwen; WANG, Jiang-Hai; PENG, Yongyi; ZHANG, Lu; WANG, Yong; HUBERT, Casey R. J.. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. **The Isme Journal**, [S.L.], v. 15, n. 8, p. 2366-2378, 1 mar. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41396-021-00932-y.

LIVINGSTONE, Paul G.; MORPHEW, Russell M.; WHITWORTH, David E.. Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. **Frontiers In Microbiology**, [S.L.], v. 9, n. 1, p. 1-13, 19 dez. 2018. Frontiers Media SA. http://dx.doi.org/10.3389/fmicb.2018.03187.

MELO, L. D. R. et al. Development of a Phage Cocktail to Control Proteus mirabilis Catheter-associated Urinary Tract Infections. **Frontiers in Microbiology**, v. 7, n. JUN, 28 jun. 2016.

MERCHANT, Samier; WOOD, Derrick E.; SALZBERG, Steven L.. Unexpected cross-species contamination in genome sequencing projects. **Peerj**, [S.L.], v. 2, p. 1-17, 20 nov. 2014. PeerJ. http://dx.doi.org/10.7717/peerj.675.

MORARU, Cristina; VARSANI, Arvind; KROPINSKI, Andrew M.. VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. **Viruses**, [S.L.], v. 12, n. 11, p. 1268, 6 nov. 2020. MDPI AG. http://dx.doi.org/10.3390/v12111268.

MUSHEGIAN, A. R.. Are There 10 31 Virus Particles on Earth, or More, or Fewer? **Journal Of Bacteriology**, [S.L.], v. 202, n. 9, p. 1-14, 9 abr. 2020. American Society for Microbiology. http://dx.doi.org/10.1128/jb.00052-20.

NADEEM, A.; WAHL, Lindi M.. Prophage as a genetic reservoir: promoting diversity and driving innovation in the host community. **Evolution**, [S.L.], v. 71, n. 8, p. 2080-2089, 20 jun. 2017. Wiley. http://dx.doi.org/10.1111/evo.13287.

National Center for Biotechnology Information. Disponível em: <https://www.ncbi.nlm.nih.gov/>. Acesso em: 9 out. 2021.

NAYFACH, Stephen; CAMARGO, Antonio Pedro; SCHULZ, Frederik; ELOE-FADROSH, Emiley; ROUX, Simon; KYRPIDES, Nikos C.. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. **Nature Biotechnology**, [S.L.], v. 39, n. 5, p. 578-585, 21 dez. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41587-020-00774-7.

NISHIMURA, Yosuke; YOSHIDA, Takashi; KURONISHI, Megumi; UEHARA, Hideya; OGATA, Hiroyuki; GOTO, Susumu. ViPTree: the viral proteomic tree server. **Bioinformatics**, [S.L.], v. 33, n. 15, p. 2379-2380, 30 mar. 2017. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btx157.

NZAKIZWANAYO, J. et al. Bacteriophage Can Prevent Encrustation and Blockage of Urinary Catheters by Proteus mirabilis. **Antimicrobial agents and chemotherapy**, v. 60, n. 3, p. 1530–6, 28 dez. 2015.

O'HARA, C. M. et al. Classification of Proteus vulgaris biogroup 3 with recognition of Proteus hauseri sp. nov., nom. rev. and unnamed Proteus genomospecies 4, 5 and 6. **International journal of systematic and evolutionary microbiology**, v. 50 Pt 5, n. 5, p. 1869–1875, 1 set. 2000.

O'HARA, C. M.; BRENNER, F. W.; MILLER, J. M. Classification, identification, and clinical significance of Proteus, Providencia, and Morganella. **Clinical microbiology reviews**, v. 13, n. 4, p. 534–46, 1 out. 2000.

OFIR, Gal; SOREK, Rotem. Contemporary Phage Biology: from classic models to new insights. **Cell**, [S.L.], v. 172, n. 6, p. 1260-1270, mar. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.cell.2017.10.045.

PAGE, Andrew J.; CUMMINS, Carla A.; HUNT, Martin; WONG, Vanessa K.; REUTER, Sandra; HOLDEN, Matthew T.G.; FOOKES, Maria; FALUSH, Daniel; KEANE, Jacqueline A.; PARKHILL, Julian. Roary: rapid large-scale prokaryote pan genome analysis. **Bioinformatics**, [S.L.], v. 31, n. 22, p. 3691-3693, 20 jul. 2015. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btv421.

PETROVA, Z. O.; BROUSSARD, G. W.; HATFULL, G. F. Mycobacteriophage-repressor-mediated immunity as a selectable genetic marker: Adephagia and BPs repressor selection. **Microbiology (Reading, England)**, v. 161, n. 8, p. 1539–1551, 1 ago. 2015.

RAMISETTY, Bhaskar Chandra Mohan; SUDHAKARI, Pavithra Anantharaman. Bacterial 'Grounded' Prophages: hotspots for genetic renovation and innovation. **Frontiers In Genetics**, [S.L.], v. 10, n. 1, p. 1-17, 12 fev. 2019. Frontiers Media SA. http://dx.doi.org/10.3389/fgene.2019.00065.

ROHWER, Forest; EDWARDS, Rob. The Phage Proteomic Tree: a genome-based taxonomy for phage. **Journal Of Bacteriology**, [S.L.], v. 184, n. 16, p. 4529-4535, 15 ago. 2002. American Society for Microbiology. http://dx.doi.org/10.1128/jb.184.16.4529-4535.2002.

ROUX, Simon; ADRIAENSSENS, Evelien M; DUTILH, Bas e; KOONIN, Eugene V; KROPINSKI, Andrew M; KRUPOVIC, Mart; KUHN, Jens H; LAVIGNE, Rob; BRISTER, J Rodney; VARSANI, Arvind. Minimum Information about an Uncultivated Virus Genome (MIUViG). **Nature Biotechnology**, [S.L.], v. 37, n. 1, p. 29-37, 17 dez. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nbt.4306.

ROUX, Simon; PÁEZ-ESPINO, David; A CHEN, I-Min; PALANIAPPAN, Krishna; RATNER, Anna; CHU, Ken; REDDY, T B K; NAYFACH, Stephen; SCHULZ, Frederik; CALL, Lee. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. **Nucleic Acids Research**, [S.L.], v. 49, n. 1, p. 764-775, 2 nov. 2020. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkaa946.

ROUX, Simon; PAUL, Blair G.; BAGBY, Sarah C.; NAYFACH, Stephen; ALLEN, Michelle A.; ATTWOOD, Graeme; CAVICCHIOLI, Ricardo; CHISTOSERDOVA, Ludmila; GRUNINGER, Robert J.; HALLAM, Steven J.. Ecology and molecular targets of hypermutation in the global microbiome. **Nature Communications**, [S.L.], v. 12, n. 1, p. 1-12, 24 maio 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41467-021-23402-7.

SEEMANN, T.. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, [S.L.], v. 30, n. 14, p. 2068-2069, 18 mar. 2014. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btu153.

SNEATH, P. H. A.; MCGOWAN, V.; SKERMAN, V. B. D. Approved Lists of Bacterial Names. **International Journal of Systematic and Evolutionary Microbiology**, v. 30, n. 1, p. 225–420, 1 jan. 1980.

STERN, A.; SOREK, R. The phage-host arms race: shaping the evolution of microbes. **BioEssays : news and reviews in molecular, cellular and developmental biology**, v. 33, n. 1, p. 43–51, jan. 2011.

SUN, Yadong; WEN, Shanshan; ZHAO, Lili; XIA, Qiqi; PAN, Yue; LIU, Hanghang; WEI, Chengwei; CHEN, Hongyan; GE, Junwei; WANG, Hongbin. Association among biofilm formation, virulence gene expression, and antibiotic resistance in Proteus mirabilis isolates from diarrhetic animals in Northeast China. **Bmc Veterinary Research**, [S.L.], v. 16, n. 1, p. 1-10, 5 jun. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s12917-020-02372-w.

SUTTLE, C. A. Marine viruses--major players in the global ecosystem. **Nature reviews. Microbiology**, v. 5, n. 10, p. 801–12, out. 2007.

TEAM, R. Core et al. R: A language and environment for statistical computing. 2013.

TORTORA, Gerard J.; FUNKE, Berdell R.; CASE, Christine L. **Microbiologia**. 10. ed. Porto

TURNER, Dann; KROPINSKI, Andrew M.; ADRIAENSSENS, Evelien M.. A Roadmap for Genome-Based Phage Taxonomy. **Viruses**, [S.L.], v. 13, n. 3, p. 506, 18 mar. 2021. MDPI AG. http://dx.doi.org/10.3390/v13030506.

VANDAMME, Erick J; MORTELMANS, Kristien. A century of bacteriophage research and applications: impacts on biotechnology, health, ecology and the economy!. **Journal Of Chemical Technology & Biotechnology**, [S.L.], v. 94, n. 2, p. 323-342, 2 nov. 2018. Wiley. http://dx.doi.org/10.1002/jctb.5810.

VIERTEL, T. M.; RITTER, K.; HORZ, H.-P. Viruses versus bacteria-novel approaches to phage therapy as a tool against multidrug-resistant pathogens. **The Journal of antimicrobial chemotherapy**, v. 69, n. 9, p. 2326–36, 1 set. 2014.

WAHL, Astrid; BATTESTI, Aurélia; ANSALDI, Mireille. Prophages in Salmonella enterica: a driving force in reshaping the genome and physiology of their bacterial host?. **Molecular Microbiology**, [S.L.], v. 111, n. 2, p. 303-316, 25 dez. 2018. Wiley. http://dx.doi.org/10.1111/mmi.14167.

WANG, G.; DUNBRACK, R. L.. PISCES: a protein sequence culling server. **Bioinformatics**, [S.L.], v. 19, n. 12, p. 1589-1591, 11 ago. 2003. Oxford University Press (OUP). http://dx.doi.org/10.1093/bioinformatics/btg224.

WICKHAM, Hadley. ggplot2. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 3, n. 2, p. 180-185, 2011.

WICKHAM, Hadley. reshape2: Flexibly reshape data: a reboot of the reshape package. **R package version**, v. 1, n. 2, 2012.

WICKHAM, Hadley. stringr: modern, consistent string processing. R J., v. 2, n. 2, p. 38, 2010.

WICKHAM, Hadley. Tidy Data. **Journal Of Statistical Software**, [S.L.], v. 59, n. 10, p. 1-23, 12 set. 2014. Foundation for Open Access Statistic. http://dx.doi.org/10.18637/jss.v059.i10.

WICKHAM, Hadley; AVERICK, Mara; BRYAN, Jennifer; CHANG, Winston; MCGOWAN, Lucy; FRANÇOIS, Romain; GROLEMUND, Garrett; HAYES, Alex; HENRY, Lionel; HESTER, Jim. Welcome to the Tidyverse. **Journal Of Open Source Software**, [S.L.], v. 4, n. 43, p. 1686, 21 nov. 2019. The Open Journal. http://dx.doi.org/10.21105/joss.01686.

WICKHAM, Hadley; FRANCOIS, R. Dplyr. In: **useR! Conference**. 2014.

WOOD, Derrick e; LIN, Henry; LEVY-MOONSHINE, Ami; SWAMINATHAN, Rajiswari; CHANG, Yi-Chien; ANTON, Brian P; OSMANI, Lais; STEFFEN, Martin; KASIF, Simon; SALZBERG, Steven L. Thousands of missed genes found in bacterial genomes and their

analysis with COMBREX. **Biology Direct**, [S.L.], v. 7, n. 1, p. 37, 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1745-6150-7-37.

ZRELOVS, Nikita; DISLERS, Andris; KAZAKS, Andris. Motley Crew: overview of the currently available phage diversity. **Frontiers In Microbiology**, [S.L.], v. 11, n. 1, p. 1-6, 29 out. 2020. Frontiers Media SA. http://dx.doi.org/10.3389/fmicb.2020.579452.

Anexos

Tabela Suplementar 1

Espécies	Completude	Profagos pré-processamento	Profagos pós-processamento
P. alimentorum	Completos	1	0
P. alimentorum	Alta completude	1	1
P. alimentorum	Baixa completude	13	13
P. alimentorum	Completude média	2	3
P. cibi	Completos	1	0
P. cibi	Alta completude	1	3
P. cibi	Baixa completude	1	1
P. cibi	Completude média	1	0
P. columbae	Completos	1	0
P. columbae	Alta completude	3	1
P. columbae	Baixa completude	10	9
P. columbae	Completude média	3	7
P. faecis	Completos	3	0
P. faecis	Alta completude	0	3
P. faecis	Baixa completude	3	3
P. faecis	Completude média	2	2
P. genomosp.	Completos	2	0
P. genomosp.	Alta completude	3	5
P. genomosp.	Baixa completude	4	4

P. genomosp.	Completude média	0	0
P. hauseri	Completos	3	0
P. hauseri	Alta completude	5	10
P. hauseri	Baixa completude	8	8
P. hauseri	Completude média	12	10
P. mirabilis	Completos	449	5
P. mirabilis	Alta completude	576	913
P. mirabilis	Baixa completude	1427	1447
P. mirabilis	Completude média	839	926
P. myxofaciens	Completos	1	0
P. myxofaciens	Alta completude	0	3
P. myxofaciens	Baixa completude	3	2
P. myxofaciens	Completude média	4	3
P. penneri	Completos	5	0
P. penneri	Alta completude	3	10
P. penneri	Baixa completude	21	21
P. penneri	Completude média	8	6
P. sp.	Completos	37	0
P. sp.	Alta completude	56	83
P. sp.	Baixa completude	156	160
P. sp.	Completude média	90	96
P. terrae	Completos	13	0
P. terrae	Alta completude	24	28
P. terrae	Baixa completude	54	56
P. terrae	Completude média	29	36
P. vulgaris	Completos	32	0
P. vulgaris	Alta completude	21	41

P. vulgaris	Baixa completude	44	45
P. vulgaris	Completude média	40	51