**UNIVERSIDADE DE SÃO PAULO**

**FACULDADE DE MEDICINA DE RIBEIRÃO PRETO**

# Título: Análise Comparativa de Classificadores Taxonômicos de Bioinformática para Avaliação da Abundância Viral em Amostras Clínicas

**Aluno: Gabriel Montenegro de Campos**

Ribeirão Preto - SP

2022

Gabriel Montenegro de Campos

# Análise Comparativa de Classificadores Taxonômicos de Bioinformática para Avaliação da Abundância Viral em Amostras Clínicas

**Monografia do Trabalho de Conclusão de Curso de Informática Biomédica desenvolvido na Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo**

**Orientador:** Dr. Svetoslav Nanev Slavov



Este trabalho foi apresentado e aprovado pela Comissão Coordenadora do Curso deInformática Biomédica em 15/03/2023.

Ribeirão Preto - SP

2022

# AGRADECIMENTOS

Gostaria de agradecer a todos que me auxiliaram no decorrer desse Trabalho de Conclusão de Curso, seja de maneira direta ou indiretamente, de perto ou de longe. Em especial gostaria de agradecer ao meu orientador Dr. Svetoslav N. Slavov por ter me aceito como seu aluno em seu laboratório, me guiado durante o ano e ter me apresentado à área da pesquisa científica, especialmente na parte de Metagenômica, e às pessoas maravilhosas que compõem sua equipe de trabalho.

Não posso deixar de agradecer à Universidade de São Paulo, com destaque para a Faculdade de Medicina de Ribeirão Preto por todos os momentos que foram me proporcionados: desde aulas e simpósios científicos que participei até as experiências que contribuíram para meu repertório pessoal. Adicionalmente, agradeço a Bruno Rossi Carmo, Carlos Eduardo Capelini, Isabela Dias Erthal e Pedro Emílio Martins, colegas de graduação e agora amigos de formação da 17ª Turma de Informática Biomédica do *campus* de Ribeirão Preto.

Trabalhando nessa monografia, agradeço também ao Hemocentro de Ribeirão Preto por toda sua infraestrutura, ao Instituto Butantan por ter me deixado utilizar o espaço em seu servidor e pelas colaborações que me ajudaram a concluir meu Trabalho de Conclusão de Curso e à FAPESP pelo fomento à minha pesquisa.

O agradecimento à minha família, por todo seu suporte, amor e paciência investidos em mim durante essa e todas as outras fases de minha vida é imensurável.

# RESUMO

Os métodos metagenômicos são uma das ferramentas mais poderosas para identificação de vírus emergentes ou pouco conhecidos. Com o advento das tecnologias de sequenciamento de nova geração (SNG), a diversidade do software que é utilizado para a análise dos dados tem crescido progressivamente. No entanto, a análise bioinformática dos dados gerados de SNG ainda apresenta um desafio significativo para a interpretação correta dos resultados. Isso deve-se principalmente a variabilidade dos programas (e pipelines) utilizados para classificação taxonômica de cada laboratório e as altas exigências computacionais para o processamento dos mesmos. Portanto, o principal objetivo deste projeto é comparar os softwares de taxonomia viral mais utilizados nas pipelines de metagenômica de vírus. Para esta finalidade os classificadores foram aplicados utilizando dados de sequenciamento obtidos de diversas amostras clínicas (plasma de pacientes com câncer de próstata, doenças febris não identificadas e doenças respiratórias negativas para SARS-CoV-2). Feito isso, também foi realizada a montagem dos genomas dos vírus identificados e os mesmos foram submetidos a análise filogenética com o intuito de investigar a epidemiologia molecular dos vírus. Com esse projeto foi possível verificar, de acordo com métricas estatísticas e questões de usabilidade, qual o melhor classificador e identificar quais linhagens de dois vírus específicos estão circulando nas amostras. Esse projeto permitiu a padronização de uma pipeline de classificação viral para a avaliação mais profunda no da abundância viral em diversos grupos de pacientes.

**Palavras-chave:** Bioinformática. Metagenômica Viral. Classificadores Taxonômicos. Abundância Viral.

# ABSTRACT

Metagenomic methods are one of the most powerful tools for identifying emerging or little-known viruses. With the advent of next-generation sequencing (NGS) technologies, the diversity of software that is used for data analysis has progressively grown. However, the bioinformatics analysis of NGS generated data still presents a significant challenge for the correct interpretation of the results. This is mainly due to the variability of the programs (and pipelines) used for the taxonomic classification of each laboratory and the high computational demands for their processing. Therefore, the main objective of this project is to compare the most used viral taxonomy software for virus metagenomics pipelines. For this purpose the classifiers were applied using sequencing data obtained from several clinical samples (plasma from patients with prostate cancer, unidentified febrile illnesses and respiratory illnesses negative for SARS-CoV-2). Once this is done, the genomes of the identified viruses were assembled and then submitted to phylogenetic analysis in order to investigate the molecular epidemiology of the viruses. With this project, it was possible to verify, according to statistical metrics and usability issues, which is the best classifier and identify which strains of two specific viruses are circulating in the samples. This project allowed the standardization of a viral classification pipeline for the deeper assessment of viral abundance in different groups of patients.



**Keywords:** Bioinformatics. Viral metagenomics. Taxonomic Classifiers. Viral Abundance.

# 1. INTRODUÇÃO

## 1.1. Comunidades Microbianas

O ser humano convive em relação simbiótica com diversos micro-organismos, representantes dos Reinos das Bactérias, dos Fungos e também dos Vírus; o ambiente formado por esses seres é chamado de comunidade microbiana, microbioma ou ainda de microbiota (quando se considera um ambiente definido). Cerca de dois terços da nossa comunidade microbiana reside no trato gastrointestinal (VIRILI *et al*., 2018), mas também são encontrados representantes em nosso plasma e vários outros tecidos e fluídos corporais (CHIU e MILLER*,* 2019). A nossa microbiota influencia a educação de nosso sistema imune ao desenvolver mecanismos complexos para identificar e destruir micróbios invasores (DOMINGUEZ-BELLO *et al.,* 2019); a homeostase nutricional, metabólica e imunológica; o metabolismo em geral; a saúde dermatológica e até mesmo a saúde mental (VIRILI *et al.,* 2018*;* MULCAHY-O'GRADY e WORKENTINE, 2016). Porém, nem toda sua relação é simbiótica, pois a microbiota e sua alteração também está relacionada à alergias, doenças inflamatórias intestinais, doenças cardiovasculares, câncer e doenças autoimunes como asma (CHIU e MILLER, 2019).

Como citado anteriormente, o sangue humano pode conter também uma vasta variedade de microorganismos e, em alguns casos, agentes virais. Apesar da grande melhora no diagnóstico das doenças transmissíveis via transfusão de sangue como HIV (Vírus da Imunodeficiência Humana), HBV (Vírus da Hepatite B) e HCV (Vírus da Hepatite C), existem outros agentes virais chamados de emergentes que da mesma maneira podem ameaçar a segurança transfusional. Tais agentes por não serem diagnosticados pelos testes de rotina nos bancos de sangue podem causar impacto significativo na área de hemoterapia pela gravidade dos sintomas causados, por exemplo febre do Nilo Ocidental (causada pelo Vírus do Nilo Ocidental), síndrome do Choque (causada pelo Vírus da Dengue), encefalite japonesa (causada pelo Vírus da Encefalite Japonesa) entre outros (STRAMER *et al*., 2009). Portanto, a

metagenômica viral fornece uma ferramenta poderosa para identificação destas doenças infecciosas e elaboração de estratégias visando a prevenção transfusional das mesmas, além da segurança de uma transfusão de sangue (STRAMER *et al*., 2009; PETERSEN e BUSCH, 2010). Em termos mais gerais, a metagenômica demonstra grande utilidade na descoberta, classificação e organização das medidas de ação contra viroses emergentes.

## 1.2. Metagenômica

Por conta da ampla intervenção do microbioma em nossas vidas, é muito importante e necessário a identificação e caracterização dessa comunidade; no ramo da Bioinformática, o sequenciamento de todo um bioma de uma amostra é chamado de Metagenômica e vem ganhando popularidade devido ao sequenciamento de última geração (*Next-Generation Sequencing - NGS*) que permite mais dados em menos tempo e por um custo menor (NOOIJ *et al*, 2018). Abordagens metagenômicas são aquelas que sequenciam todo o DNA ou RNA de uma amostra (CHIU e MILLER, 2019), seja o solo tóxico, seja o ambiente marinho, seja o corpo humano, permitindo a análise de todo o microbioma presente em amostras de pacientes.

As ferramentas metagenômicas portanto são usadas para a detecção de todos os genomas dos microrganismos (bactérias, fungos e vírus) simultaneamente em uma amostra; o que nos permite descobrir, por exemplo, novos patógenos virais, caracterizar o metaviroma (conjunto de todos os vírus) humano em estados saudáveis e patológicos, utilizar em aplicações forenses e inferir sobre o perfil microbiótico de uma amostra; tudo para melhorar o diagnóstico de doenças (DE VRIES *et al.,* 2021). A metagenômica também remove a necessidade de projetar e sintetizar *primers* de PCR (*Polymerase Chain Reaction)* para vírus específicos (KISELEV *et al.,* 2020), uma vez que todos os representantes da amostra serão encontrados. Através da metagenômica, também, são identificados agentes virais ou microorganismos em geral que demonstram difícil ou impossível cultivo em meios de cultura comerciais ou até mesmo diagnóstico laboratorial (CHIU e MILLER, 2019).

### 1.2.1. Metagenômica Viral

Dentro da Metagenômica, a parte de metagenômica viral, ou metaviroma, busca identificar apenas os vírus presentes na amostra; o que permite o diagnóstico de pessoas com doenças infecciosas suspeitas mas com agente etiológico desconhecido. Como por exemplo encefalite de causa desconhecida, infecções virais, vírus respiratórios em crianças, vírus patógenos presentes no pulmão, entre outras síndromes clínicas (DE VRIES *et al.,* 2021).

A metagenômica viral pode revelar vírus reemergentes que não podem ser cultivados e emergentes que podem infectar pessoas de diversos perfis e situações que exigem diversos testes para o descobrimento dos mesmos (NOOIJ, *et al., 2018)*. Um exemplo é, em situações oncológicas, descobrir vírus associados a diversos tipos de câncer (herpesvírus, papilomavírus, poliomavírus) ou seu envolvimento nas vias de sinalização (CHIU e MILLER, 2019).

## 1.3. Análise do Metaviroma

Um dos principais problemas enfrentados pela análise do metaviroma é o desenvolvimento de métodos computacionais robustos para a identificação dos micro-organismos devido ao alto número de *reads* (leituras), fragmentos curtos de DNA, de comprimento geralmente de 150 pares de base (pb) presentes no sequenciamento de uma amostra e por conta do crescente número de genomas de microrganismos sequenciados (YE *et al.,* 2019).

A análise começa com a extração do DNA da amostra de diversas origens (desde o solo até amostras clínicas) e termina com a investigação taxonômica, incluindo vários passos durante o processo. Esses passos combinados são chamados de *pipelines* e servem como instruções do que a máquina deve realizar. Há diversos programas e pipelines que servem para a classificação taxonômica do metaviroma, cujas métricas mais usadas para sua avaliação são acurácia, velocidade e requerimentos computacionais (YE *et al.,* 2019; NOOIJ, *et al.,* 2018).

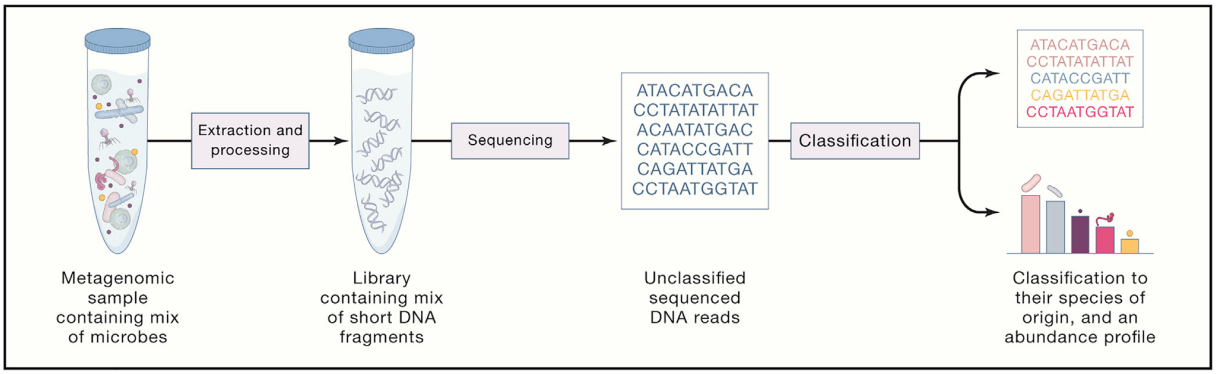



Figura 1: Etapas de processamento das amostras desde a extração do material genômico até anotação do perfil de abundância de conteúdo de amostra. (YE *et al.,* 2019).

Entretanto, ainda há alguns problemas na classificação do metaviroma, como a alta variabilidade dos dados taxonômicos, banco de dados utilizados na classificação muito abrangentes e estáticos, além de alto uso de memória computacional (YE *et al.,* 2019).

### 1.4. Classificação Taxonômica

Vários *softwares* foram desenvolvidos para classificar taxonomicamente os dados metagenômicos e estimar a abundância dos táxons. Para poder compará-los, é necessário saber como funcionam e as melhores maneiras de avaliá-los (YE *et al*., 2019).

Ye *et al.* e Buffet-Bataillon *et al.* sugerem para efeitos comparativos usar a Curva Precisão-Revocação feita pela linguagem de programação Python. A precisão é a proporção de espécies positivas verdadeiras identificadas na amostra dividida pelo número total de espécies identificadas pelo método, enquanto o *recall* (revocação ou sensibilidade) é definido como a proporção de espécies positivas verdadeiras dividida por o número de espécies distintas realmente na amostra.

## 2. OBJETIVOS

### 2.1. Objetivo Geral

O objetivo principal deste projeto, é, portanto, comparar os classificadores taxonômicos mais utilizados na atualidade para classificação de

leituras virais provenientes de arquivos de sequenciamento de última geração (formato fastq, *paired-end*) obtidos de diversos grupos de pacientes.

### 2.2. Objetivos Específicos

2.2.1. Implementação de pipeline base para análise dos dados de sequenciamento de última geração;

2.2.2. Análises comparativas de classificadores taxonômicos na base de sequências nucleotídicas (Kraken2 e CLARK);

2.2.3. Análises comparativa de classificadores taxonômicos na base de aminoácidos (DIAMOND e Kaiju);

2.2.4. Apresentar a aplicação de um bom perfil taxonômico gerado a partir de algum classificador

2.2.4.1. Montagem de novo utilizando o programa SPAdes v. 3.15.4 e

2.2.4.2. Vigilância genômica de alguns genomas completos pertencentes a vírus com importância para a saúde pública.

## 3. MATERIAIS E MÉTODOS

### 3.1. Fluxo de Trabalho

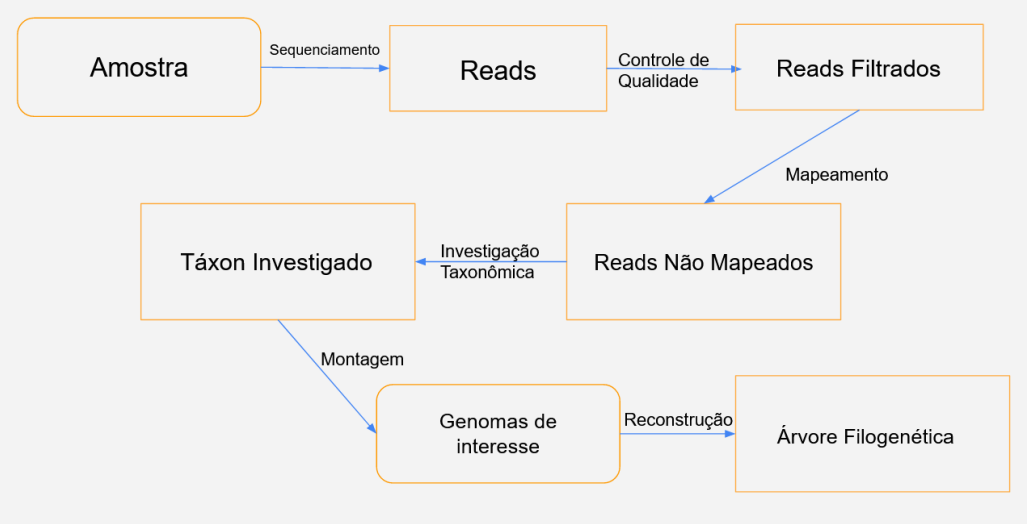



Figura 2: Visão geral do fluxo de trabalho ao ser realizado; da preparação da amostra até a análise. Adaptado de https://genomics.sschmeier.com/index.html

3.1.1. Sequenciamento de Última Geração para Obtenção de Leituras e Controle de Qualidade

O Fluxo de Trabalho começa com o sequenciamento das amostras clínicas (detalhadas em 3.2) pelo Sequenciador NextSeq 2000 localizado no Laboratório Estratégico de Sequenciamento de SARS-CoV-2 no Instituto Butantan para a obtenção dos reads, os reads são armazenados no Banco de Dados Illumina BaseSpace em formato FASTQ. O sequenciador usa a tecnologia de *paired-end*, que inclui sequenciar ambas as extremidades (pontas) dos fragmentos de DNA em uma biblioteca e o alinhar das leituras diretas (*forward*) e reversa (*reverse*) como pares de leitura, exemplificado na Figura 3.

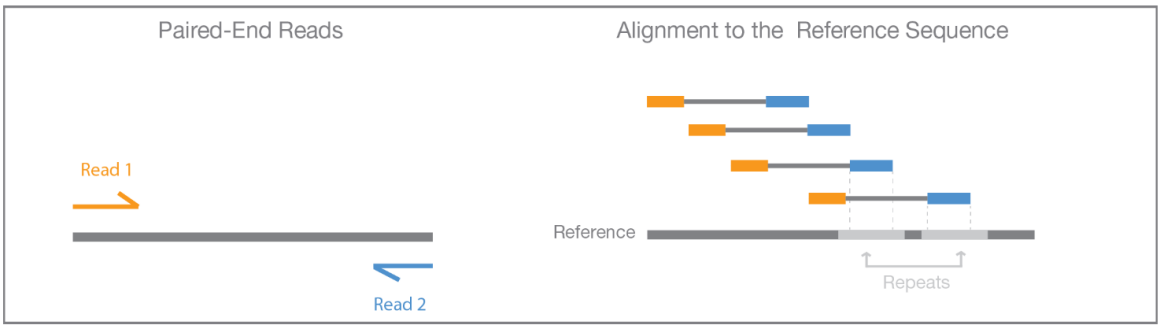



Figura 3: Sequenciamento e Alinhamento de leituras *paired-end* (PE): permite que ambas as extremidades do fragmento de DNA sejam sequenciadas. Como a distância entre cada leitura emparelhada é conhecida, os algoritmos de alinhamento podem usar essas informações para mapear as leituras sobre regiões repetitivas com mais precisão. Isso resulta em um melhor alinhamento das leituras, especialmente em regiões repetitivas e de difícil sequenciação do genoma (ILLUMINA - Paired-End vs Single-Read, 2022).

Os reads passam por um controle de qualidade para filtrar aqueles de baixa qualidade (presença de adaptadores, caudas poli-N, poli-A ou similares) e que podem prejudicar a análise (BATUT *et al.,* 2022). Deste modo obteremos as leituras *reads* filtradas. Para o controle de qualidade utiliza-se os programas *FastQC* ("Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data", 2019) e *fastp* (CHEN, *et al.,* 2018). Em geral, é feita uma análise inicial dos dados brutos usando o *FastQC*, e, em seguida, o fastp para detectar e remover, das sequências, caudas poli-G, poli-X, regiões sobrepostas e adaptadores, além de filtrá-las pelo escore de qualidade com valor 30, seguindo a fórmula $Q = -10 \log P$. Isso

significa que a probabilidade de uma base incorreta é 1 em 1000, virtualmente perfeito (ILLUMINA, 2022).

### 3.1.2. Remoção de Sequências do Hospedeiro: Mapeamento

A partir desse momento, como na amostra há uma quantidade significativa de *reads* do hospedeiro e de outros microrganismos, é realizado um processo de mapeamento para identificar quais são provenientes de um genoma de referência (no caso o genoma humano - GRCh38, disponível em https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40. Acesso em 14 de Fev. de 2022) e quais não são. Os programa mais recomendados para o mapeamento são os BWA (LI; DURBIN, 2009) e o Bowtie2 (LANGMEAD et al., 2009), e, junto com o programa samtools (LI et al., 2009), é possível gerar um arquivo apenas das leituras não mapeadas, ou seja, aquelas que não foram classificadas como genoma humano. Ambos utilizam como algoritmo a Transformada de Burrows-Wheeler (descrito em 3.2.2) e possuem performance semelhante quando usados com *paired-end*, porém, o BWA consegue recuperar informações sobre as pontas, pois utiliza o alinhamento local nesse caso, diferente do Bowtie2 (LANGMEAD e SALZBERG, 2012), o que pode melhorar seu mapeamento.

Uma vez que na amostra encontram-se apenas leituras não-mapeadas é necessário realizar a classificação taxonômica, com a finalidade de identificar os vírus presentes na amostra utilizando os classificadores *Kaiju, Clark, Kraken2* e *DIAMOND* podendo revelar uma variação nas leituras classificadas (MOUSTAFA *et al*., 2017). Em seguida, os vírus com interesse clínico são selecionados e montados para se fazer a reconstrução filogenética deles.

## 3.2. Classificação taxonômica: Classificadores Taxonômicos

Para o projeto, os classificadores a serem comparados são: Kraken 2 e CLARK, na base de dados de nucleotídeos, Kaiju e DIAMOND na base de dados de sequências proteicas.

### 3.2.1. Kraken 2

O Kraken é um classificador taxonômico de 2014 escrito em C++ e em *Pearl* por Derrick Wood e Jennifer Lu do Centro para Biologia Computacional (*Center for Computational Biology* - CBB) da Universidade John Hopkins (WOOD e SALZBERG, 2014). O Kraken utiliza um algoritmo de memória intensivo que usa *substrings* de nucleotídeos curtas (*K-mers*) com o táxon de referência do ancestral comum mais distante (*Last Common Ancestor* - *LCA*), consultado, por padrão e recomendação, o banco de dados RefSeq do NCBI (WOOD e SALZBERG, 2014). É possível também que o usuário crie um banco de dados personalizado para diversos fins, não necessariamente utilizando o RefSeq.

Os *K-mers* exigem grande requisito de memória, portanto, foi criado o Kraken 2, em 2018. O novo programa utiliza uma tabela *hash* compacta e probabilística; a tabela faz com que haja ganho em especificidade e acurácia dos resultados, mais velocidade e menos uso de memória de execução em relação ao Kraken 1 (cerca de $1/3$ de memória reduzida e cinco vezes mais rápido) por armazenar um par chave-valor de 32 bits (sendo 17 bits dedicados aos ID 's do banco de dados e 15 bits para o código *hash*). Além disso, o novo programa só armazena minimizadores de comprimento $l$, (sendo $l < k$), enquanto o Kraken armazenava todos os *K-mers* (WOOD *et al.,* 2019).

Sua documentação está disponível em <https://github.com/DerrickWood/kraken2/blob/master/docs/MANUAL.markdown >. Acesso em 03 de Junho de 2022.

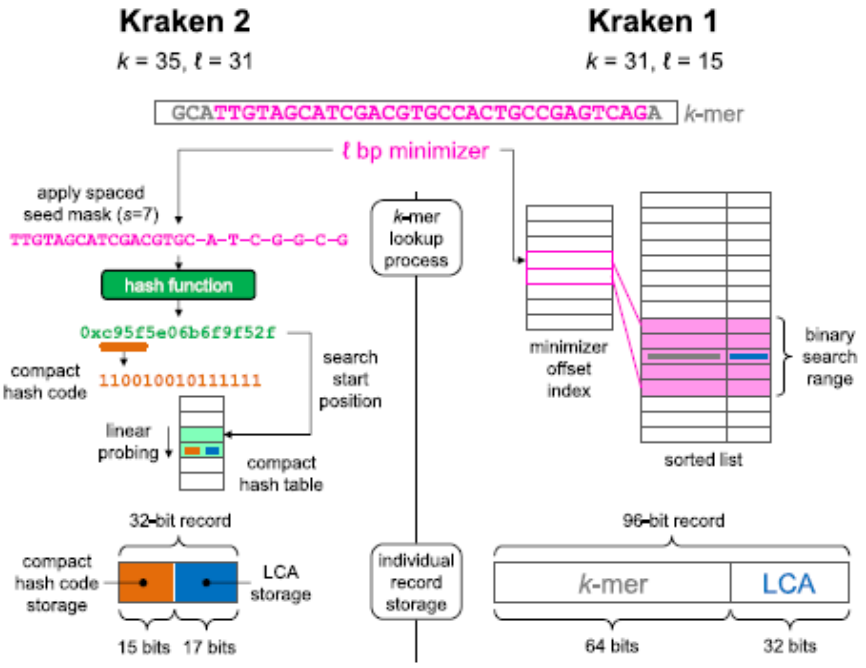



Figura 4: Algoritmo do classificador Kraken, mostrando as diferenças entre a versão 1 e 2 (WOOD *et al.,* 2019).

### 3.2.2. Kaiju

Outro classificador taxonômico muito usado na análise metagenômica, o Kaiju, foi escrito em C/C + + por Peter Menzel e Anders Krogh no ano de 2016. O classificador, utiliza, ao contrário do Kraken e de vários outros classificadores, a base de proteína ao invés da base de nucleotídeos, pois as sequências de proteínas são mais conservadas do que o DNA e os genomas microbianos e virais são tipicamente embalados com genes codificadores de proteínas (MENZEL *et al.,* 2016). Diferente dos outros classificadores, o Kaiju procura por correspondências exatas máximas (*Maximum Exact Matches* - MEM's) da sequência de aminoácidos em um determinado banco de dados de proteínas como referência, para assim alcançar alta precisão e sensibilidade.

Como os *reads* são sequências de nucleotídeos, o que é feito pelo Kaiju inicialmente é traduzir os *reads* em quadros de leitura de aminoácidos, que são divididos em fragmentos nos *stop* códons, os fragmentos são então ordenados por tamanhos e, a partir do maior,

começa a consulta no banco de dados utilizando a busca inversa do algoritmo *Burrow-Wheeler Transform* (BWT) para encontrar os MEM's; aqueles de maior comprimento são retidos e quando a busca se encerra, o identificador do táxon da sequência de banco de dados correspondente é recuperado da matriz de sufixos e impresso na saída (MENZEL *et al.,* 2016). O método BWT é um processamento estatístico que utiliza a teoria de que dado um símbolo, há uma grande probabilidade desse símbolo ser sempre precedido do mesmo conjunto de símbolos (BURROWS; WHEELER, 1994); a mesma teoria é aplicada no alinhamento.

Sua documentação está disponível em <https://github.com/bioinformatics-centre/kaiju>. (Acesso em 05 de Junho de 2022).

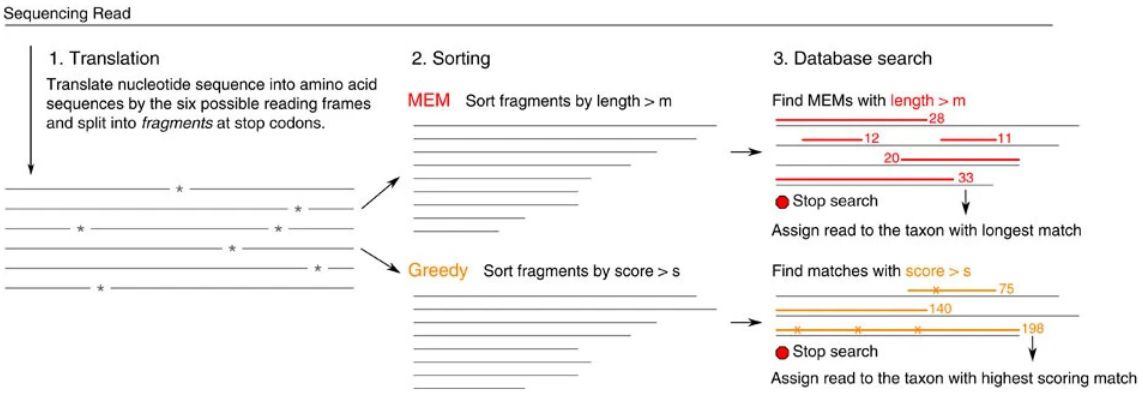



Figura 5: Algoritmo do classificador Kaiju (MENZEL *et al.,* 2016).

### 3.2.3. DIAMOND

O *Double Index AlignMent Of Next-generation sequencing Data* ou DIAMOND é um classificador taxonômico que usa um banco de dados proteico, assim como o Kaiju, escrito em C++ por Benjamin Buchfink e Daniel H. Husonque (Departamento de Ciência da Computação e Centro de Bioinformática, Universidade de Tübingen - Alemanha). Surgiu para substituir o BLASTX (considerado como padrão ouro em alinhar leituras contra um banco de dados proteico) no quesito de alto rendimento (BUCHFINK *et al.,* 2015). Para tanto, utiliza da abordagem de semeadura e extensão junto com algoritmos

adicionais de ganho de performance como redução do alfabeto, sementes espaçadas e, como o nome diz, dupla indexação (BUCHFINK *et al.,* 2015).

O programa faz uso do paradigma de semeadura e extensão, que consiste em utilizar uma semente (*seed* - pequenas palavras de comprimento fixo) para encontrar correspondências exatas na sequência de referência, caso encontre, a semente é estendida, até, se possível, achar a sequência por completo; enquanto a semente está pequena, há aumento na sensibilidade, conforme ela começa a crescer, a velocidade aumenta. Para aumentar ainda mais a velocidade, o DIAMOND também utiliza um alfabeto reduzido composto por 11 palavras: [KREDQN] [C] [G] [H] [ILV] [M] [F] [Y] [W] [P] [STA]; e para aumentar a sensibilidade, utiliza-se sementes espaçadas, que consistem em sementes mais longas nas quais apenas um subconjunto de posições é usado. Por padrão, são 4 sementes de tamanho variado entre 15 e 24 caracteres com peso 12 (Figura 6), é o peso que indica o número de posições (BUCHFINK *et al.,* 2015).

A dupla-indexação diz respeito em, tanto a semente quanto a sequência de referência ser indexada, ou seja, ambas são ordenadas lexicograficamente, permitindo o cálculo do alinhamento local nas localizações de sementes correspondentes e reduzindo o uso de memória (BUCHFINK *et al.,* 2015).

Sua documentação está disponível em <https://github.com/bbuchfink/diamond/wiki>. Acesso em: 16 de Junho de 2022.

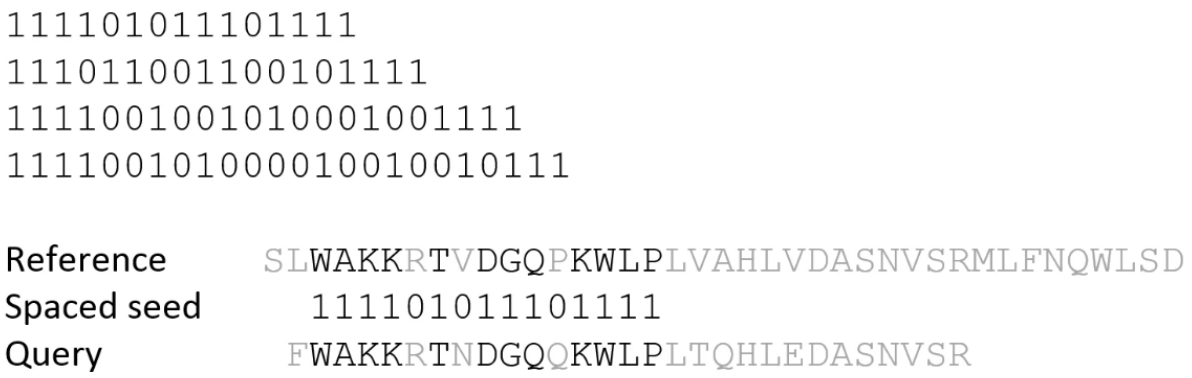



Figura 6: Algoritmo do classificador DIAMOND (BUCHFINK *et al.,* 2015).

Para o melhor entendimento da classificação taxonômica gerada pelo DIAMOND, recomenda-se utilizar, em conjunto, o software MEGAN (que se encontra, até o momento em sua 6ª versão), desenvolvido, também, pela Universidade de Tübingen e escrito por Daniel H. Huson (Bağcı *et al.,* 2021). O MEGAN é capaz de ler o arquivo gerado pelo DIAMOND (formato .daa) e associar os ID 's aos nomes científicos do organismo, presente no banco de dados, pela função *meganizer*. Tanto o manual quanto o tutorial se encontram em <Megan6>

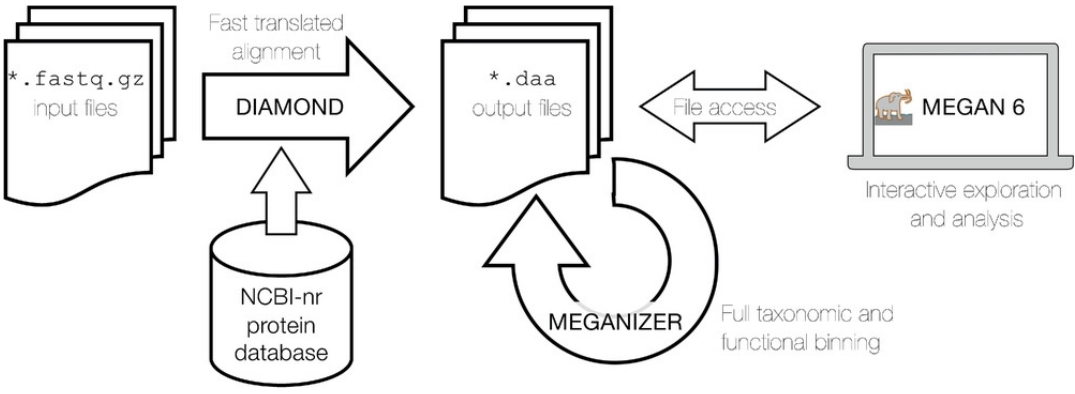



Figura 7: *Pipeline* conjunto DIAMOND-MEGAN6 (Bağcı *et al.,* 2021).

### 3.2.4. CLARK

O CLARK (ou *CLAssifier based on Reduced K-mers*) é um classificador taxonômico que utiliza banco de dados de nucleotídeos, desenvolvido pelo Departamento de Engenharia e Ciência da

Computação da Universidade da Califórnia, em 2015 (OUNIT *et al.,* 2015).

Assim como o Kraken, também utiliza, como padrão, o banco CLARKde dados RefSeq do NCBI e uma tabela *hash* para fazer a busca (OUNIT *et al.,* 2015). O que o diferencia do Kraken, e de outros classificadores, é o fato de usar *K-mers* discriminativos, portanto não utilizar todos os *K-mers*. O CLARK funciona indexando as sequências-alvo (*inputs*), ou seja, armazenando-as em uma tabela *hash*, para cada *K-mer* distinto $w$, as seguintes informações: o ID do alvo, o número de alvos distintos e o número de ocorrências de $w$. Assim, dado um número inteiro $k$ e $m$ genomas de referência $\{g1, g2, \ldots, gm\}$, o *K-mers* discriminativos $wi$ para o genoma $gi$ são o conjunto de todos os *K-mers* em $gi$ que não ocorrem (exatamente) em nenhum outro genoma, removendo qualquer *K-mer* que tenha ocorrido em outro alvo; deixando o CLARK mais sensível e com menos uso de memória (OUNIT *et al.,* 2015; OUNIT e LONARDI, 2016).

Sua documentação está disponível em <http://clark.cs.ucr.edu/>. Acesso em 16 de Junho de 2022.

### 3.2.5. qPCR

Para as análises de Correlação foi realizada amplificação por meio de PCR de tempo real para obtenção do valor valor de Ct, ou *Ct-value (Cycle-Threshold),* também chamado de valor de Cq (*Cycle of Quantification*) para os principais vírus identificados através da utilização dos classificadores. Essa técnica foi escolhida pois detecta através do aumento da fluorescência amostras positivas e o parâmetro Ct é diretamente relacionada a sua carga viral cujo valor foi utilizado na análise de regressão. Por outro lado, através da confirmação molecular teremos uma ideia se de fato os vírus classificados taxonomicamente são verdadeiros ou podem ser referidos como artefatos. Por esse motivo é considerado o Padrão Ouro nas análises de detecção de organismos nas amostras (ALI, 2020).

O valor de Ct é o número de ciclos necessários para o início da amplificação da sequência gênica-alvo presente no DNA de cada amostra, ou seja, é o ponto obtido quando a amplificação positiva (o ciclo) cruza o

*threshold* (limiar), por isso o nome *Cycle Threshold* ou *Ct*. O Ct é diretamente proporcional ao logaritmo da quantidade inicial do gene-alvo em uma determinada amostra, assim, quanto menor for o número inicial do Ct obtido na amostra, maior a quantidade deste gene na amostra que também pode ser referida como carga viral (ALI, 2020).

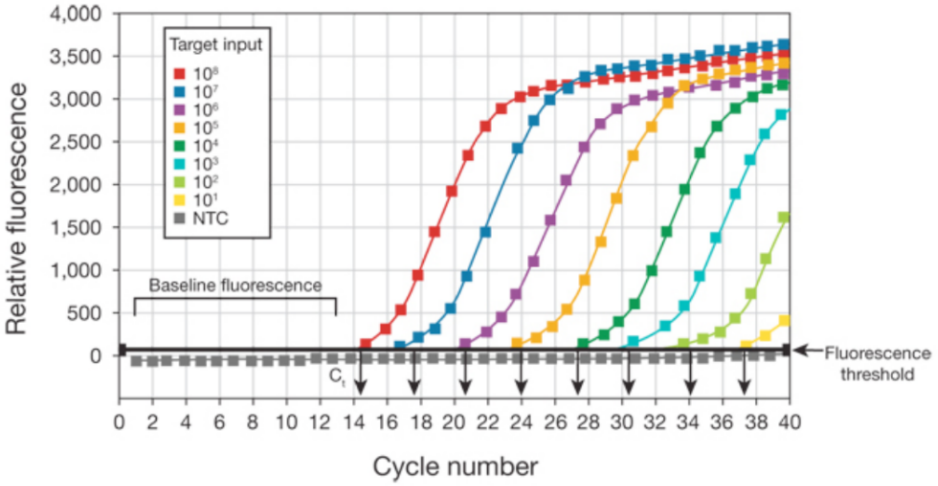



Figura 8: Corrida de amplificação, amplification plot em inglês de cargas virais decimais. Esquematizando o gráfico de amplificação do PCR em tempo real . No esquema, o *threshold* é representado pela linha cinza e no momento em que esse limite é ultrapassado, é obtido o valor de Ct. No caso, o gene-alvo cuja fluorescência está em vermelho apresenta menor Ct e, consequentemente, maior quantidade na amostra. Fonte: *Real Time PCR Handbook ThermoFisher Scientific*

## 3.3. Montagem do Genoma Completo

Os reads classificados serão montados em sequências maiores chamados de contigs com a finalidade de obter dados mais aprofundados sobre o genoma dos vírus identificados através de análise filogenética.. Essa etapa permite corrigir erros de leituras únicas e reduzir o tamanho dos dados; para a montagem, pode-se usar um genoma de referência ou fazer a montagem *de novo* (NOOIJ, *et al.,* 2018).

Um dos programas mais recomendados para a montagem é o SPAdes (St. Petersburg genome assembler) (BANKEVICH *et al*., 2012; NOOIJ, *et al.,* 2018). O SPAdes surgiu inicialmente para suprir a necessidade de entender as bactérias em diversos projetos, como o Projeto Microbioma Humano e descoberta de antibióticos (BANKEVICH *et al*., 2012); porém, com o passar do tempo, o SPAdes ampliou para se especializar em outros microrganismos, como fungos e vírus.

As etapas do montador são a criação de grafos de Bruijn multidimensionais com valor de $K$ para as arestas e $K-1$ para vértices que detectam protuberâncias e evitam a formação de leituras quiméricas. O SPAdes estima a distância dos K-mers e em seguida aglutina (para maiores detalhes observar marcação em vermelho na figura 9.A) os vértices, resultando em 9.B, obtendo $h$ caminhos possíveis (no caso $h=3$). A figura 9.C mostra, no círculo exterior, cada um dos 9 reads de 4-mers ($K=4$), no próximo círculo mais interior, o grafo de Bruijn e no mais interior o genoma de referência que deseja montar. O grafo multidimensional é mostrado em 9.D.

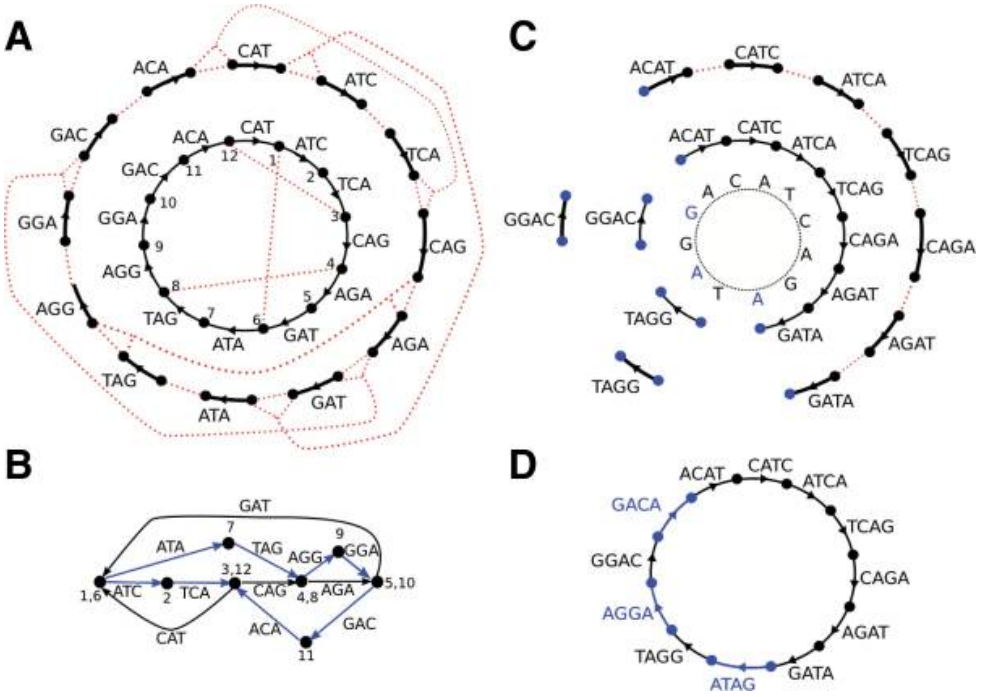



Figura 9: Etapas do SPAdes (BANKEVICH *et al*., 2012)

Sua documentação está disponível em <https://github.com/ablab/spades>. Acesso em 11 de Junho de 2022.

## 3.4. Análise Filogenética

Com o genoma dos vírus de interesse montados a partir das leituras obtidas, a última parte do projeto incluiu a análise filogenética para entender a epidemiologia molecular de vírus com interesse clínico. Os contigs obtidos de diversos *Pools* foram alinhados entre si, portanto um alinhamento múltiplo, utilizando o programa MAFFT v.7.455, escrito na linguagem C e que utiliza

Transformadas Rápidas de Fourier cuja eficiência é de O(N log N); reduzindo o tempo de CPU quando comparado com outros programas, como por exemplo o ClustalW (KATOH *et al.,* 2002). As árvores filogenéticas serão reconstruídas a partir do modelo Maximum Likelihood utilizando o software IQ-Tree (NGUYEN *et al.*, 2015). A vizualização das árvores sua edição será realizada utilizando tanto o *software* FigTree, disponível em <https://github.com/rambaut/figtree> e o pacote ggtree, escrtito em R, do Bioconductor, disponível em <https://bioconductor.org/> e <https://bioconductor.org/packages/release/bioc/html/ggtree.html>, ambos com acesso dia 11 de Junho de 2022, escritos em R (YU, 2020).

O IQ-Tree aplica o método de máxima verossimilhança, ou seja, utiliza um modelo probabilístico de evolução para a montagem de sua árvore capaz de corrigir vários eventos mutacionais e, portanto, com menos erros de amostragem (CALDART, *et al.,* 2016). Desse modo, as árvores geradas representam alta probabilidade e com menos tempo de geração (NGUYEN *et al.*, 2015).

O comando para o alinhamento é **mafft** --auto --inputorder "viruses.fasta" > "viruses_aln.fasta". O parâmetro --auto indica que o próprio programa irá escolher a melhor opção para o mapeamento. Para a geração do arquivo TREE, o comando é "**iqtree** -s viruses_aln.fasta -lmap 1000 -bb 1000 -m MFP"; o comando indica o *input* (-s), o número de pontos que serão desenhados no mapeamento de verossimilhança (-lmap), o modelo (-m) automático, MFP e o valor de *bootstrap* (-bb) escolhido foi 1000.

## 3.5. *Pipeline* I

A pipeline inicial utilizada foi baseada na pipeline de metagenômica viral do Instituto Craig Venter nos EUA. Ela foi utilizada, também, como treinamento dos programas e do uso da Linha de Comando. A mesma contou com os seguintes programas, na ordem descrita:

- fastp para o controle de qualidade das sequências;
- bowtie2 para o mapeamento das leituras humanas (*NCBI* - GRCh38) e
- Kaiju, inicialmente, para a classificação taxonômica das leituras não mapeadas.

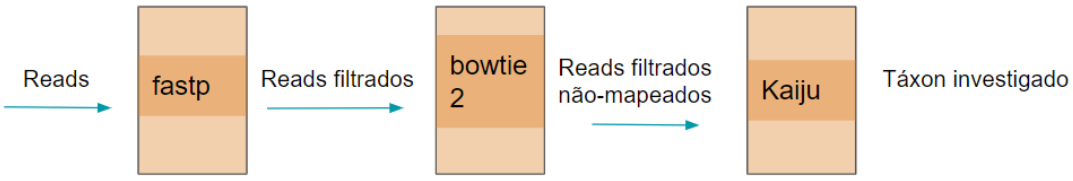



Figura 10: Programas utilizados na *Pipeline* I, indicando seus *inputs* e *outputs*.

### 3.5.1. Linha de Comando

**fastp** -i rawData/Pool_R1.fastq.gz -o Pool_trimmed_R1.fastq -I rawData/Pool_R2.fastq.gz -O Pool_trimmed_R2.fastq

**bowtie2** -x /reference/human_as -1 Pool_trimmed_R1.fastq -2 Pool_trimmed_R2.fastq Pool_map.sam --un-conc Pool_reads_unmapped.fastq

**kaiju** -t nodes.dmp -f kaiju_db_viruses.fmi -i Pool_reads_unmapped.1.fastq -j Pool_reads_unmapped.2.fastq -o Pool_reads_kaiju.out

**kaiju2table** -t nodes.dmp -n names.dmp -r species -u -e -o Pool_kaiju.tsv Pool_reads_kaiju.out

## 3.6. Amostragem

Para a realização do projeto, os dados brutos foram obtidos de sequenciamentos já realizados pelo sequenciador NextSeq 2000, do Laboratório Estratégico de Sequenciamento de SARS-CoV-2, Instituto Butantan, armazenados no banco de dados do mesmo, o Illumina BaseSpace. Foram incluídos dados brutos dos seguintes grupos de interesse na forma de *Pools*: amostras de soro obtidas de casos não identificados de infecções arbovirais, grupo de amostras de pacientes com câncer de próstata e amostras de crianças com sintomatologia respiratória aguda porém com resultado negativo para SARS-CoV-2. Os *Pools* estão divididos em:

- Grupo 1:
  - Pools 2 e 3: Amostras de plasma de pacientes com câncer de próstata do Hospital Erasto Gaertner da Universidade Federal do Paraná;

- Pool 25: Amostra de pool de soros de pacientes com infecções arbovirais desconhecidas provenientes do Laboratório Central ( LACEN) de Alagoas em Maceió.

- Grupo 2:
  - Pools 1 ao 15: de amostras de *swab* bucal de pacientes pediátricos com sintomatologia respiratória aguda porém com resultado negativo para SARS-CoV-2

### 3.7. Regressão Linear dos Classificadores

Para a avaliação do melhor classificador, foi realizada a correlação entre o logaritmo natural do número de *reads* obtido pelos 4 classificadores pelo logaritmo do valor de Ct. O método linear para a correlação foi a técnica de Regressão Linear, conforme CARBO *et al.,* 2022 e foi avaliada utilizando o Coeficiente de Determinação ou Explicativo ($R^2$).

A regressão linear é utilizada para estimar um valor esperado de uma variável dependente (y) a partir de uma ou várias variáveis independentes (Xi), assume-se que a relação da resposta às variáveis é uma função linear (MORETTIN *et al.,* 2017). Sua fórmula é $y_i = \alpha + \beta X_i + \varepsilon_i$, onde:

$y_i$: variável dependente, a qual o modelo tentará prever;

$\alpha$: o intercepto da reta com o eixo vertical quando x = 0;

β: coeficiente angular;

Xi: variável independente e

εi: resíduos e erros de medição.

O coeficiente de determinação é a medida de ajuste de um modelo estatístico linear generalizado (como no caso, a regressão linear); seu valor varia entre 0 e 1, quanto mais próximo de 1, mais explicativo é o modelo, ou seja, melhor se ajusta à amostra (MORETTIN *et al.,* 2017).

## 4. RESULTADOS

No desenvolvimento do projeto foi realizada a comparação visual, em forma de Barplot, entre os 4 classificadores, utilizando tanto a Linguagem R quanto Python, comparando os resultados de abundância. Além disso, um gráfico de correlação, feito em R, apresentado por Regressão Linear do classificador com o Padrão Ouro, ou seja, resultados obtidos através de qPCR (CARBO, *et al.,* 2022).

## 4.1. *Pipeline* I

A primeira *pipeline* foi utilizada com os programas Kaiju e Kraken2 para efeito de comparação (como será mostrado a seguir). Da forma que foi executada a pipeline, nós observamos que o controle de qualidade não foi muito rigoroso, uma vez que não foram obtidas informações sobre as taxas de contaminação dos dados brutos antes de ser executado o fastp, ou seja, informações referentes à adaptadores, fitas poli-A e poli-N, entre outros contaminantes possíveis. Ao mesmo tempo não é feita a verificação se há melhora nas leituras após o uso do fastp.

O mapeador bowtie2, utilizado nessa pipeline utiliza o genoma humano (*NCBI* - GRCh38) como referência para o mapeamento, entretanto, pela grande presença de leituras classificadas erroneamente com o Kaiju (Figura 11.A), é possível que haja erros no mapeamento, deixando passar leituras humanas e causando erros de classificação.

### 4.1.1. Abundância Viral *Pipeline* I

Na execução da *pipeline*, foi utilizado o classificador Kaiju, cujo banco de dados foi o *NCBI - Viruses,* disponível em <https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239>, dessa maneira, apenas táxons virais estão presentes na base de dados e somente eles são identificados. Para avaliar a precisão de análise e a comparação dos dois classificadores utilizando essa pipeline foram usados dois Pools: i) um pool de amostras de plasma obtidas de pacientes com perfil indeterminado de amplificação por PCR de doenças arbovirais (Dengue, Zika e Chikungunya) do estado de Alagoas - Brasil, o Pool 25 e ii) um pool de amostras respiratórias de pacientes pediátricos com sintomas agudos de infecção respiratória porém negativas para SARS-CoV-2, o Pool 7.

Nessa análise comparativa do pool de amostras de plasma de Alagoas, a pipeline com o programa Kaiju gerou resultados pouco relevantes em aspecto clínico (Figura 11.A). Os vírus mais abundantes no Pool 25 foram o *BeAn 58058 virus*, conhecido por infectar roedores brasileiros do gênero *Oryzomys* e Macacos-do-Velho-Mundo (*Chlorocebus aethiops*) (WANZELLER, *et al.,* 2017), o

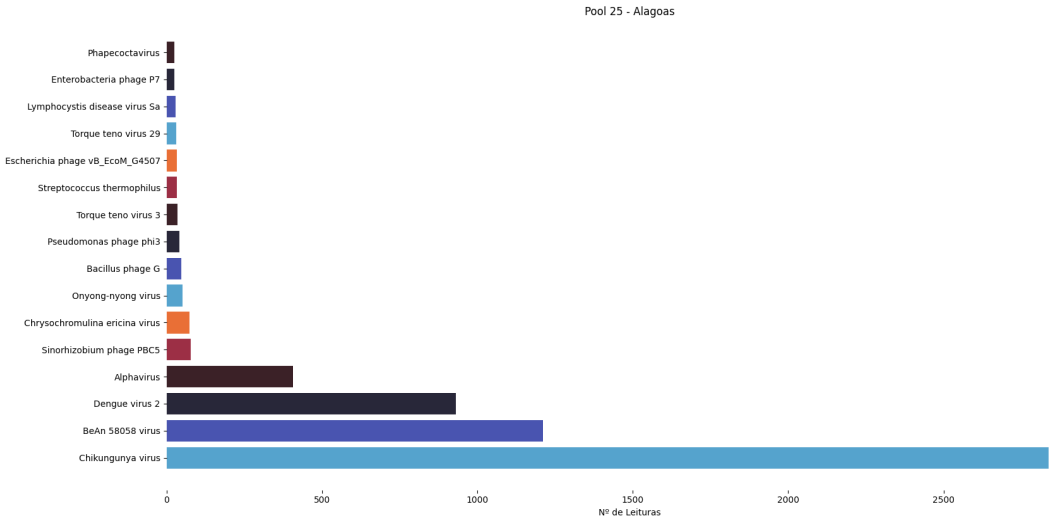
*Chrysochromulina ericina virus,* que infecta microalgas marinhas (GALLOT-LAVALÉE *et al.,* 2017) e o *Lymphocystis disease virus* que causa doença de pele em peixes (LABELLA *et al.,* 2019). No Pool 7, também houve a presença do *BeAn 58058 virus*, múltiplos fagos, *Lymphocystis disease virus* e *Glypta fumiferanae ichnovirus*, cujo hospedeiro são as vespas parasitas (BIGOT, *et al.,* 2008). Apesar desses resultados, como também houve a detecção dos vírus da Dengue 2 e da Chikungunya no Pool 25 e *Enterovírus* no Pool 7, pode se indicar que a análise não foi tão precária e a pipeline apenas precise apenas de algumas melhoras.

Diante desses resultados, foi necessário a elaboração de outra *pipeline* visto que a primeira é básica e sem a informação sobre o controle de qualidade. Além disso, a classificação demonstrou vírus sem importância clínica, principalmente no Pool 25, possivelmente errôneas advindas da alta presença de material genético humano não filtrado e outros contaminantes. Para verificar se houve a presença de alguma leitura humana, foi feita a análise desse Pool com o Kraken2 devido ao seu banco de dados (*NCBI - RefSeq*, disponível em <https://ftp.ncbi.nlm.nih.gov/genomes/refseq/>) conter táxons de *Homo sapiens*. Após a confirmação de contaminação do material genético do hospedeiro, foi optado por incluir um mapeador mais robusto utilizando outro programa de alinhamento, o BWA.

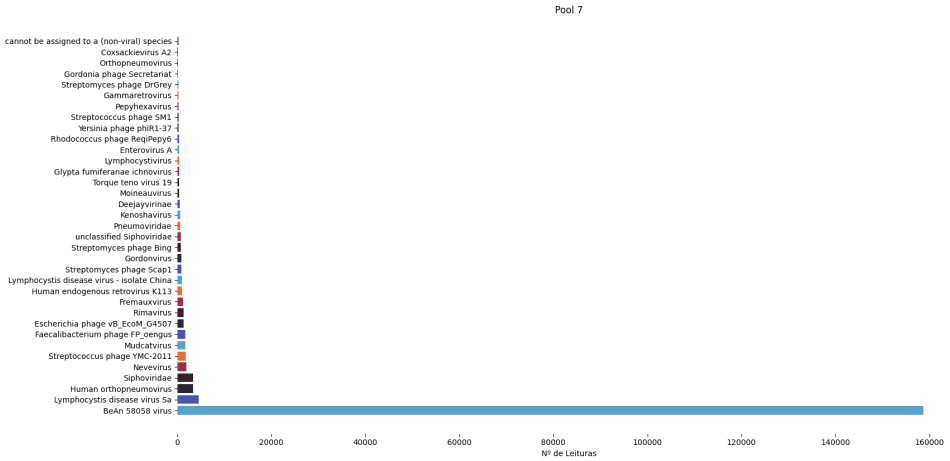
Para melhor representação visual, os vírus foram filtrados entre aqueles com leituras maiores que 25 pb no Pool 25, obtendo 16 resultados, e maiores que 200 no Pool 7, obtendo assim, 35 resultados.

Os gráficos foram montados com Python e contou com as bibliotecas pandas (https://pandas.pydata.org/), matplotlib (https://matplotlib.org/) e seaborn (https://seaborn.pydata.org/).

A. Pool 25 - Kaiju




B. Pool 7 - Kaiju




C. Pool 25 - Kraken2

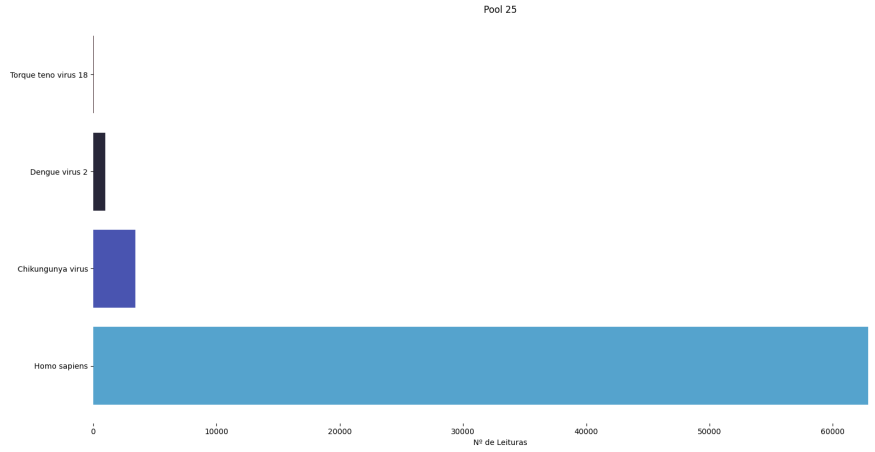



Figura 11. Barplots de abundância viral utilizando como amostra de referência um Pool de plasma obtido de amostras com amplificação inconclusiva por PCR de arboviroses circulando no estado de Alagoas (Pool 25) e outro utilizando como amostras de referência amostras pediátricas negativas para SARS-CoV-2 porém com sintomatologia aguda respiratória (Pool 7). Nós podemos observar a grande diferença tanto na abundância quanto nas espécies classificadas. Com a utilização do programa Kaiju nós tivemos grandes quantidades de vírus como BeAn em ambos os Pools, Chrysochromulina e múltiplos fagos, destes nenhum com importância clínica. A figura C confirmou que há contaminação de leituras humanas (azul) na amostra e portanto deveriam haver mudanças na *pipeline*, principalmente do mapeador.

## 4.2. Edição da *Pipeline*

Para suprir a necessidade de ter informações mais robustas sobre o *trimming* das sequências e do controle de qualidade foi utilizado o FastQC antes de depois do uso do fastp. Junto a isso, após a leitura da documentação do fastp, novos parâmetros foram incorporados:

- -q: valor da qualidade das leituras, definido em 30;
- -g e -x : detectar sequências poli-G e poli-X, respectivamente, e removê-las;
- -c: correção de regiões sobrepostas e
- -h: gerar *report* em formato .html

Aplicando assim um controle melhor sobre a qualidade; na segunda pipeline foi adicionado o programa SPAdes e com ela foi realizada a comparação entre os 4 classificadores propostos, conforme mostra a Figura 12. O banco de dados utilizados em cada classificador é descrito a seguir:

- Kaiju: *NCBI - Viruses,* disponível em <https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239>;
- Kraken2: *NCBI - RefSeq*, que contém informações sobre vírus, humanos, archeas e bactérias; disponível em <https://ftp.ncbi.nlm.nih.gov/genomes/refseq/>;
- DIAMOND: banco de dados de proteínas virais do NCBI, o mesmo do BLAST, disponível em <https://ftp.ncbi.nlm.nih.gov/blast/db/>, identificado com o prefixo nr e
- CLARK: o *NCBI - RefSeq Viruses*; recuperado pelo próprio programa utilizando o comando *./set_targets.sh DIR_DB viruses.*

  Todos os bancos foram acessados em Junho de 2022.

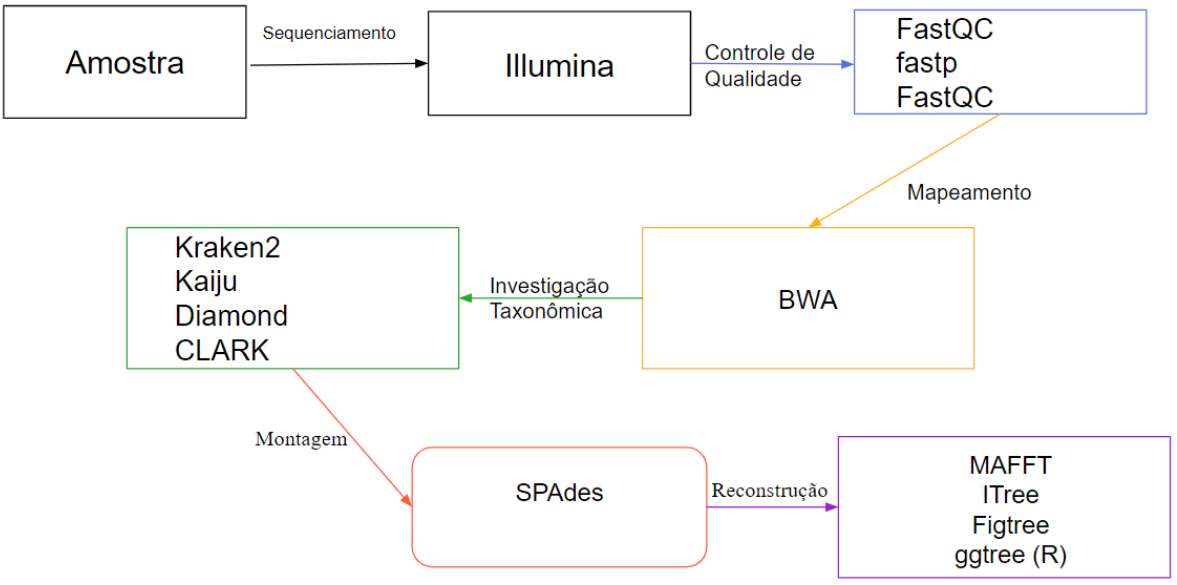



Figura 12: Representação da nova *pipeline*, indicando os programas em cada etapa. Os blocos coloridos indicam as etapas realizadas por esse projeto

### 4.2.1. Linha de Comando

A linha de comando foi escrita em linguagem *bash* e executada no terminal do Linux.

**fastqc** rawData/Pool_R1.fastq.gz /rawData/Pool_R2.fastq.gz -o /QC/Reports/before

**fastp** -i /rawData/Pool_R1.fastq.gz -o /QC/Pool_trimmed_R1.fastq.gz -I /rawData/Pool_R2.fastq.gz -O QC/Pool_trimmed_R2.fastq.gz -q 30 -g -x -c -h QC/after/fastp.html

**fastqc** /QC/Pool_trimmed_R1.fastq.gz /QC/Pool_trimmed_R2.fastq.gz -o /QC/Reports/after

**bwa** mem -P genomes/Human/GRCh38_latest_genomic.fna /QC/Pool_trimmed_R1.fastq.gz QC/Pool_trimmed_R2.fastq.gz > Pool_mapping.bam

**samtools** view -b -f 4 /mapping/Pool_mapping.bam > /mapping/Pool_unmapped.bam

**samtools** fastq -1 Pool_unmapped.R1.fastq.gz -2 Pool_unmapped.R2.fastq.gz /mapping/Pool_unmapped.bam

**kraken2** -db /database/kraken2/db/ --paired /mapping/Pool_unmapped.R1.fastq.gz /mapping/Pool_unmapped.R2.fastq.gz --report Pool_kraken.tsv

**kaiju** -z 12 -t nodes.dmp -f kaiju_db_viruses.fmi -i Pool_unmapped.fastq.gz -o Pool_kaiju.out

**kaiju2table** -t nodes.dmp -n names.dmp -r species -u -e -o Pool_kaiju.tsv Pool_kaiju.out

**./classify_metagenome.sh** --gzipped -n 12 -O mapping/Pool_unmapped.fastq.gz -R Pool_3_clark

**./estimate_abundance.sh** -F /CLARK/Pool_clark.csv -D /database/clark_db/Viruses/ --mpa

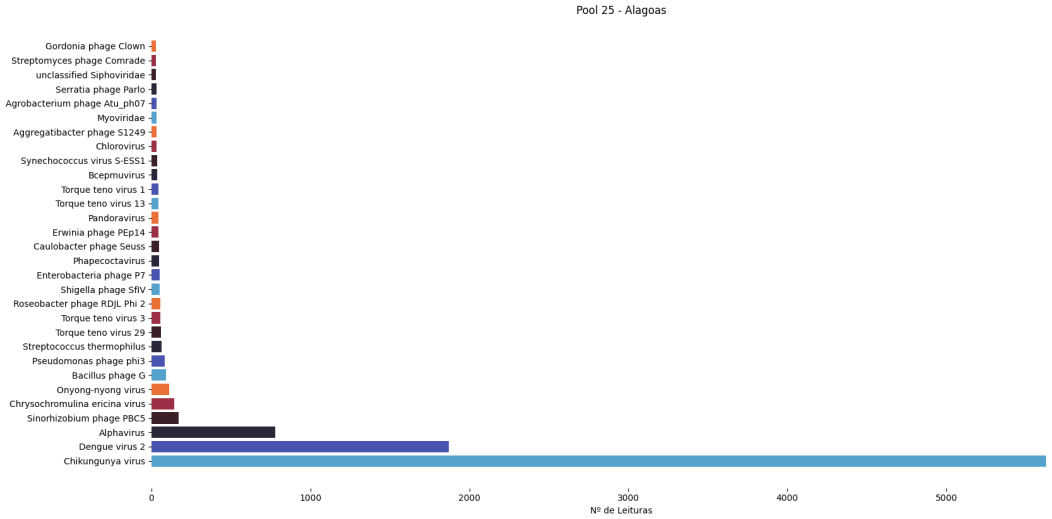
**diamond** blastx -d /database/diamond_db/viruses.dmnd --verbose --outfmt 100 -q /mapping/Pool_unmapped.fastq.gz -o Pool_Diamond.daa/

### 4.2.2. Abundância Viral

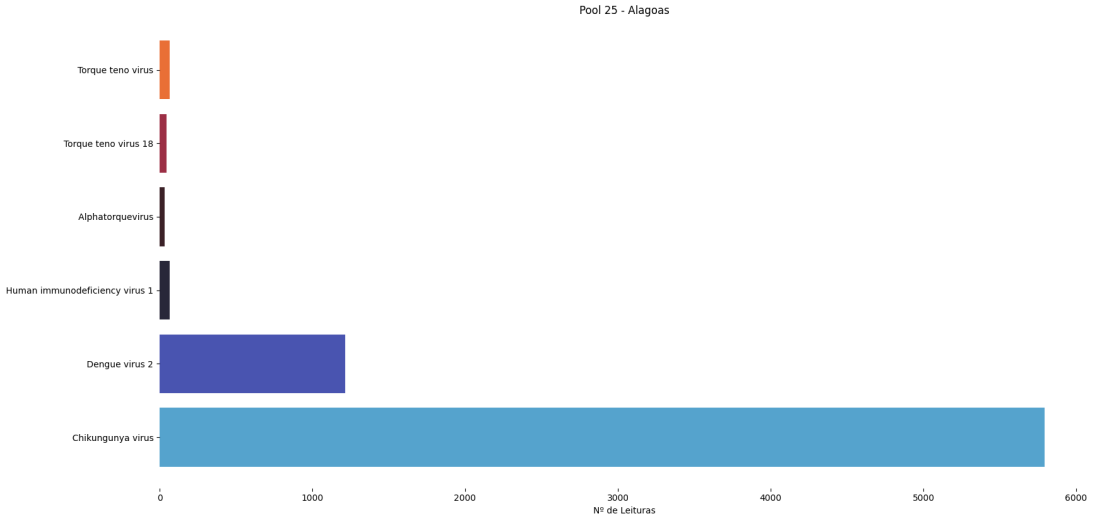
Com essa nova pipeline foi feita uma análise dos Pools 25 - Grupo 1 e 7 - Grupo 2 para verificar se houve melhorias ao executar mudanças na *pipeline* e uma análise comparativa dos classificadores Kraken 2 e Kaiju no Pool 7 do Grupo 2, apenas para verificar se houve alguma divergência dos vírus identificados. Os resultados são

apresentados nas figuras 13.A e 13.B, referentes ao Pool 25, e 13.C e 13.D, referentes ao Pool 7.

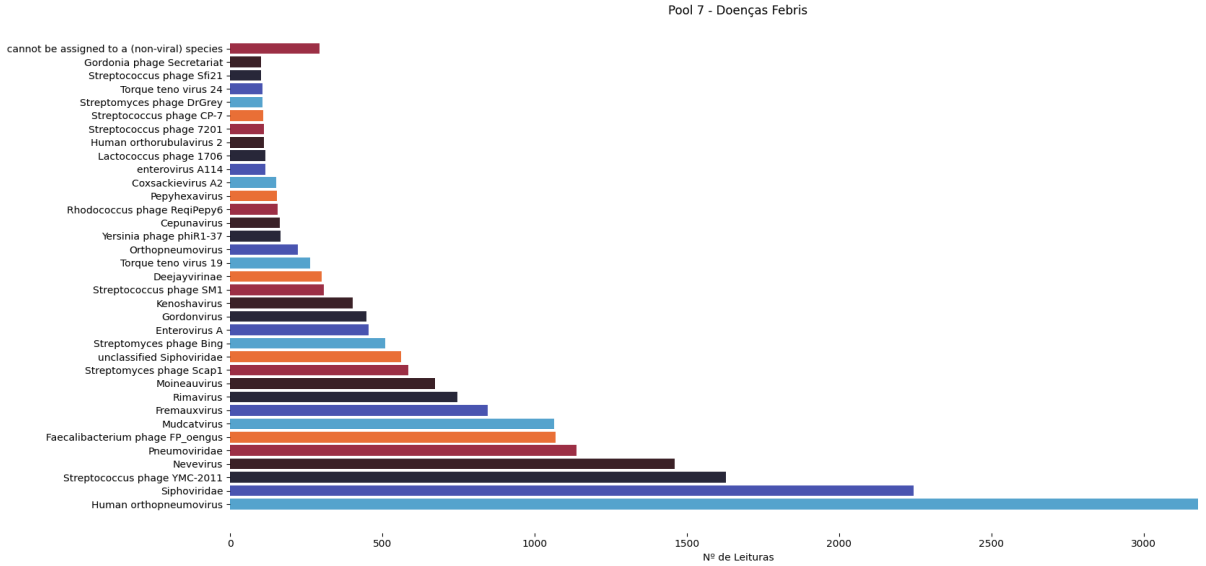
A. Pool 25 - Kaiju - nova *pipeline*




B. Pool 25 - Kraken2 - nova *pipeline*




C. Pool 7 - Kaiju - nova *pipeline*




D. Pool 7 - Kraken2 - nova *pipeline*

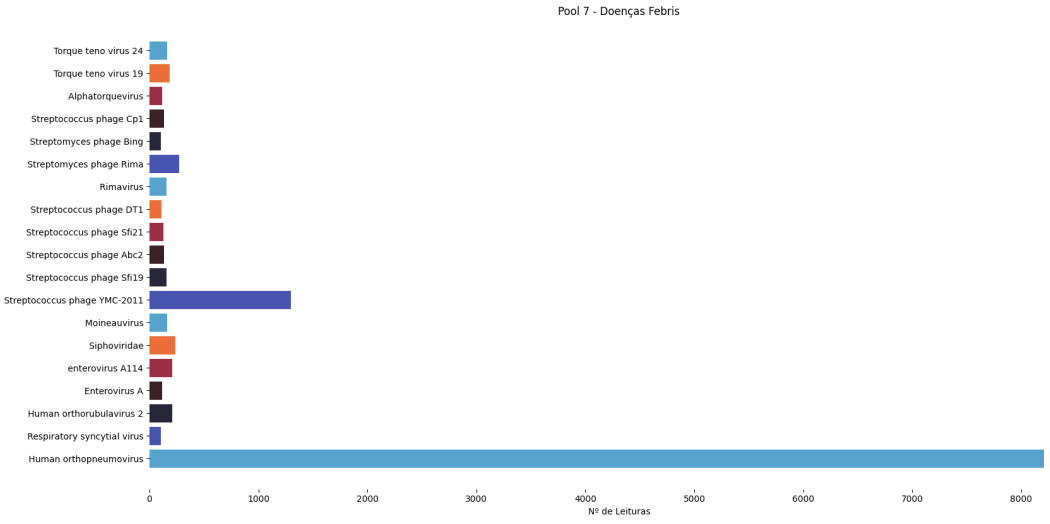



Figura 13: Barplots de comparação entre os Classificadores utilizando como amostras de referência amostras pediátricas negativas para SARS-CoV-2 porém com sintomatologia aguda respiratória (Pool 7) e amostras com amplificação inconclusiva por PCR de arboviroses circulando no estado de Alagoas (Pool 25). (A) Com a nova *pipeline*, os vírus *BeAn 58058* e o *Lymphocystis disease virus* deixaram de ser classificados, porém ainda houve classificação do *Chrysochromulina ericina virus* e do *Chlorovirus*, cujo hospedeiros também são algas (VAN ETTEN *et al.,* 2022). *O chikungunya virus* continuou sendo o vírus sendo mais abundante junto com o vírus da Dengue 2. É possível perceber a

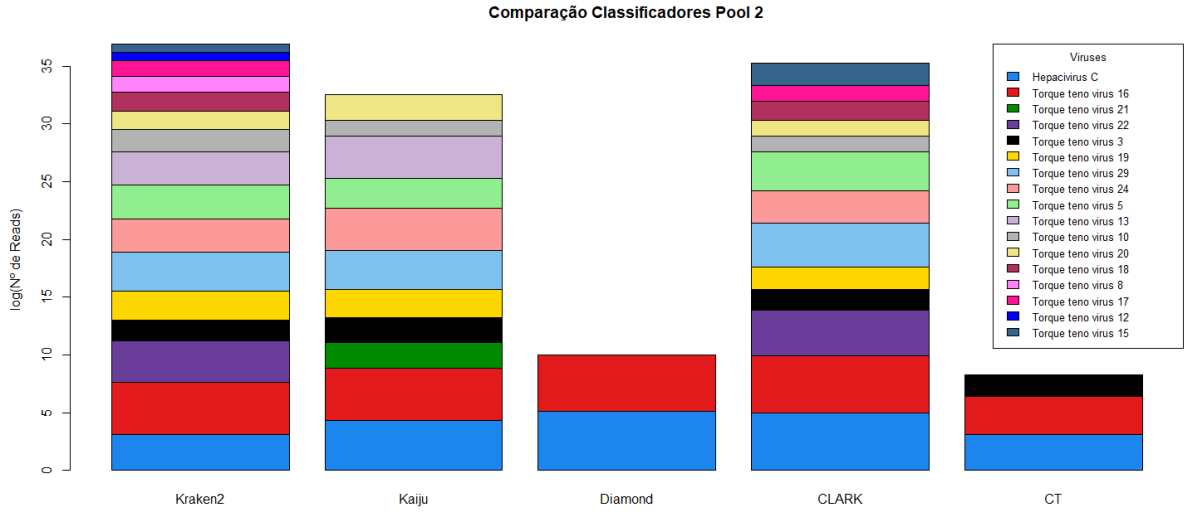
grande riqueza de vírus encontrados. (B) Utilizando a nova pipeline e o Kraken2, houve redução significativa no número de vírus com leituras superiores a 25. Dessa vez, os vírus da Dengue 2 e da Chikungunya foram os mais abundantes junto com os Torque Teno Vírus (TTV); a presença do *Human Immunodeficiency virus* (HIV) pode ter ocorrido devido à integração de elementos do HIV ao DNA humano (LUGANINI e GRIBAUDO, 2020). (C) (D) Utilizando o Kraken2 para a análise, houve redução na riqueza dos vírus, porém os vírus classificados possuem maior relevância clínica, como o respiratório sincicial, enterovírus A e A114 e orthorubulavirus.

Por conta dos resultados obtidos, foi escolhida para a análise comparativa dos classificadores, a nova *pipeline,* apelidada de *Pipeline* II.

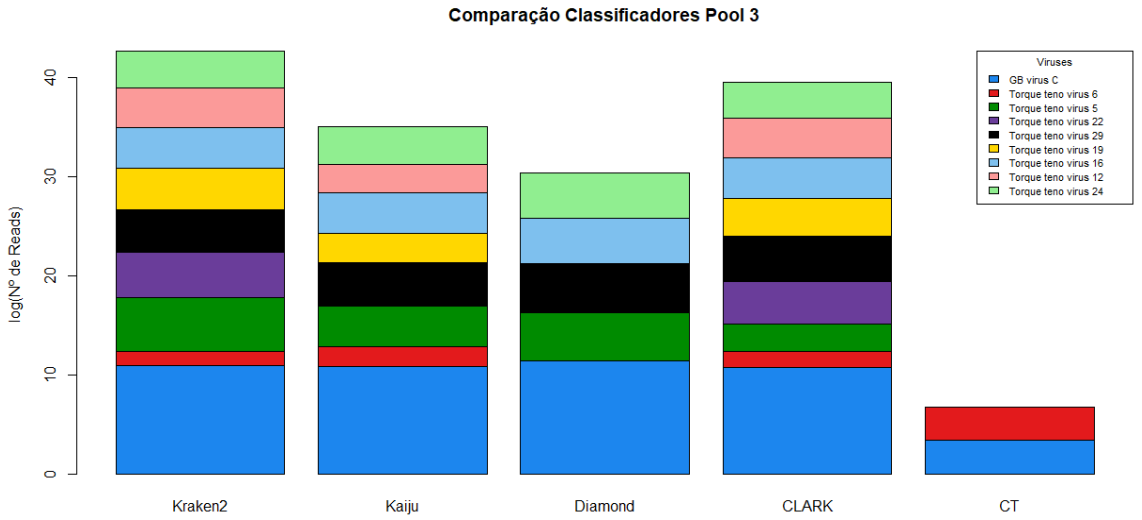
## 4.3. Comparação Visual dos Classificadores

Para a comparação dos quatro classificadores, foram escolhidos os Pools 2, 3 e 25 do Grupo 1 e os Pools 8 e 9 do Grupo 2, sendo que, houve confirmação molecular de alguns vírus do Grupo 1 (indicados pelo valor de Ct). Também foram filtrados vírus de interesse clínico para saber como os classificadores se comportavam. Por conta da grande diferença de leituras (variando por exemplo entre 331000 para 200 leituras em um único classificador) foi escolhida a normalização logarítmica natural, $log_e(Nº\ de\ leituras)$, padrão da linguagem R.

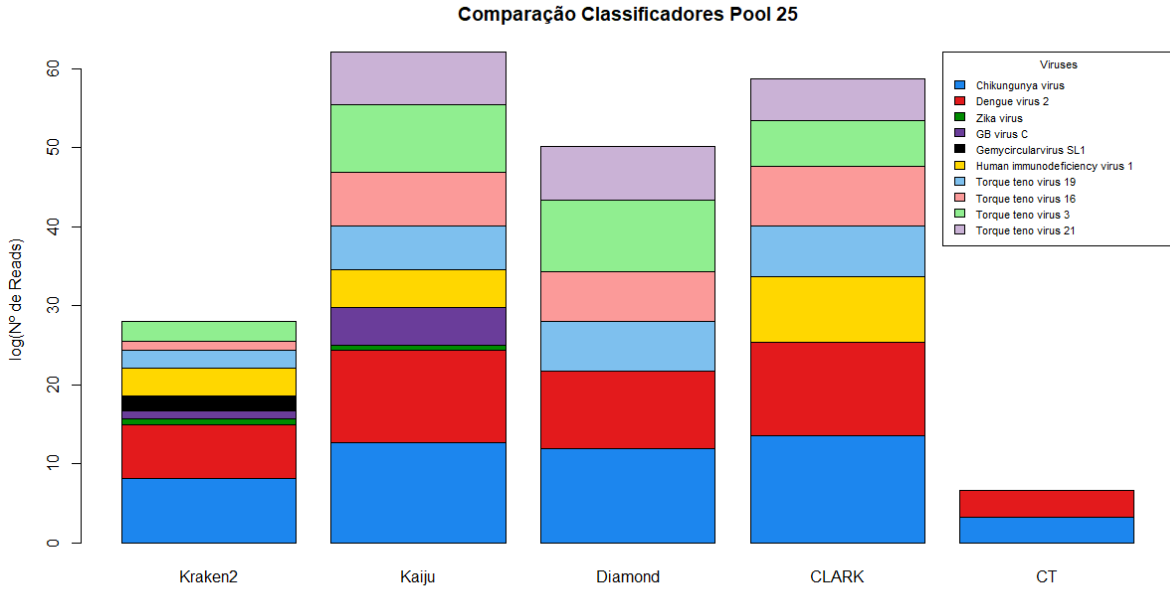
A.



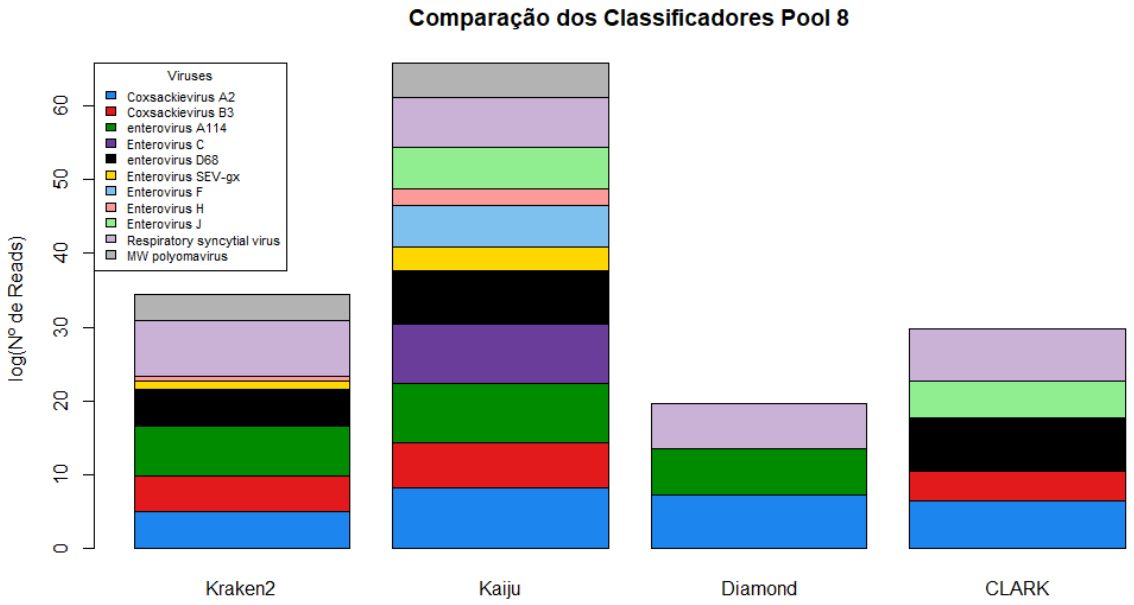

B.


Comparação Classificadores Pool 3
log(Nº de Reads)
0
10
20
30
40
Kraken2
Kaiju
Diamond
CLARK
CT
Viruses
GB virus C
Torque teno virus 6
Torque teno virus 5
Torque teno virus 22
Torque teno virus 29
Torque teno virus 19
Torque teno virus 16
Torque teno virus 12
Torque teno virus 24

C.


Comparação Classificadores Pool 25
log(Nº de Reads)
0
10
20
30
40
50
60
Kraken2
Kaiju
Diamond
CLARK
CT
Viruses
Chikungunya virus
Dengue virus 2
Zika virus
GB virus C
Gemycircularvirus SL1
Human immunodeficiency virus 1
Torque teno virus 19
Torque teno virus 16
Torque teno virus 3
Torque teno virus 21

D.




E.

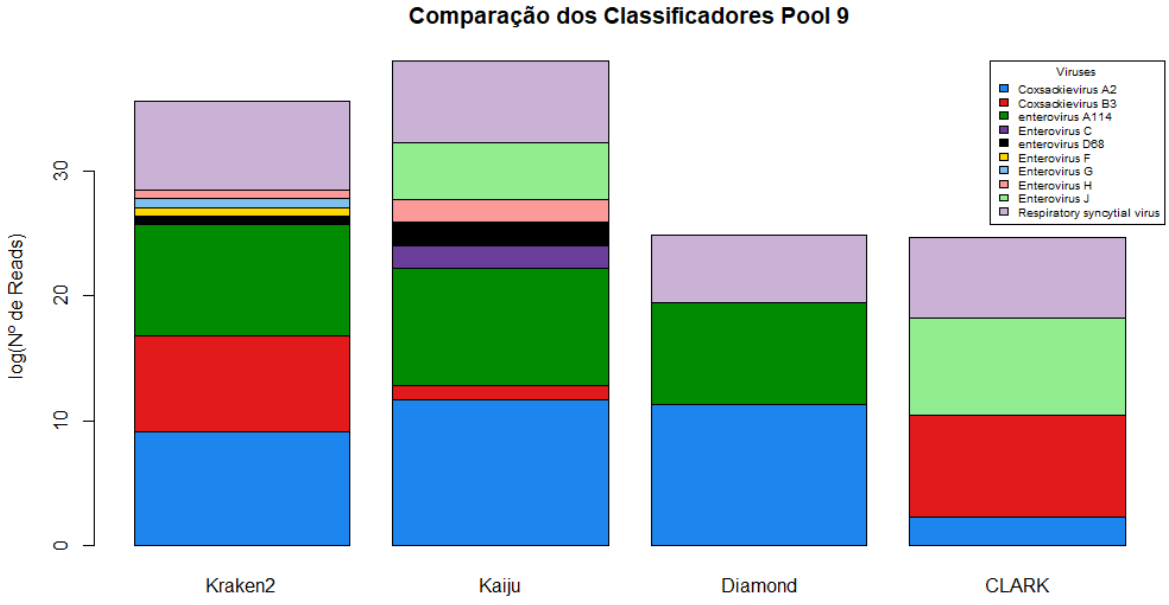



Figura 14: Barplots de comparação entre os 4 Classificadores utilizando como amostras de referência amostras pediátricas negativas para SARS-CoV-2 porém com sintomatologia aguda respiratória (Pool 8 e 9); amostras com amplificação inconclusiva por PCR de arboviroses circulando no estado de Alagoas (Pool 25) e amostras de pacientes com Câncer de Próstata (Pool 2 e 3). Houve confirmação molecular (representado pelo valor de CT dos Pools 2, 3 e 25). (A) É possível perceber a

grande riqueza de vírus em todos os classificadores (com exceção do Diamond), porém apenas o Hepacivírus C foi confirmado. (B) No Pool 3, com exceção com Diamond, todos os classificadores indicaram a presença dos vírus HPV-1(GB vírus C) e do TTV 16, que foram confirmados posteriormente. (C) No Pool 25, os vírus da Dengue 2 e Chikungunya foram identificados em todos os Classificadores e confirmados.

(D e E) Comparação dos resultados de cada classificador dos Pools 8 e 9 do Grupo II, respectivamente.

Os vírus que foram submetidos à confirmação molecular para se obter o valor de CT foram os TTV's 3, 16, 19, 21 e 22; Pegivirus humano do tipo 1 (GB virus C/HPgV-1); vírus da hepatite C (HCV) e vírus da febre Chikungunya e Dengue do tipo 2. Houve confirmação dos Anellovírus TTV 16 nos Pools 2 e 3; TTV 3 no Pool 2; HCV no Pool 2; HPgV-1 no Pool 3 e os vírus da Dengue 2 e da Chikungunya no Pool 25.

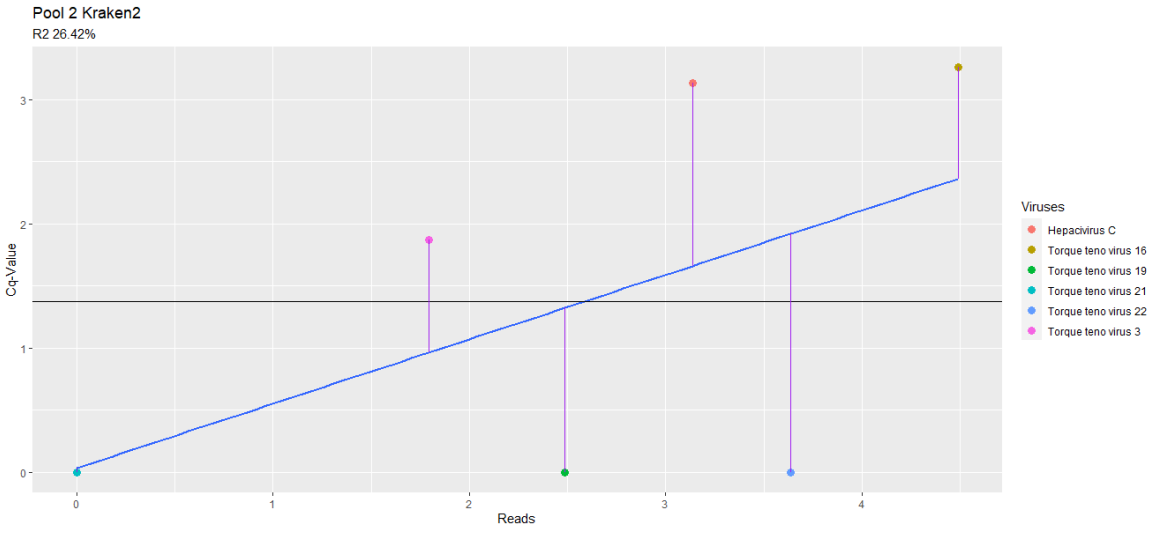
## 4.4. Regressão Linear

Com os valores de CT obtidos dos vírus TTV's 3, 16, 19, 21, 22, Hepacivírus C, HPgV-1, Chikungunya e Dengue 2 dos Pools 2, 3 e 25 do Grupo I, foi realizada a Regressão Linear entre esses valores e os obtidos pelos classificadores. Para avaliar qual o melhor classificador, utilizou-se como métrica o Coeficiente de Determinação ($R^2$ %), assim, quanto mais perto de 1, mais explicativo o modelo e melhor sua performance, conforme realizado no estudo de CARBO *et al.,* 2022.

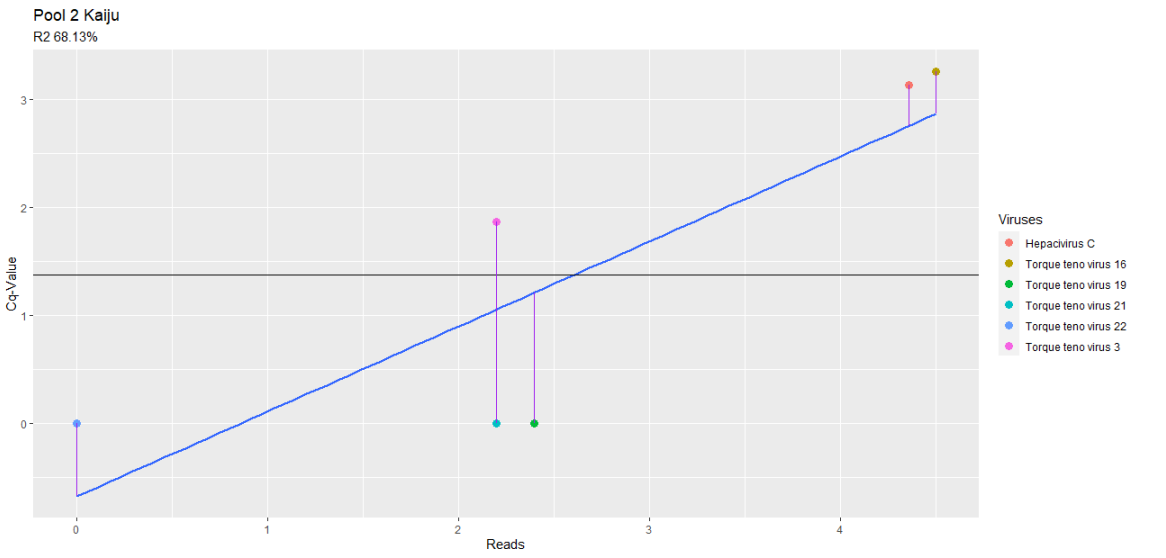
As regressões foram realizadas na linguagem R com o suporte do pacote ggplot2.

### 4.4.1. Performance Classificadores x Pool 2

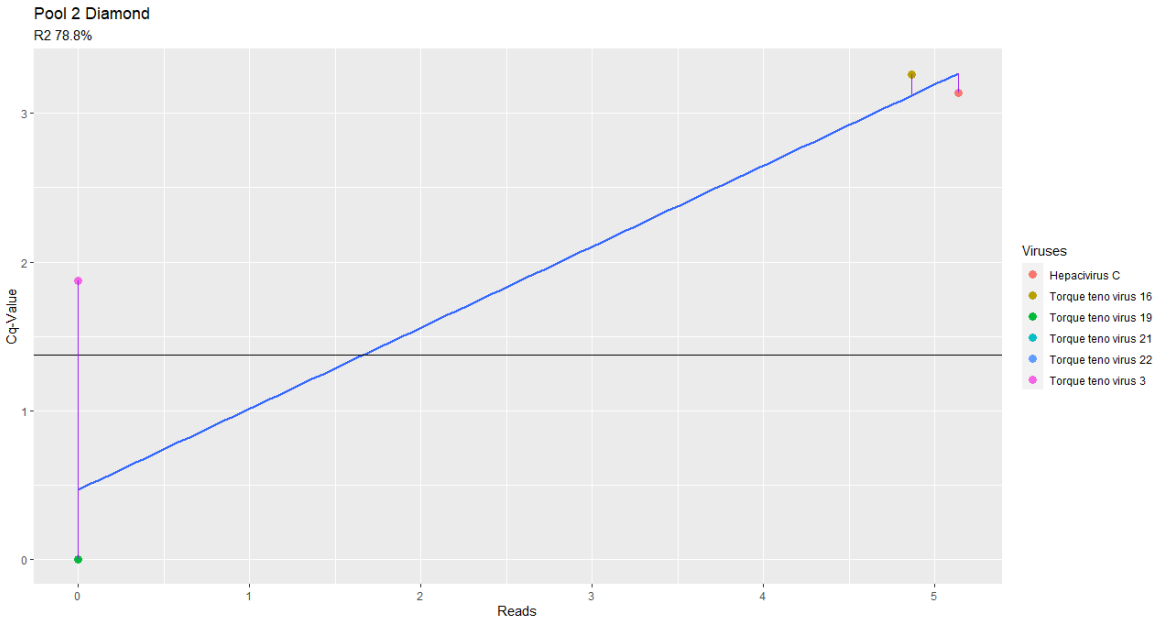
A.


Pool 2 Kraken2
R2 26.42%
Cq-Value
Reads
Viruses
Hepacivirus C
Torque teno virus 16
Torque teno virus 19
Torque teno virus 21
Torque teno virus 22
Torque teno virus 3

B.


Pool 2 Kaiju
R2 68.13%
Cq-Value
Reads
Viruses
Hepacivirus C
Torque teno virus 16
Torque teno virus 19
Torque teno virus 21
Torque teno virus 22
Torque teno virus 3

C.




D.

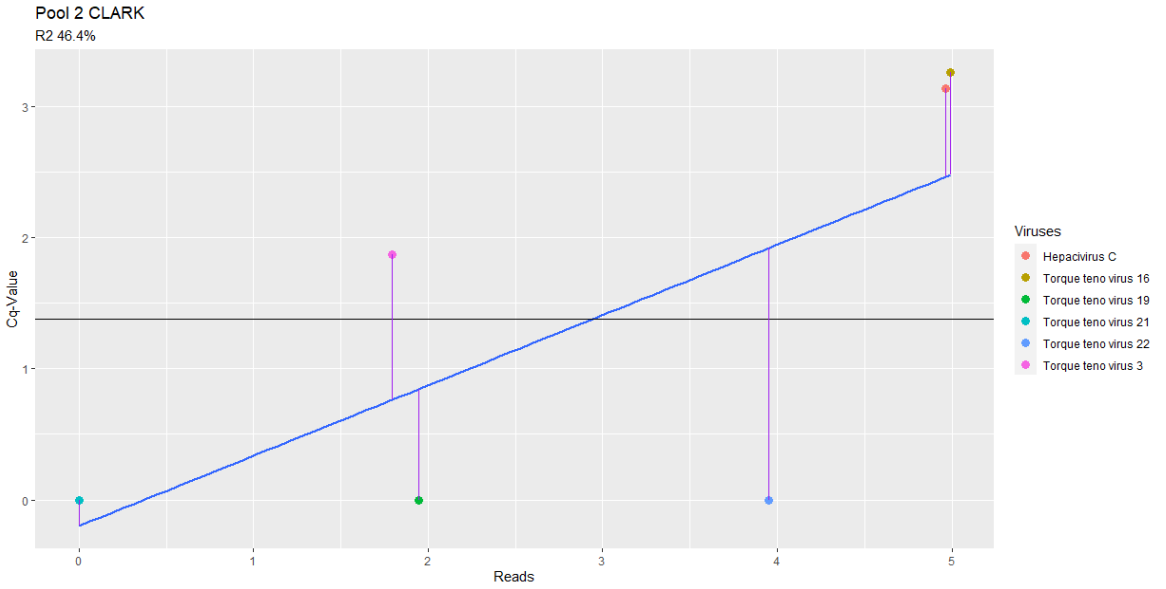
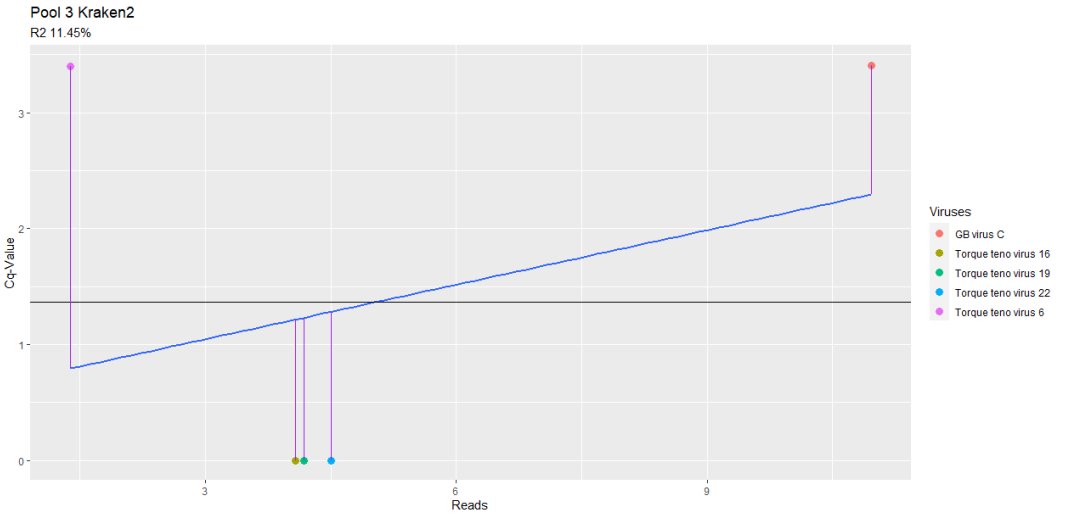



Figura 15: Retas de Regressão Linear Cq-Value x Nº de Reads do Pool 2 obtidos pelos Classificadores. (A) Kraken2. (B) Kaiju. (C) Diamond. (D) CLARK.

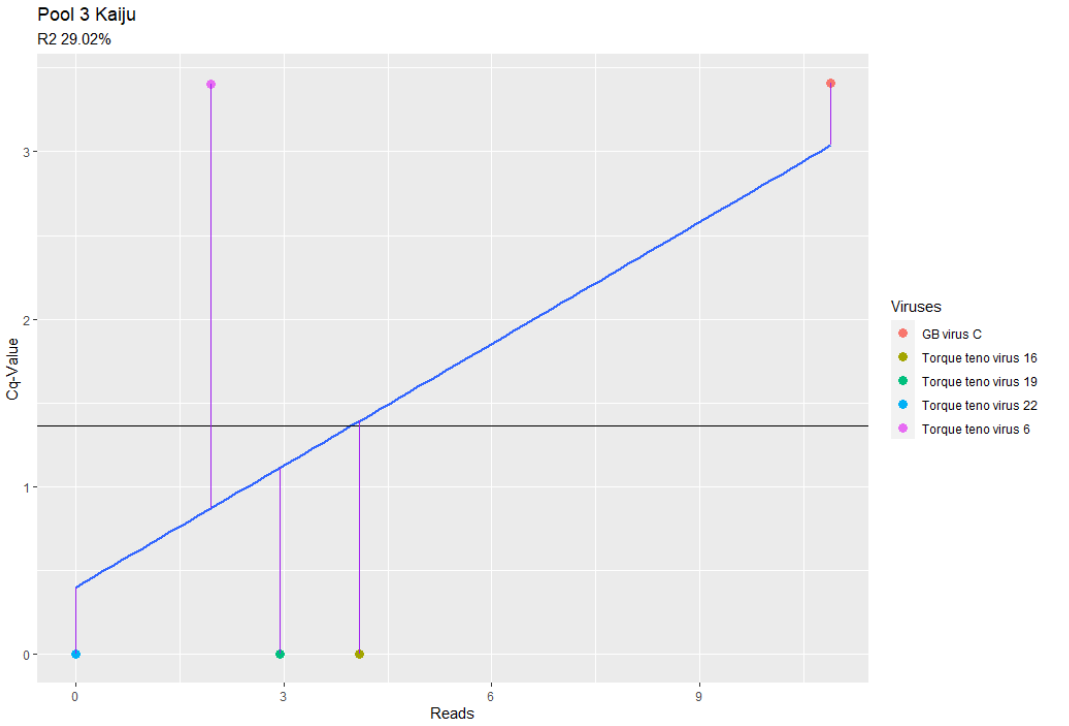
Para o Pool 2, os $R^2$ dos Classificadores foram: Diamond com 78,8%; Kaiju com 68,13%; CLARK com 46,4% e Kraken2 com 26,42%.

### 4.4.2. Performance Classificadores x Pool 3

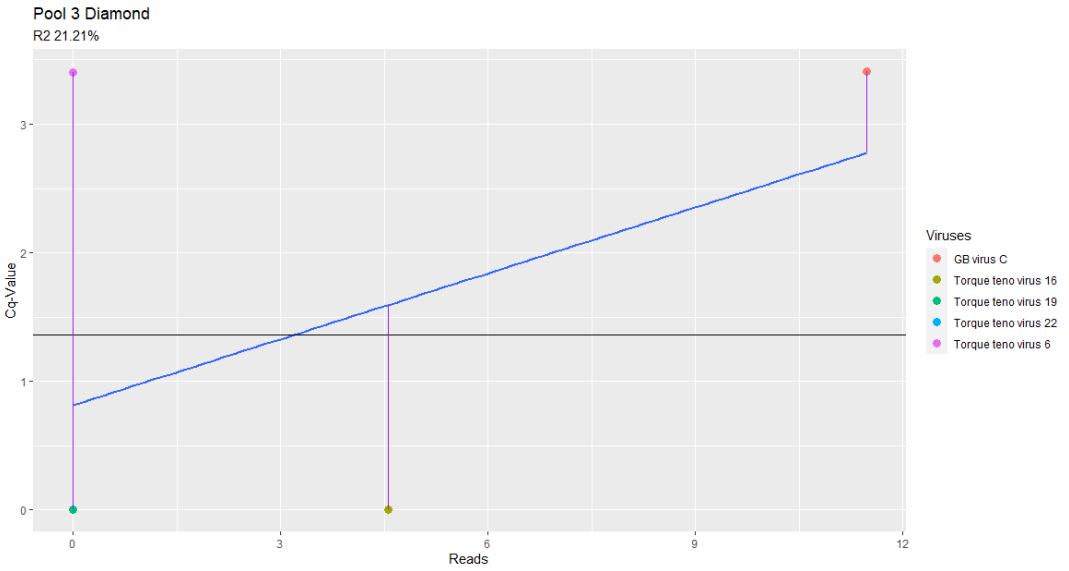
A.


Pool 3 Kraken2
R2 11.45%
Cq-Value
Reads
Viruses
GB virus C
Torque teno virus 16
Torque teno virus 19
Torque teno virus 22
Torque teno virus 6

B.


Pool 3 Kaiju
R2 29.02%
Cq-Value
Reads
Viruses
GB virus C
Torque teno virus 16
Torque teno virus 19
Torque teno virus 22
Torque teno virus 6

C.




D.

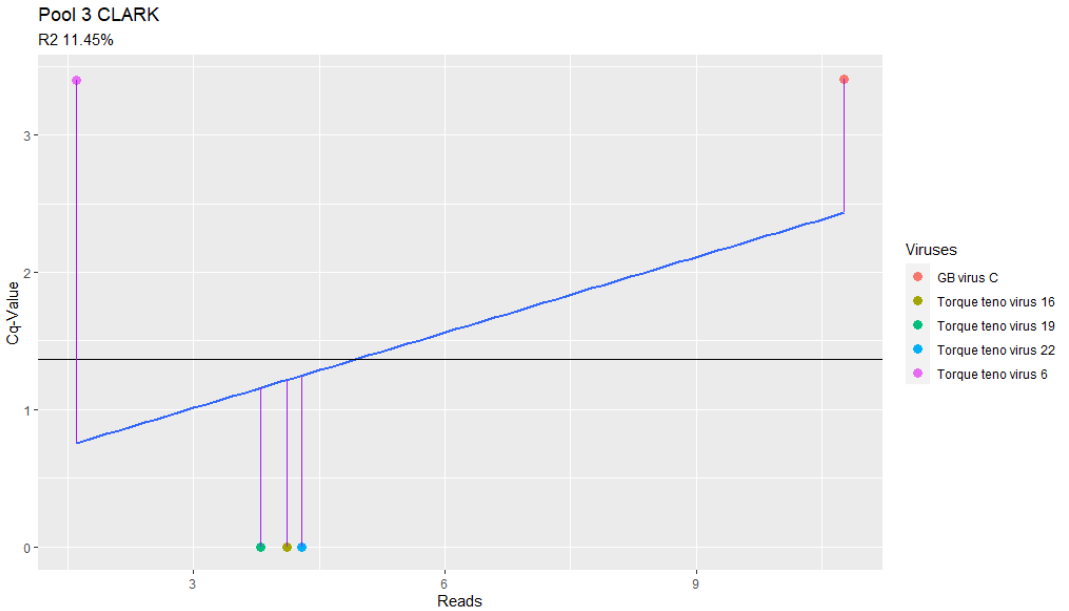
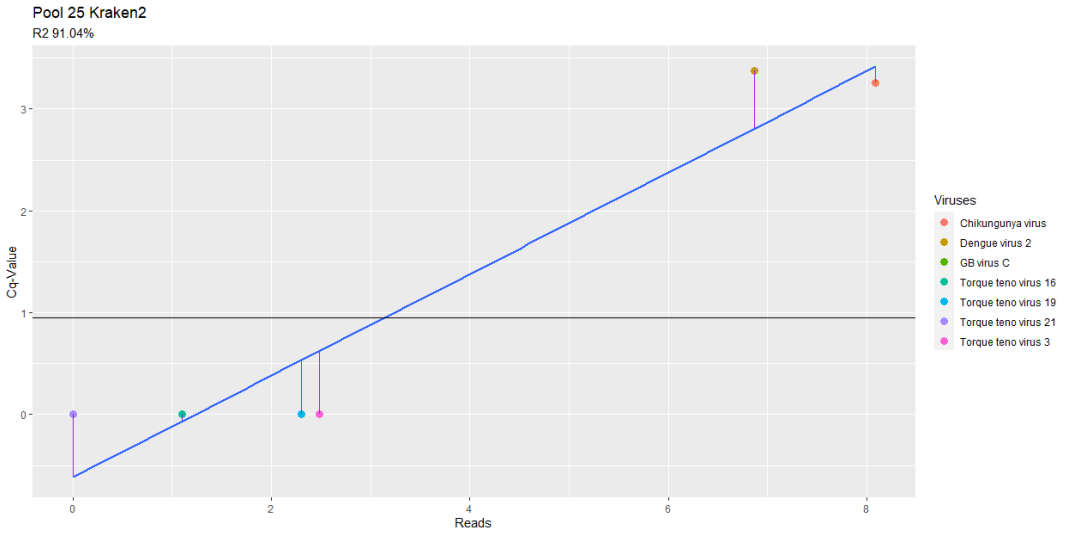



Figura 16: Retas de Regressão Linear Cq-Value x Nº de Reads do Pool 3 obtidos pelos Classificadores. (A) Kraken2. (B) Kaiju. (C) Diamond. (D) CLARK.

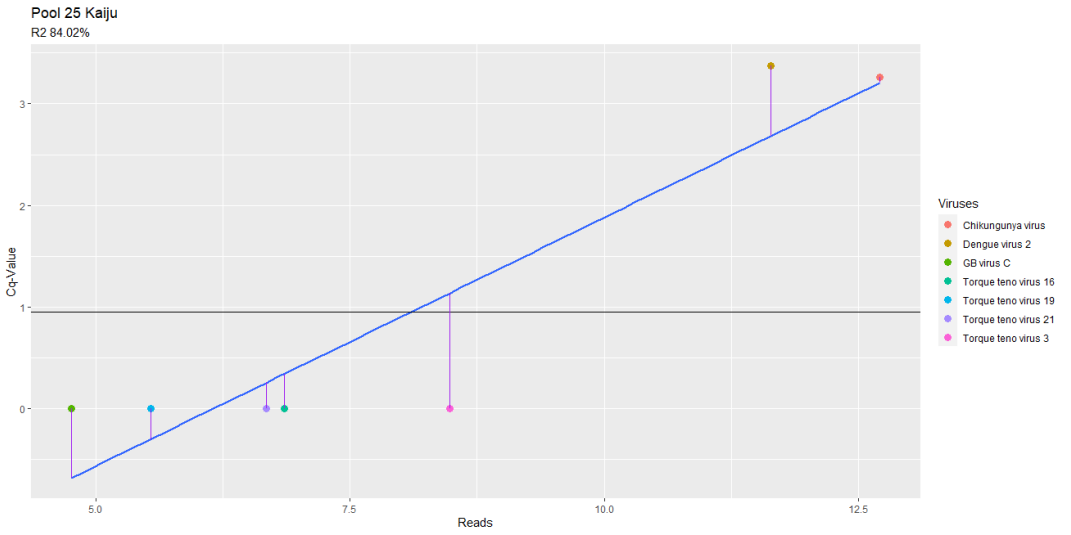
Para o Pool 3, os $R^2$ dos Classificadores foram: Kaiju com 29,02%; Diamond com 21,21%; CLARK com 11,46% e Kraken2 com 11,45%.

### 4.4.3. Performance Classificadores x Pool 25

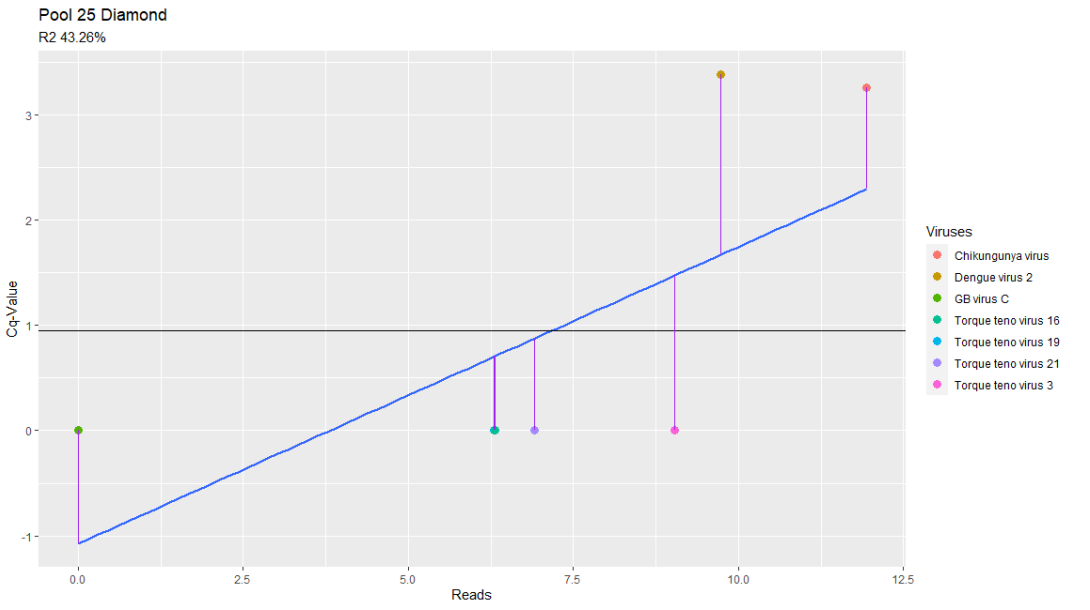
A.


Pool 25 Kraken2
R2 91.04%
Cq-Value
Reads
Viruses
Chikungunya virus
Dengue virus 2
GB virus C
Torque teno virus 16
Torque teno virus 19
Torque teno virus 21
Torque teno virus 3

B.


Pool 25 Kaiju
R2 84.02%
Cq-Value
Reads
Viruses
Chikungunya virus
Dengue virus 2
GB virus C
Torque teno virus 16
Torque teno virus 19
Torque teno virus 21
Torque teno virus 3

C.


Pool 25 Diamond
R2 43.26%
Cq-Value
Reads
Viruses
Chikungunya virus
Dengue virus 2
GB virus C
Torque teno virus 16
Torque teno virus 19
Torque teno virus 21
Torque teno virus 3

D.

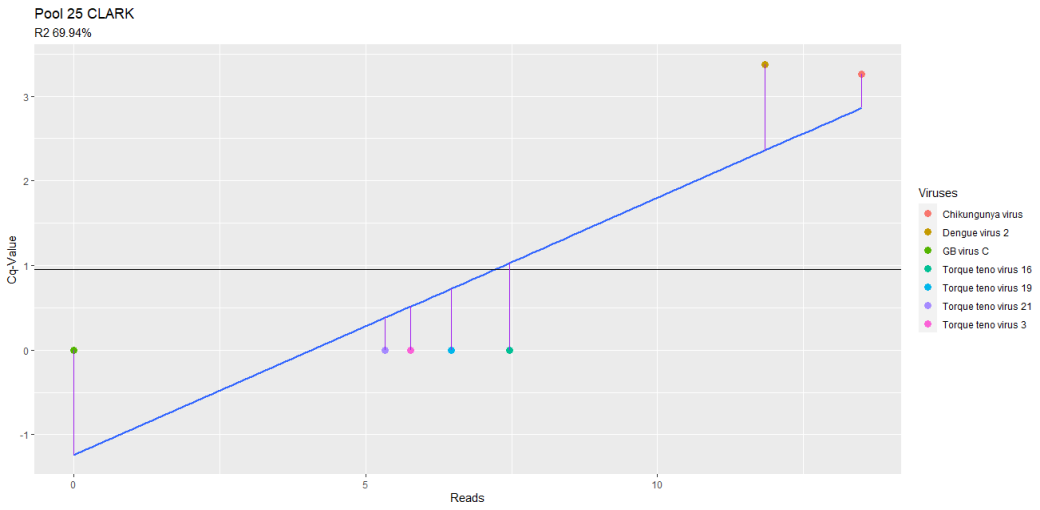



Figura 17: Retas de Regressão Linear Cq-Value x Nº de Reads do Pool 25 obtidos pelos Classificadores. (A) Kraken2. (B) Kaiju. (C) Diamond. (D) CLARK.

Para o Pool 25, os R² dos Classificadores foram: Kraken2 com 91,04%; Kaiju com 84,02%; CLARK com 69,94% e Diamond com 43,26%.

Os resultados dos Pools para cada classificador foram compilados na Tabela 1.

| | **Pool 2** | **Pool 3** | **Pool 25** |
|---|---|---|---|
| **Kaiju** | 68.13% | 29.02% | 84.02% |
| **CLARK** | 46.4% | 11.46% | 69.94% |
| **Kraken2** | 26.42% | 11.45% | 91.04% |
| **Diamond** | 78.8% | 21.21% | 43.26% |

Tabela 1: Valor do R² dos Pools para cada Classificador

## 4.5. Vírus de Interesse Montados

Depois de uma revisão na literatura, foram escolhidos, especificamente, os vírus *Coxsackievirus A2* e o *enterovírus D68,* que, conforme mostra a figura 18, estão presentes em quase todos os pools. Ambos vírus estão relacionados com recentes epidemias, sendo o Coxsackievirus A2, causador da doença mão-pé-boca, tendo ocorrido principalmente na China entre 2008 e 2009 (HU,

*et al.,* 2011) e o enterovírus D68, responsável por sintomas respiratórios graves, em 2014, nos Estados Unidos e na Europa (SUN, *et al.,* 2019).

Os táxons foram investigados com o Kraken2 e a Pipeline II.

A abundância dos vírus presentes no Grupo II pode ser observada na Figura 18.

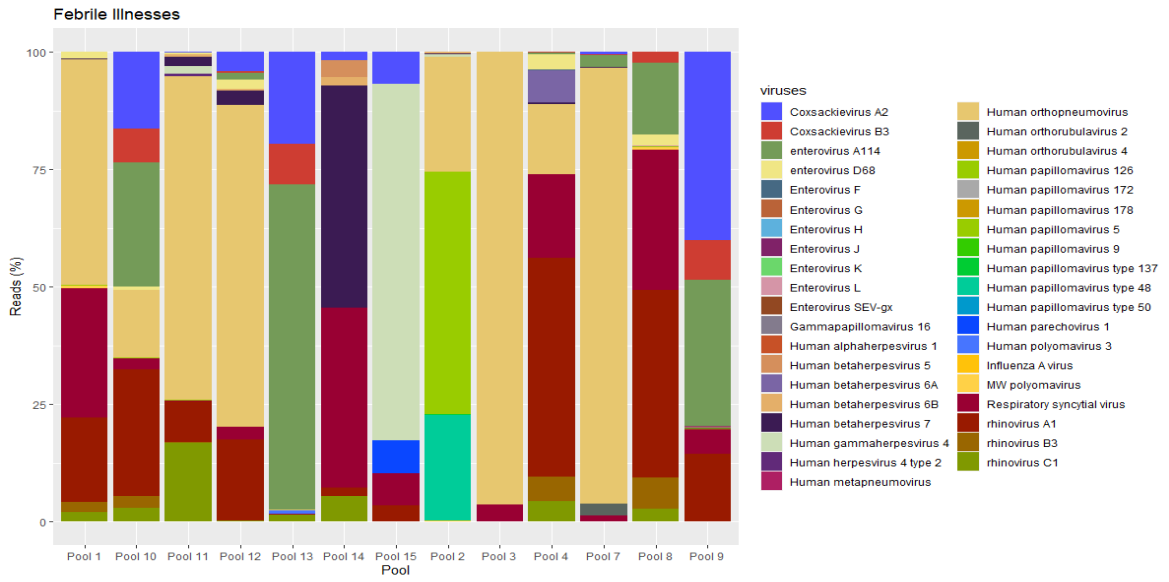



Figura 18: Porcentagem dos vírus de interesse obtidos pela *pipeline* II com o Kraken2 dos Pools do Grupo 2.

### 4.5.1. Linha de Comando

**spades.py** -1 Pool_R1.fq.gz -2 Pool_R2.fq.gz -s Pool.fq.gz -k 21,33,55,77 --trusted-contigs /genomes/reference.fasta -o spades_output/

O parâmetro –trusted-contigs indica que será usado um genoma de referência, não para a montagem, mas sim para o fechamento de *gaps*. Para ambos os vírus, a referência foi obtida pela busca no *NCBI-Nucleotide* pelo genoma completo. O genoma de referência do Coxsackievirus A2 possui 7363 pares de base e é de 2018, disponível em <https://www.ncbi.nlm.nih.gov/nuccore/MF281257.1>. O genoma de referência do enterovírus D68 possui 7367 pares de base, também de 2018, disponível em <https://www.ncbi.nlm.nih.gov/nuccore/NC_038308.1/>. Ambos com acesso em 21 de nov. de 2022.

## 4.6. Análise Filogenética

### *4.6.1. Coxsackievirus* A2

Foram montados 5 scaffolds de Coxsackievirus A2 provenientes dos Pools 3, 9, 10, 11 e 12, estes foram agregados em um arquivo do tipo FASTA contendo genomas de subtipos de Coxsackievirus A:

- 17 genomas de A10;
- 15 genomas de A16;
- 10 genomas de A6;
- 7 genomas de A2;
- 5 genomas de A4;
- 4 genomas de A5, A7, A8 e de A12;
- 3 genomas de A14 e de A71;
- 1 de A3 e
- 1 de *Simian enterovírus* 19 e 13.

Esse arquivo foi chamado “coxsackie.fasta” com 73 genótipos no total. O alinhamento foi feito pelo comando **mafft** --auto --inputorder "coxsackie.fasta" > "coxsackie_aln.fasta”. Todos os genótipos compondo o dataset de alinhamento citados foram obtidos do NCBI utilizando o seguinte parâmetro: “Coxsackievirus ***subtipo*** complete genome”.

Com o *output* gerado, partiu-se para a montagem do arquivo TREEFILE pelo programa IQ-TREE a partir da linha de comando “**iqtree** -s coxsackie_aln.fasta -lmap 1000 -bb 1000 -m MFP”. O mapa de verossimilhança é apresentado na Figura 19. Quanto mais quartetos distribuídos pelos vértices, mais informativos eles são. Recomenda-se que o valor central seja menor que 30%, pois são pouco informativos e que os vértices somem mais de 60% (Strimmer & von Haeseler, 1997).

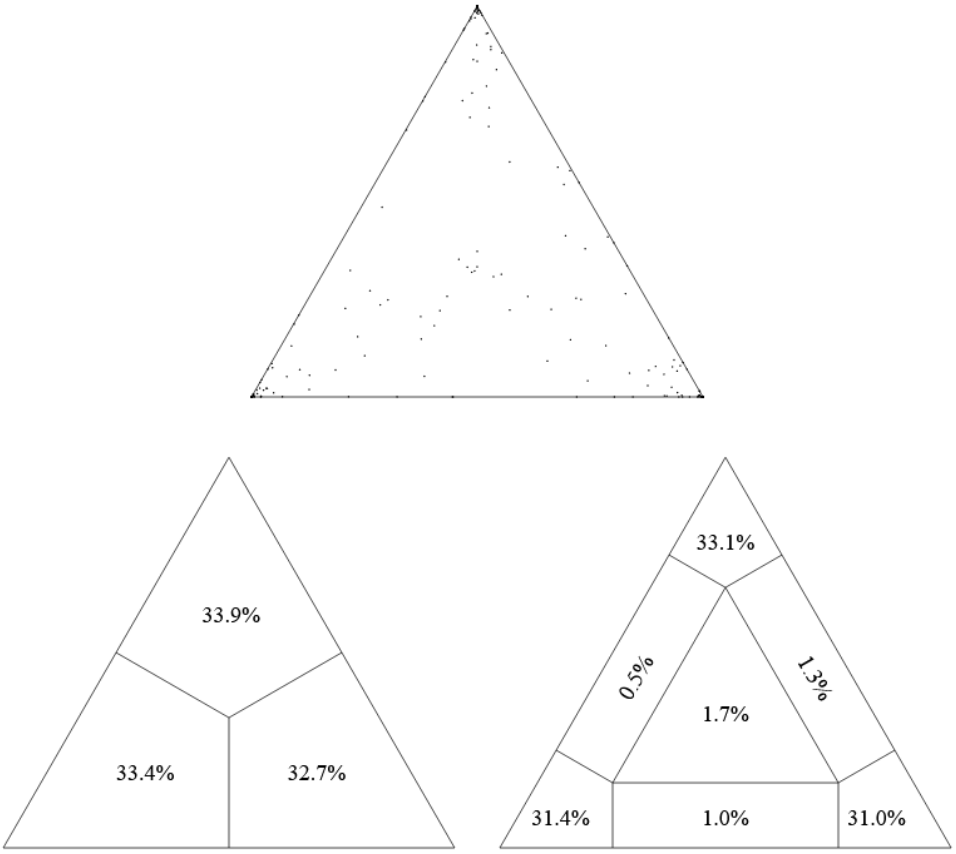



Figura 19: Distribuição dos quartetos pelo mapa de verossimilhança dos subtipos de Enterovírus A

O Mapa de Verossimilhança nos mostra que a árvore gerada é informativa, uma vez que seu valor central é de 1,71% (inferior a 30%) e o valor dos vértices soma mais de 60%. Com isso, foi possível reconstituir a árvore filogenética dos genomas montados de Coxsackievirus A. Com o arquivo do tipo TREEFILE, sua edição foi feita utilizando o programa FigTree, conforme a Figura 20.

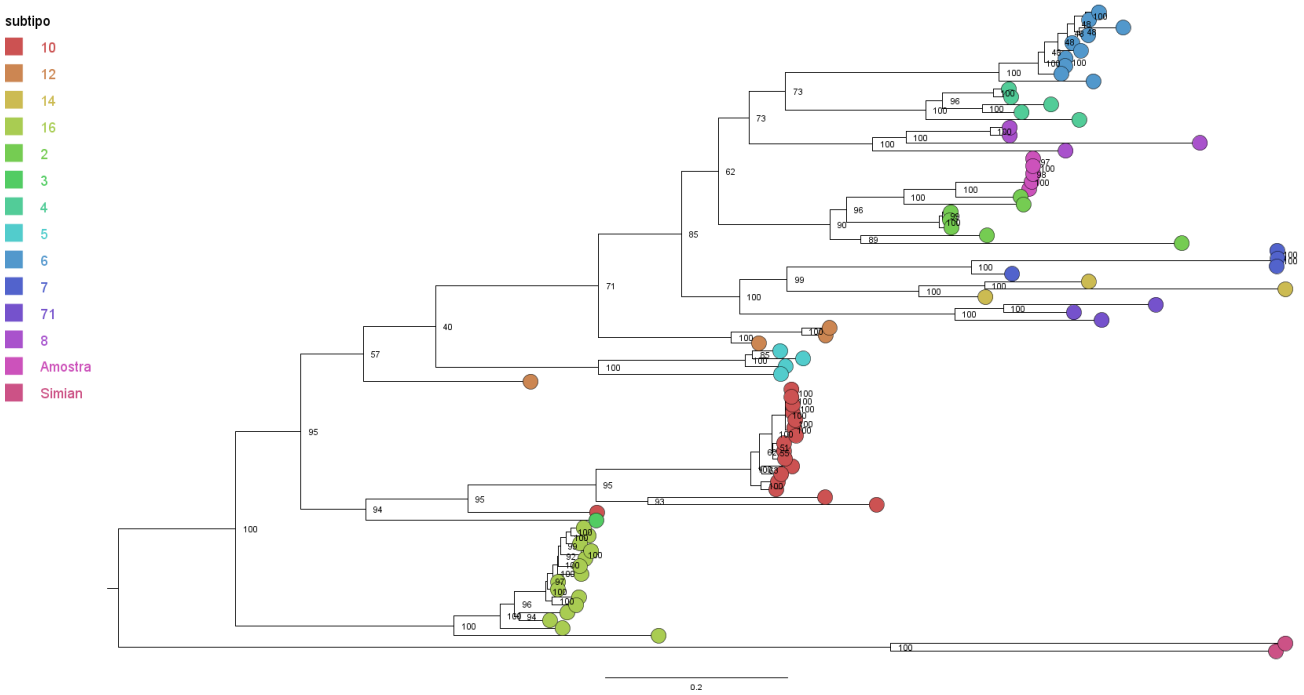



Figura 20: Árvore Filogenética dos subtipos de CoxsackieA: A10, A12, A14, 16, A2, A3, A4, A5, A6, A7, A71, A8, *Simian enterovírus 13* e *Simian enterovírus 19*

### *4.6.2. enterovírus D68*

Foram montados 5 scaffolds de enterovírus D68 provenientes dos Pools 1, 9, 10, 11 e 12 e agregados em um arquivo do tipo FASTA contendo genótipos de subtipos de Enterovírus D:

- 13 genomas de D68 e
- 2 genomas de enterovírus D111, D70 e de D94

Esse arquivo foi chamado “entero.fasta”. O alinhamento dos 19 genótipos foi feito pelo comando **mafft** --auto --inputorder "entero.fasta" > "entero_aln.fasta". Todos os genótipos compondo o dataset de alinhamento citados foram obtidos do NCBI utilizando o seguinte parâmetro: "enterovírus ***subtipo*** complete genome”

Após o alinhamento, percebeu-se que todas as sequências pertencem ao mesmo genótipo e, portanto, foi escolhido apenas um genoma completo Pool 1, para verificar o

genótipo deste vírus e fazendo parte a árvore filogenética contendo os representantes do grupo de Enterovírus D.

O comando do IQ-TREE para os Enterovírus D foi "**iqtree** -s entero_aln.fasta -lmap 1000 -bb 1000 -m MFP"

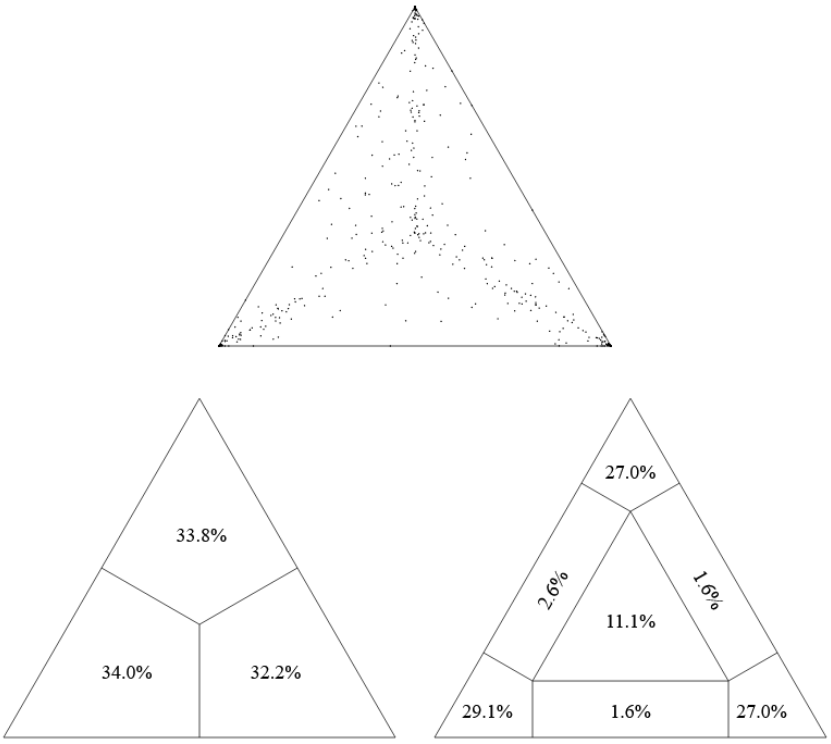



Figura 21: Mapa de Máxima Verossimilhança da árvore filogenética dos subtipos dos Enterovírus D

O Mapa de Verossimilhança nos mostra que a árvore gerada é informativa, uma vez que seu valor central é de 11.1% (inferior a 30%) e o valor dos vértices soma mais de 60%. Com isso, foi possível formar a Árvore dos Enterovírus D, mostrada na Figura 22.

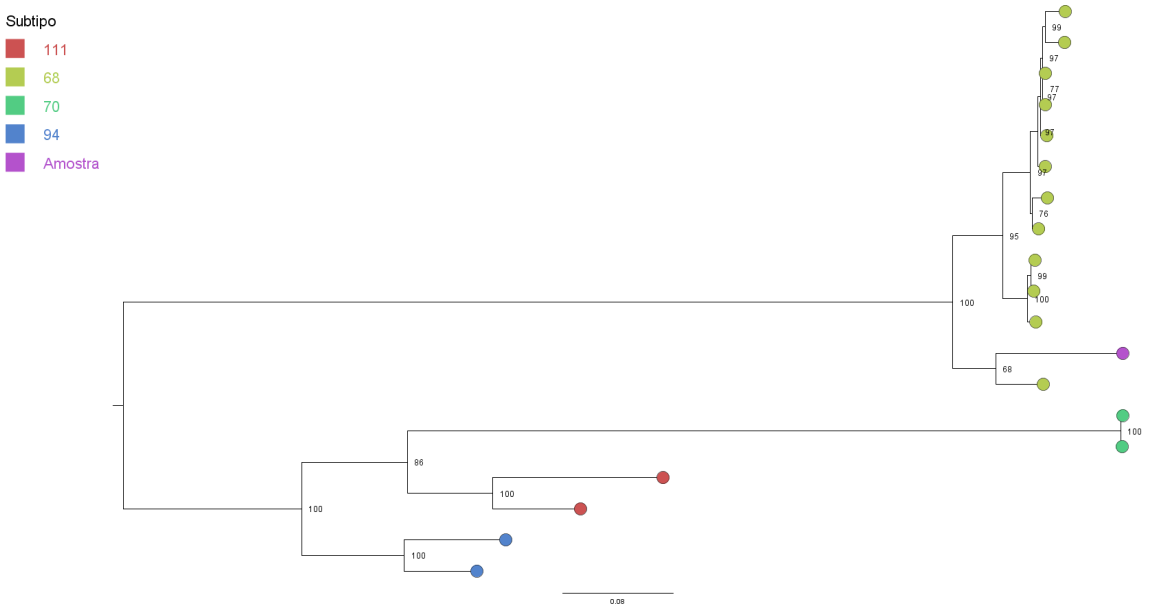



Figura 22: Árvore Filogenética referente aos subtipos de Enterovírus D: D111, D68, D70 e D94

# 5. DISCUSSÃO

Neste trabalho foram comparados 4 classificadores taxonômicos citados na literatura (CARBO *et al.,*2022; DE VRIES *et al.,* 2021; YE *et al.,* 2019) para a anotação do viroma em amostras clínicas; para tal, foram utilizadas as técnicas de comparação visual por meio de barplots e de regressão linear dos resultados dos classificadores pelo valor de Ct obtido ao se realizar o qPCR dos vírus detectados pelos classificadores. A avaliação foi feita por meio do Coeficiente de Determinação ($R^2$ %). Complementar a isso, foi feita a montagem e análise filogenética dos vírus Coxsackievirus A2 e enterovírus D68 cujos genomas completos foram recuperados devido ao grande número de leituras identificados em amostras de pacientes pediátricos negativas para SARS-CoV-2. Estes dados demonstram que a metagenômica viral é uma ferramenta poderosa para avaliar a evolução molecular dos vírus, além de verificar a abundância viral de uma amostra.

## 5.1. Comparação Visual

Como se pode observar a partir dos resultados obtidos (4.3), tanto o Kraken2 quanto o Kaiju apresentaram resultados mais notáveis, devido à quantidade de vírus identificados, principalmente os referentes ao Pool 25, onde, ambos encontraram o vírus da febre Zika, *Gemycircularvirus* e *Human*

*Immunodeficiency virus*, o que indica que a metagenômica é método robusto e muito adequado para identificar vírus desconhecidos ou pouco caracterizados em amostras clínicas. Por outro lado, o programa Diamond identificou apenas vírus confirmados por qPCR.

O vírus da febre *Zika* é um flavivírus cujos vetores são mosquitos do gênero *Aedes* (AGUMADU e RAMPHUL, 2018), o mesmo que transmite outros arbovírus como Dengue e Chikungunya. Os sintomas clínicos das 3 infecções são parecidos e portanto a identificação destes agentes virais na base clínica é praticamente impossível (AGUMADU e RAMPHUL, 2018). No ano de 2015, houve uma epidemia de Zika no Brasil, resultando em vários casos de microcefalia em recém-nascidos pelo país, principalmente na região Nordeste (ZANOTTO e LEITE, 2018; AMBROGI, BRITO e DINIZ, 2020). A identificação do vírus pelos classificadores Kraken2 e Kaiju, especialmente de um Pool vindo do Nordeste pode indicar a reemergência deste agente na região. Portanto, a identificação correta destes agentes virais é de suma importância para o diagnóstico e até mesmo para o manejo dessas infecções.

Outro vírus identificado pelo programa Kraken2 foi o *gemycircularvirus SL1,* que é um vírus de DNA de fita simples, pertencente à família *Genomoviridae.* Estes vírus são amplamente distribuídos na natureza e podem infectar desde fungos, plantas, diversos animais até humanos; e já foram identificados em fezes e amostras de sangue brasileiras (PHAN, *et al.,* 2015). Não se sabe ao exato qual é o significado da detecção de genomas dos vírus gemycircular em amostras clínicas. Seu diagnóstico pode ser relacionado devido ao surgimento desse vírus na população humana, ou apenas por uma infecção fúngica que poderia liberar este vírus na corrente sanguínea, ou fungos/plantas consumidos na dieta ou mesmo contaminação de partículas flutuando no ar (PHAN, *et al.,* 2015).

Nós demonstramos também a presença de leituras pertencentes ao HIV no Pool 25, pelo Kraken2, Kaiju e CLARK. É conhecido que o genoma humano é composto de genes oriundos da família de retrovírus, a qual o HIV pertence (LUGANINI e GRIBAUDO, 2020) e portanto sua presença nos resultados pode indicar que se trata de algum vetor viral ou sequências humanas contendo elementos de retrovírus endógenos que passaram o filtro de qualidade. Nós observamos que as leituras pertencem ao gene comum de todos

os retrovírus através do qual é realizada integração no genoma do hospedeiro: a região terminal longa (LTR). A presença de um único gene sem cobertura em outras partes do genoma do HIV é indicação de contaminação durante o preparo de bibliotecas e especificamente de vetores de clonagem cujos sítios de integração contém elementos de LTR (LUGANINI e GRIBAUDO, 2020). Uma indicação forte, que essas sequências são provenientes de contaminação, foi que conseguimos realizar testes confirmatórios da presença de HIV tanto por meio de ensaios sorológicos como por meio de amplificação dos ácidos nucleicos. Em ambos os casos as reações foram negativas, o que é indicação de ausência de infecção por este retrovírus. Em suma, independentemente do grande potencial da metagenômica para a identificação de novos vírus, os mesmos precisam ser confirmados diretamente na amostra para verificar a veracidade dos resultados obtidos, ainda mais quando se trata de uma infecção de suma importância para a saúde pública como a do HIV.

Nos Pools 2 e 3 obtidos de pacientes com câncer de próstata, foi possível identificar a presença de *HCV* e HPgV-1, respectivamente. HCV é o vírus causador da hepatite viral do tipo C, pertencente à família *Flaviviridae*, transmitido através de vias parenterais (PISANO *et al.,* 2021). Embora o HCV seja estabelecido como agente causador de câncer hepatocelular, sua associação com o câncer de próstata ainda não é totalmente entendida (MAHALE, *et al.,* 2017). O HPgV-1 que, também pertence aos *Flaviviridae,* é transmissível também através de vias parenterais e principalmente transfusão de sangue e drogas injetáveis (RESHETNYAK *et al.,* 2008). HPgV-1 é considerado vírus comensal que não tem relação com quadro clínico. No entanto, em pacientes com HIV e coinfecção com HPgV-1 foi observada uma relação positiva entre a presença deste vírus e número dos linfócitos CD4+/DD8+ deste modo postergando a progressão da infecção a síndrome da imunodeficiência adquirida (YU *et al.,* 2022).

Foi interessante perceber, também, como os classificadores se comportam em relação aos TTV's, visto que o Kraken2, Kaiju e CLARK identificaram mais espécies desse vírus, principalmente os TTV's 3, 5 e 16 (confirmado) no Pool 2 e os TTV`s 5 (confirmado), 6 e 24 no Pool 3. Os Torque Teno Vírus pertencem à família dos Anellovírus, sua grande abundância das espécies de TTV's pode ter ocorrido por sua pouca

variabilidade genética e grande quantidade de espécies no plasma (SARAIRAH, *et al.,* 2020), levando à confusão do classificador.

Considerando as amostras obtidas de pacientes com síndrome respiratória aguda porém negativos para SARS-CoV-2 foi observada uma grande abundância de espécies virais pelos classificadores Kraken2 e Kaiju e pouca no Diamond, o que era esperado. Nos Pools 8 e 9, nós observamos principalmente os vírus *Coxsackie A2, Coxsackie B3, enterovírus A114* e *enterovírus D68* que causam doença aguda em crianças e demonstram transmissão respiratória. Os vírus citados pertencem ao gênero dos *Enterovírus*, são caracterizados por sua fita simples de RNA; possuem grande prevalência mundial, infectando tanto crianças (principalmente) quanto adultos e causam ampla gama de doenças, incluindo encefalite, meningite, miocardite, doença mão-pé-boca, conjuntivite, doenças respiratórias e doença gastrointestinal (BROUWER *et al.,* 2021).

## 5.2. Análises de Regressão Linear

No geral, os classificadores que utilizam a mesma base (de nucleotídeos ou aminoácidos) apresentam performance parecida. Nos Pools 2 e 3, os classificadores a base de aminoácidos demonstraram desempenho melhor do que os com base de nucleotídeos. Segundo Ye *et al.*, a similaridade das classificações pode indicar que há outros fatores como recursos computacionais que não podem ser ignorados, como a composição e tamanho tanto do banco de dados quanto da amostra e a usabilidade do classificador pelo usuário (discutido em 5.3). Ainda segundo o mesmo autor, um *database* customizado é melhor pois é possível selecionar as espécies. Neste trabalho, todos os 4 classificadores possuem a opção de customizar o banco de dados; a customização foi feita para todos os classificadores, com exceção do Kraken2 que utilizou o *RefSeq* completo, identificando apenas táxons virais.

A classificação robusta do Diamond no Pool 2 pode ser explicada pois o classificador apenas indicou os vírus que realmente estavam presentes na amostra, mesmo que não tenha identificado o TTV 3. Embora o Kraken2 tenha identificado várias espécies de TTV's que não estão presentes nos outros pools através de reação molecular, nós conseguimos confirmar apenas 2  e, portanto o valor de $R^2$ foi relativamente baixo, de 26,42%. Por outro lado, nós não

possuímos testes moleculares para todos os tipos de TTV, que também pode ser contribuído para o valor baixo de R2.

No Pool 3 (pacientes de câncer de próstata), todos os classificadores apresentaram baixo desempenho, visto que nenhum alcançou um coeficiente de determinação superior a 30%. Novamente os classificadores à base de aminoácidos apresentaram resultados melhores, entretanto o Diamond foi o único que não identificou o TTV 6, confirmado nesse Pool. A baixa performance de CLARK e Kraken2 novamente se dá pela alta quantidade de espécies virais encontradas e tendo apenas 2 sendo confirmadas através de métodos moleculares.

As regressões referentes ao Pool 25 (doença arboviral aguda) foram as que apresentaram resultados mais expressivos, sendo que apenas o Diamond demonstrou $R^2$ inferior a 50%, porém, foi o único que não identificou a presença de HIV. A baixa performance pode ser explicada, pois os vírus confirmados foram os que mais se distanciaram da reta de regressão. O Kraken2, apesar de ter identificado mais espécies, apresentou baixo número de leituras, o que fez com que mais se aproximasse de seu valor verdadeiro (nulo), assim, adaptando-se melhor a reta de regressão; comportamento semelhante foi identificado no Kaiju, cuja leituras foram inferiores as identificadas nos outros dois classificadores e portanto alcançou $R^2$ de 84,02%,

No estudo de Ye *et al.,* os classificadores a base de nucleotídeos apresentam resultados melhores pois os de aminoácidos podem apresentar ausência de sequências não codificantes em seu banco de dados; esse resultado foi diferente do observado em nosso estudo, onde o Kaiju e o Diamond apresentaram bons resultados no geral. Porém, é importante ressaltar que Ye considerou, além de Kaiju e Diamond, o MMseqs2 como classificador a base de aminoácidos e fez a classificação não só de vírus, mas também de bactérias.

Carbo *et al.* fez seu estudo apenas em amostras virais, o que aproxima aos nossos resultados, pois em sua análise os classificadores DNA-Proteína (a base de aminoácidos) são mais sensíveis para sequências novas ou altamente variáveis devido a pouca taxa de mutação nas proteínas em relação ao nucleotídeos. Isso explica o porquê do Diamond apresentar pouca abundância para os TTVs e enterovírus.

Ambos estudos concordam que a presença da sequência de *Homo sapiens* no banco de dados dos classificadores pode ajudar a identificar possíveis contaminações, isso foi confirmado e importante para substituição da *Pipeline* I para a II (Figura 11.C)

## 5.3. Usabilidade dos Classificadores

Em termos de usabilidade, o Diamond e o CLARK não apresentaram bom desempenho. Conforme Ye *et al.,* e comprovado neste estudo, o Diamond não apresenta o perfil de abundância e por isso deve ser utilizado junto com o MEGAN, isso faz com que haja um custo de tempo ao realizar a análise metagenômica. O CLARK, por sua vez, não possui um bom suporte, sua página está desatualizada e não é intuitiva, principalmente para a montagem de seu banco de dados.

## 5.4. Análise Filogenética dos Vírus de Importância para a Saúde Pública

Os vírus escolhidos para a montagem e, posteriormente, para a análise filogenética, são pertencentes aos Pools do Grupo II. Os *Enterovírus* são um gênero da família dos *Picornaviridae* que possuem prevalência de 2,9% em crianças com bronquiolite (BROUWER *et al., 2021*). São transmitidos principalmente por secreções respiratórias da pessoa infectada e podem causar uma ampla gama de condições clínicas, incluindo meningite asséptica, doenças respiratórias e miocardite, embora muitos casos sejam assintomáticos ou benignos (LIN *et al.,* 2018; KENMOE *et al.*, 2020; BROUWER *et al., 2021*).

Os vírus escolhidos para a montagem dos contigs e para análise filogenética foram o *Coxsackievirus A2* e o *enterovírus D68*. O *Coxsackievirus A2* (CVA2) é uma subespécie do Enterovírus A. É mais presente na Europa e na maioria dos casos causa problemas respiratórios mas também pode causar a doença mão-pé-boca, caracterizada por infectar principalmente crianças e causar pequenas feridas na cavidade oral e erupções nas mãos e nos pés, ocorrendo principalmente no verão e outono (BROUWER *et al.,* 2021; CDC/hand-mouth-foot, 2022). Coxasckievirus A2 causa doença aguda em pacientes pediátricos mas devido a similaridade dos sintomas com viroses respiratórias sua presença na maioria dos casos não é suspeita. Por este motivo, podemos afirmar a grande importância da metagenômica para identificação de

vírus não suspeitos bem como para a vigilância molecular destes agentes o que aconteceu com o vírus supracitado o qual foi analisado filogeneticamente.

O enterovírus D68 (EV-D68) é uma subespécie de Enterovírus D. É mais comum na Europa e é conhecido por causar problemas respiratórios; ocorre principalmente do verão e outono, mas não se descarta a possibilidade de circulação durante todo o ano (BROUWER *et al.,* 2021; CDC/ev-d68, 2022). Como comentado anteriormente, houve uma epidemia de EV-D68 nos Estados Unidos de Agosto de 2014 até Janeiro de 2015, infectando 1395 pessoas, principalmente crianças (SUN *et al*., 2019; CDC/ev-d68, 2022).

Em relação aos genomas completos dos *Coxsackievirus* A obtidos a partir dos Pools foram agrupados conforme o esperado, ou seja, agrupados junto com os genótipos de CVA2. Segundo a relação filogenética, a linhagem que se encontra nos Pools é mais próxima da linhagem de *Coxsackievirus* A2 que circulou Estados Unidos de 2014, o que indica que a linhagem das Américas é mais similar entre si do que as outras linhagens (provenientes da China) analisadas. O EV-D68 montado a partir dos Pools também foi agrupado corretamente, junto aos D68. Assim como o CVA2, a linhagem que circula nas amostras é semelhante à dos Estados Unidos, mais especificamente de Minnesota de 1989; portanto, a linhagem que aqui circula não é semelhante à que causou a epidemia de 2014.

# 6. CONCLUSÃO

A partir dos resultados obtidos e da discussão apresentada, para uma análise em que se deseja obter qual o vírus mais abundante em uma amostra, pode-se utilizar qualquer classificador, pois todos identificaram o vírus mais abundante confirmado molecularmente, desde que o mapeador consiga remover tudo ou grande parte do genoma do hospedeiro. O Diamond falhou em indicar o TTV 3 no Pool 2 e o TTV 6 no Pool 3, ambos confirmados. Kraken2, Kaiju e CLARK apresentaram maior riqueza nos resultados, isso é bom caso se deseja fazer uma análise de possíveis vírus emergentes ou reemergentes em uma determinada população ou região. Além disso, Kaiju se mostrou um bom classificador em relação a performance do $R^2$, conforme nos mostra a Tabela 1.

Kraken2 e Kaiju foram os melhores em termos de usabilidade e também se mostraram bons classificadores.

Foi possível demonstrar a aplicação do que um bom resultado de perfil taxonômico é capaz de fazer, uma vez que foi possível montar 10 genomas virais (comprovando que estavam presentes na amostra) e fazer sua reconstrução filogenética.

# 7. REFERÊNCIAS


AGUMADU, Vivian C.; RAMPHUL, Kamleshun. Zika Virus: A Review of Literature. **Cureus**, *[S. l.]*, 2018. DOI: 10.7759/cureus.3025.

AMBROGI, Ilana G.; BRITO, Luciana; DINIZ, Debora. The vulnerabilities of lives: Zika, women and children in Alagoas State, Brazil. **Cadernos de Saúde Pública**, *[S. l.]*, v. 36, n. 12, 2020. DOI: 10.1590/0102-311x00032020.

ALI, Taccyanna. **O quê é a técnica de RT-qPCR para detecção dos casos de COVID-19?** 2020. Disponível em: https://www.igenomix.com.br/blog/o-que-e-a-tecnica-de-rt-qpcr-para-deteccao-dos-casos-de-covid-19/. Acesso em: 9 nov. 2022.

Babraham Bioinformatics - **FastQC A Quality Control tool for High Throughput Sequence Data**. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Acesso 24 de Abril de 2022.

BAĞCI, Caner; PATZ, Sascha; HUSON, Daniel H. DIAMOND+MEGAN: Fast and Easy Taxonomic and Functional Analysis of Short and Long Microbiome Sequences. **Current Protocols**, *[S. l.]*, v. 1, n. 3, 2021. DOI: 10.1002/cpz1.59. Acesso em: 28 ago. 2022.

BANKEVICH, Anton et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. **Journal of Computational Biology**, [S. l.], v. 19, n. 5, 2012. DOI: 10.1089/cmb.2012.0021.

BATUT, Bérénice *et al.*, 2022 **Quality Control** (Galaxy Training Materials). Disponível em <https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html> Acesso 06 Maio 2022.

BIGOT, Yves; SAMAIN, Sylvie; AUGÉ-GOUILLOU, Corinne; FEDERICI, Brian A. Molecular evidence for the evolution of ichnoviruses from ascoviruses by symbiogenesis. **BMC Evolutionary Biology**, *[S. l.]*, v. 8, n. 1, p. 253, 2008. DOI: 10.1186/1471-2148-8-253.

BROUWER, Lieke; MORENI, Giulia; WOLTHERS, Katja C.; PAJKRT, Dasja. World-Wide Prevalence and Genotype Distribution of Enteroviruses. **Viruses**, *[S. l.]*, v. 13, n. 3, p. 434, 2021. DOI: 10.3390/v13030434.

BUCHFINK, Benjamin; XIE, Chao; HUSON, Daniel H. Fast and sensitive protein alignment using DIAMOND. **Nature Methods**, *[S. l.]*, v. 12, n. 1, p. 59–60, 2014. DOI: 10.1038/nmeth.3176.

BUFFET-BATAILLON, Sylvie; RIZK, Guillaume; CATTOIR, Vincent; SASSI, Mohamed; THIBAULT, Vincent; DEL GIUDICE, Jennifer; GANGNEUX, Jean-Pierre. Efficient and Quality-Optimized Metagenomic Pipeline Designed for Taxonomic Classification in Routine Microbiological Clinical Tests. **Microorganisms**, [S. l.], v. 10, n. 4, p. 711, 2022. DOI: 10.3390/microorganisms10040711.

BURROWS, M.; WHEELER, D. J. A block-sorting lossless data compression algorithm. Palo Alto, Calif.: Digital, **Systems Research Center**, 1994.

CARBO, Ellen C. et al. **Performance of five metagenomic classifiers for virus pathogen detection using respiratory samples from a clinical cohort**. [s.l.] : Cold Spring Harbor Laboratory, 2022. Disponível em: http://dx.doi.org/10.1101/2022.01.21.22269647.

CDC. **Causes and Transmission of Hand, Foot, and Mouth disease**. 2022. Disponível em: https://www.cdc.gov/hand-foot-mouth/about/transmission.html. Acesso em: 25 nov. 2022.

CDC**. Enterovirus D68 (EV-D68)**. 2022. Disponível em: https://www.cdc.gov/non-polio-enterovirus/about/ev-d68.html. Acesso em: 25 nov. 2022.

CALDART, Eloiza Teles; MATA, Helena; CANAL, Cláudio Wageck; RAVAZZOLO, Ana Paula. Phylogenetic Analysis: Basic Concepts and Its Use as a Tool for Virology and Molecular Epidemiology. **Acta Scientiae Veterinariae**, *[S. l.]*, v. 44, n. 1, p. 20, 2016. DOI: 10.22456/1679-9216.81158.

CHEN, S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. **Bioinformatics**, v. 34, n. 17, p. i884–i890, 1 set. 2018.

CHIU, Charles Y.; MILLER, Steven A. Clinical metagenomics. **Nature Reviews Genetics**, [S. l.], v. 20, n. 6, p. 341–355, 2019. DOI: 10.1038/s41576-019-0113-7. Disponível em: http://dx.doi.org/10.1038/s41576-019-0113-7.

DE VRIES, Jutte J. C. et al. Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. **Journal of Clinical Virology,** [S. l.], v. 141, n. June, 2021. DOI: 10.1016/j.jcv.2021.104908.

DOMINGUEZ-BELLO, Maria Gloria; GODOY-VITORINO, Filipa; KNIGHT, Rob; BLASER, Martin J. Role of the microbiome in human development. **Gut**, [S. l.], v. 68, n. 6, 2019. DOI: 10.1136/gutjnl-2018-317503.

GALLOT-LAVALLÉE, Lucie; BLANC, Guillaume; CLAVERIE, Jean-Michel. Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. **Journal of Virology**, *[S. l.]*, v. 91, n. 14, 2017. DOI: 10.1128/jvi.00230-17.

HU, Y. F.; YANG, Fan; DU, J.; DONG, J.; ZHANG, T.; WU, Z. Q.; XUE, Y.; JIN, Qi. Complete Genome Analysis of Coxsackievirus A2, A4, A5, and A10 Strains Isolated from Hand, Foot, and Mouth Disease Patients in China Revealing Frequent Recombination of Human Enterovirus A. **Journal of Clinical Microbiology**, *[S. l.]*, v. 49, n. 7, p. 2426–2434, 2011. DOI: 10.1128/jcm.00007-11.

KATOH, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic Acids Research,** v. 30, n. 14, p. 3059–3066, 15 jul. 2002.

KENMOE, Sebastien; KENGNE-NDE, Cyprien; EBOGO-BELOBO, Jean Thierry; MBAGA, Donatien Serge; FATAWOU MODIYINJI, Abdou; NJOUOM, Richard. Systematic review and meta-analysis of the prevalence of common respiratory viruses in children . **PLOS ONE**, *[S. l.]*, v. 15, n. 11, p. e0242302, 2020. DOI: 10.1371/journal.pone.0242302.

KISELEV, Daniel; MATSVAY, Alina; ABRAMOV, Ivan; DEDKOV, Vladimir; SHIPULIN, German; KHAFIZOV, Kamil. Current trends in diagnostics of viral infections of unknown etiology. **Viruses**, 2020. DOI: 10.3390/v12020211.

LABELLA, Am; LEIVA-REBOLLO, R.; ALEJO, A.; CASTRO, D.; BORREGO, Jj. Lymphocystis disease virus (LCDV-Sa), polyomavirus 1 (SaPyV1) and papillomavirus 1 (SaPV1) in samples of Mediterranean gilthead seabream. **Diseases of Aquatic Organisms**, *[S. l.]*, v. 132, n. 2, p. 151–156, 2019. DOI: 10.3354/dao03311.

LANGMEAD, Ben; TRAPNELL, Cole; POP, Mihai; SALZBERG, Steven L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology**, *[S. l.]*, v. 10, n. 3, p. R25, 2009. DOI: 10.1186/gb-2009-10-3-r25. Acesso em: 25 abr. 2022.

LANGMEAD, Ben; SALZBERG, Steven L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, *[S. l.]*, v. 9, n. 4, p. 357–359, 2012. DOI: 10.1038/nmeth.1923.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 2009.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, *[S. l.]*, v. 25, n. 14, p. 1754–1760, 2009. DOI: 10.1093/bioinformatics/btp324.

LIN, Gu-Lung; MCGINLEY, Joseph P.; DRYSDALE, Simon B.; POLLARD, Andrew J. Epidemiology and Immune Pathogenesis of Viral Sepsis. **Frontiers in Immunology**, *[S. l.]*, v. 9, 2018. DOI: 10.3389/fimmu.2018.02147.

LUGANINI, Anna; GRIBAUDO, Giorgio. Retroviruses of the Human Virobiota: The Recycling of Viral Genes and the Resulting Advantages for Human Hosts During Evolution. **Frontiers in Microbiology**, *[S. l.]*, v. 11, 2020. DOI: 10.3389/fmicb.2020.01140.

MAHALE, Parag; TORRES, Harrys A.; KRAMER, Jennifer R.; HWANG, Lu-Yu; LI, Ruosha; BROWN, Eric L.; ENGELS, Eric A. Hepatitis C virus infection and the risk of cancer among elderly US adults: A registry-based case-control study. **Cancer**, *[S. l.]*, v. 123, n. 7, p. 1202–1211, 2017. DOI: 10.1002/cncr.30559.

MENZEL, Peter; NG, Kim Lee; KROGH, Anders. Fast and sensitive taxonomic classification for metagenomics with Kaiju. **Nature Communications**, [S. l.], v. 7, 2016. DOI: 10.1038/ncomms11257.

MORETTIN, PEDRO ALBERTO; BUSSAB, WILTON OLIVEIRA. **ESTATÍSTICA BÁSICA**. [s.l.] : Saraiva Educação S.A., 2017.

MOUSTAFA, A. et al. The blood DNA virome in 8,000 humans. **PLOS Pathogens**, v. 13, n. 3, p. e1006292, 22 mar. 2017

MULCAHY-O'GRADY, Heidi; WORKENTINE, Matthew L. The challenge and potential of metagenomics in the clinic. **Frontiers in Immunology**, [S. l.], v. 7, n. FEB, p. 1–8, 2016. DOI: 10.3389/fimmu.2016.00029

NOOIJ, Sam; SCHMITZ, Dennis; VENNEMA, Harry; KRONEMAN, Annelies; KOOPMANS, Marion P. G. Overview of virus metagenomic classification methods and their biological applications. **Frontiers in Microbiology**, [S. l.], v. 9, n. APR, 2018. DOI: 10.3389/fmicb.2018.00749.

NGUYEN, Lam Tung; SCHMIDT, Heiko A.; VON HAESELER, Arndt; MINH, Bui Quang. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. **Molecular Biology and Evolution**, [S. l.], v. 32, n. 1, 2015. DOI: 10.1093/molbev/msu300.

OLSON, Nathan D. et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. **Frontiers in Genetics**, 2015. DOI: 10.3389/fgene.2015.00235.

OUNIT, Rachid; WANAMAKER, Steve; CLOSE, Timothy J.; LONARDI, Stefano. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. **BMC Genomics**, *[S. l.]*, v. 16, n. 1, 2015. DOI: 10.1186/s12864-015-1419-2.

OUNIT, Rachid; LONARDI, Stefano. Higher classification sensitivity of short metagenomic reads with CLARK-S. **Bioinformatics**, *[S. l.]*, v. 32, n. 24, p. 3823–3825, 2016. DOI: 10.1093/bioinformatics/btw542.

**Paired-End vs. Single-Read Sequencing Technology**. [s.d.]. Disponível em: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html. Acesso em: 7 dez. 2022.

PHAN, Tung Gia et al. Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. **Virology**, *[S. l.]*, v. 482, p. 98–104, 2015. DOI: 10.1016/j.virol.2015.03.011.

PETERSEN, L. R.; BUSCH, M. P. Transfusion‐transmitted arboviruses. **Vox sanguinis,** v. 98, n. 4, p. 495-503, 2010. https://doi.org/10.1111/j.1423-0410.2009.01286.x

PISANO, María B.; GIADANS, Cecilia G.; FLICHMAN, Diego M.; RÉ, Viviana E.; PRECIADO, María V.; VALVA, Pamela. Viral hepatitis update: Progress and perspectives. **World Journal of Gastroenterology**, *[S. l.]*, v. 27, n. 26, p. 4018-4044, 2021. DOI: 10.3748/wjg.v27.i26.4018.

RAMBAU, Andrew. **FigTree**. 2007. Disponível em: http://tree.bio.ed.ac.uk/software/figtree/. Acesso em: 11 set. 2022.

RESHETNYAK, Vasiliy Ivanovich; KARLOVICH, Tatiana Igorevna; ILCHENKO, Ljudmila Urievna. Hepatitis G virus. **World Journal of Gastroenterology**, *[S. l.]*, v. 14, n. 30, p. 4725, 2008. DOI: 10.3748/wjg.14.4725.

SARAIRAH, Haneen; BDOUR, Salwa; GHARAIBEH, Waleed. The Molecular Epidemiology and Phylogeny of Torque Teno Virus (TTV) in Jordan. **Viruses**, *[S. l.]*, v. 12, n. 2, p. 165, 2020. DOI: 10.3390/v12020165.

SCHMEIER Sebastian. Computational Genomics Tutorial. **Genomics Tutorial**, 2020. Disponível em <https://genomics.sschmeier.com/index.html>. Acesso dia 24 de Abril de 2022.

**Sequencing Quality Scores**. [s.d.]. Disponível em: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html. Acesso em: 28 mar. 2022.

STRAMER, S. L. et al. Emerging infectious disease agents and their potential threat to transfusion safety. **Transfusion**, v. 49, p. 1S-29S, 2009. https://doi.org/10.1111/j.1537-2995.2009.02279.x

STRIMMER, Korbinian; VON HAESELER, Arndt. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. **Proceedings of the National**

**Academy of Sciences**, *[S. l.]*, v. 94, n. 13, p. 6815–6819, 1997. DOI: 10.1073/pnas.94.13.6815. Acesso em: 21 nov. 2022.

SUN, Jing; HU, Xiao-Yi; YU, Xiao-Fang. Current Understanding of Human Enterovirus D68. **Viruses**, *[S. l.]*, v. 11, n. 6, p. 490, 2019. DOI: 10.3390/v11060490.

VAN ETTEN, James L.; AGARKOVA, Irina V.; DUNIGAN, David D. Chloroviruses. **Viruses**, *[S. l.]*, v. 12, n. 1, p. 20, 2019. DOI: 10.3390/v12010020.

VIRILI, Camilla; FALLAHI, Poupak; ANTONELLI, Alessandro; BENVENGA, Salvatore; CENTANNI, Marco. Gut microbiota and Hashimoto 's thyroiditis. **Reviews in Endocrine and Metabolic Disorders**, 2018. DOI: 10.1007/s11154-018-9467-y

WANZELLER, A. L. M.; SOUZA, A. L. P.; AZEVEDO, R. S. S.; JÚNIOR, E. C. Sousa; FILHO, L. C. F.; OLIVEIRA, R. S.; LEMOS, P. S.; JÚNIOR, J. V.; VASCONCELOS, P. F. C. Complete Genome Sequence of the BeAn 58058 Virus Isolated from Oryzomys sp. Rodents in the Amazon Region of Brazil. **Genome Announcements**, *[S. l.]*, v. 5, n. 9, 2017. DOI: 10.1128/genomea.01575-16.

WOOD, Derrick E.; LU, Jennifer; LANGMEAD, Ben. Improved metagenomic analysis with Kraken 2. **Genome Biology**, [S. l.], v. 20, n. 1, p. 1–13, 2019. DOI: 10.1186/s13059-019-1891-0.

WOOD, Derrick E.; SALZBERG, Steven L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. **Genome Biology**, [S. l.], v. 15, n. 3, 2014. DOI: 10.1186/gb-2014-15-3-r46.

YE, Simon H.; SIDDLE, Katherine J.; PARK, Daniel J.; SABETI, Pardis C. Benchmarking Metagenomics Tools for Taxonomic Classification. **Cell**, [S. l.], v. 178, n. 4, p. 779–794, 2019. DOI: 10.1016/j.cell.2019.07.010. Disponível em: https://doi.org/10.1016/j.cell.2019.07.010.

YU, Guangchuang. Using ggtree to Visualize Data on Tree-Like Structures. **Current Protocols in Bioinformatics**, [S. l.], v. 69, n. 1, 2020. DOI: 10.1002/cpbi.96.

YU, Guangchuang; LAM, Tommy Tsan-Yuk; ZHU, Huachen; GUAN, Yi. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. **Molecular Biology and Evolution**, *[S. l.]*, v. 35, n. 12, p. 3041–3043, 2018. DOI: 10.1093/molbev/msy194.

YU, Guangchuang; SMITH, David K.; ZHU, Huachen; GUAN, Yi; LAM, Tommy Tsan‐Yuk. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. **Methods in Ecology and Evolution,** [S. l.], v. 8, n. 1, p. 28–36, 2016. DOI: 10.1111/2041-210x.12628.

YU, Yaqi; WAN, Zhenzhou; WANG, Jian-Hua; YANG, Xianguang; ZHANG, Chiyu. Review of human pegivirus: Prevalence, transmission, pathogenesis, and clinical implication. **Virulence**, *[S. l.]*, v. 13, n. 1, p. 323–340, 2022. DOI: 10.1080/21505594.2022.2029328.

ZANOTTO, Paolo Marinho de Andrade; LEITE, Luciana Cezar de Cerqueira. The Challenges Imposed by Dengue, Zika, and Chikungunya to Brazil. **Frontiers in Immunology**, *[S. l.]*, v. 9, 2018. DOI: 10.3389/fimmu.2018.01964.