FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

Programa de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

MESTRADO ACADÊMICO

METAGENÔMICA COMO FERRAMENTA DIAGNÓSTICA E IDENTIFICAÇÃO DE PATÓGENOS EMERGENTES

ALESSANDRA GONZALEZ DO NASCIMENTO

Salvador – Bahia
2023

FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

Programa de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

METAGENÔMICA COMO FERRAMENTA DIAGNÓSTICA E IDENTIFICAÇÃO DE PATÓGENOS EMERGENTES

ALESSANDRA GONZALEZ DO NASCIMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa para obtenção do grau de Mestra.

Orientador: Luciano Kalabric Silva
Coorientadora: Taryn Ariadna Castro Cuesta

Salvador – Bahia
2023

Ficha Catalográfica elaborada pela Biblioteca do
Instituto Gonçalo Moniz/ FIOCRUZ – Bahia - Salvador

N244m Nascimento, Alessandra Gonzalez do

Metagenômica como ferramenta diagnóstica e identificação de patógenos emergentes. / Alessandra Gonzalez do Nascimento. _Salvador, 2023.

98 f.: il.: 30 cm

Orientador: Luciano Kalabric Silva
Coorientadora: Taryn Ariadna Castro Cuesta

Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, 2023.

1. Metagenômica. Sequenciamento de nova geração. 3. Viroma. 4. Sequenciamento por nanoporos. 5. Protocolo de validação. I. Título.

CDU 575.113

“METAGENÔMICA COMO FERRAMENTA DIAGNÓSTICA E IDENTIFICAÇÃO DE PATÓGENOS EMERGENTES”.

ALESSANDRA GONZALEZ DO NASCIMENTO

FOLHA DE APROVAÇÃO

Salvador, 03 de fevereiro de 2023.

COMISSÃO EXAMINADORA

Dra. Jaqueline Goes de Jesus
Professora
USP

Dra. Luciane Amorim Santos
Pesquisadora
IGM/FIOCRUZ

Dr. Tiago Gräf
Pesquisador
IGM/FIOCRUZ/BA

FONTES DE FINANCIAMENTO

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Dedico este trabalho ao meu bichinho que me ensinou o que é amor de verdade. Nino, meu neném, sinto sua falta todo santo dia. Vou te amar para sempre.

AGRADECIMENTOS

Aos meus pais, Roqueline Gonzalez Mendes do Nascimento e José Alexandre Gomes do Nascimento pelo amor incondicional, incentivo a correr atrás dos meus sonhos e apoio nessa vida tão difícil e encantadora que escolhi para mim. Sou eternamente grata por todo esforço que ambos fizeram para me proporcionar uma educação de qualidade, se estou aqui academicamente hoje é porque tive meus pais me ajudando direta e indiretamente na minha jornada. Agradeço pelos valores que carrego comigo e que me fazem ser quem eu sou de verdade. Sou grata por me derem a vida e se fazerem de fato presentes em minha vida. Vocês são minha base, o meu tudo!! Não há palavras suficientes para meu amor e minha gratidão, espero um dia devolver todo bem que me fizeram. Amo vocês para todo o sempre.

A minha avó, Roselita Rocha Falcão, que me ama incondicionalmente desde que eu ainda não estava nesse mundo, que acompanhou e acompanha meu crescimento, que celebra a cada nova fase da minha vida, que me mostrou o que é ter o amor de vó e que me ensinou que família vai muito mais além do que simples laços de sangue. Obrigada por escolher ser minha avó e por me permitir ser sua neta. Obrigada por me mostrar que o amor de verdade supera qualquer adversidade e que o amor quando é dado de bom coração sempre encontra um caminho de volta até nós.

Ao meu primo, Caio Gonzalez, por ser meu irmão, por me escolher como sua irmã, pelo apoio, por todas as nossas tardes jogando videogame e conversando, pelas boas memórias na minha infância e adolescência, por todo amor de irmão envolvido todos esses anos, por ser sempre a minha melhor companhia e por estar sempre juntinho comigo. Sinto muito orgulho de acompanhar teu crescimento pessoal e profissional, sinto orgulho do ser humano gentil, carinhoso, dedicado e amável que você é, saiba que te ter em minha vida é sinônimo de benção e de nunca estar só, pois sempre estarei com você, irmãozinho.

Ao meu querido orientador, Dr. Luciano Kalabric, por me escolher e me acolher como sua mestranda e ao longo desses dois anos de pós-graduação, por me mostrar que é possível sim ter uma relação saudável entre orientador e orientando, por ser minha inspiração na pesquisa. Aprendi muito ao seu lado, sobre minha pesquisa e principalmente sobre a vida. Muito obrigada pela oportunidade de me tornar uma pesquisadora e, principalmente, uma pessoa melhor. Obrigada também por toda confiança depositada em mim.

A minha coorientadora, Dra. Taryn Castro, por ser minha parceira de laboratório, minha mentora e minha motivação diária, por me mostrar humildade e gentileza mesmo diante dos nossos desafios laboratoriais. Obrigada por estar ao meu lado nas longas horas de experimento

e por continuar ao meu lado mesmo quando não estávamos sob o mesmo fuso horário. Sempre irei torcer pelo seu sucesso, obrigada por me coorientar maravilhosamente.

Ao meu amigo, Lucas Burgos, por me aturar desde o primeiro dia da faculdade, por ter me mostrado a oportunidade de fazer esse mestrado, por suportar todos os meus surtos todos esses anos, pela conexão inexplicável que temos e acima de tudo, por sempre estar ao meu lado e nunca desistir ou enjoar da gente. Sua amizade foi o que me ajudou a ficar de pé nos dias mais escuros, foi o que me fez sorrir ainda mais em dias felizes, foi o que me motivou a encarar todas as dificuldades e concluir mais esse capítulo no meu livro da vida.

A minha amiga, Gabriela Grimaldi, por ser minha eterna parceira de "bora? bora!", por estar ao meu lado e literalmente no laboratório ao lado do meu também, por sempre me dar forças, por sempre torcer por mim, por dividir tantos momentos especiais juntas, por aturar minhas crises de realismo com a graciosidade de um raio de sol nas primeiras horas da manhã. Nossa amizade é linda demais, sou muito mais feliz por te chamar de amiga.

As minhas amigas, Alice Oliveira e Natália Soares, por essa amizade linda, doida e inabalável desde que nos conhecemos 7 anos (e contando) atrás lá no cursinho pré-vestibular. Vocês fazem parte da minha vida e da minha história, sou muito grata por ter conhecido vocês e mais ainda por ver nossa amizade florescendo e evoluindo ao longo dos anos, cada uma em um caminho, mas o sentimento verdadeiro não muda.

Aos meus poucos e bons amigos de escola, Allan Maciel, Elvis Pfaffenseller e Leonardo Felipe, por essa amizade pura e verdadeira que ultrapassa distâncias, séries, escolas e muitos anos. Sou muito feliz por ter vocês em minha vida, por saber que o sentimento não se altera, independente da circunstância de nossas vidas. A amizade não muda, ela permanece e se fortalece. Obrigada pelo carinho e companheirismo de sempre.

A todos os meus familiares, conhecidos e amigos (já que não conseguiria incluir o nome de todos aqui), que torceram e acompanharam a saga desse mestrado, vocês sabem muito bem quem são. Obrigada por toda paciência com meus sumiços, compreensão ao me ver desmarcar compromissos por precisar estar no laboratório ou escrevendo, pelo apoio nessa vida difícil de ser pesquisador no Brasil e pelo incentivo a alcançar meus objetivos.

Aos meus colegas do grupo Virologia Molecular, por partilharem conhecimento, vivências e muitas horas em reuniões, por formarem um grupo coeso e dedicado em todas as atividades de pesquisa que estamos envolvidos.

Ao laboratório de patologia e biologia molecular (LPBM), por me acolher, tanto em estrutura física quanto em auxílio dos estudantes e pesquisadores nesses anos dedicados à minha pesquisa.

Ao Dr. Federico Costa, Dr. Nathan Grubaugh e Dr. Joseph Fauver pelo apoio financeiro e instrumental para a realização desse projeto.

Ao Instituto Gonçalo Moniz (IGM) – Fiocruz Bahia pela infraestrutura de laboratórios e plataformas que permitiram a realização desta pesquisa.

Ao Programa de Pós-graduação em Biotecnologia em Saúde e Medicina Investigativa (PgBSMI) pela oportunidade de cursar uma pós-graduação stricto sensu.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo fornecimento da minha bolsa de mestrado que me permitiu seguir na área de pesquisa apesar de todas as adversidades no nosso país nesse período.

Por fim, gratidão a todos que puderam, direta ou indiretamente, contribuir com o início, andamento e finalização desse mestrado.

“E um homem não me define
Minha casa não me define
Minha carne não me define
Eu sou meu próprio lar.”

Francisco, el Hombre – Triste, Louca ou Má

NASCIMENTO, Alessandra Gonzalez do. **Metagenômica como ferramenta diagnóstica e identificação de patógenos emergentes**. 2023. 98 f. il. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) - Fundação Oswaldo Cruz, Centro de Pesquisas Gonçalo Moniz, Salvador, 2023.

RESUMO

INTRODUÇÃO: A análise metagenômica de dados de tecnologias de sequenciamento de nova geração (NGS) tem fornecido oportunidades para o diagnóstico e vigilância em saúde pública de patógenos conhecidos e ou emergentes em amostras clínicas. **OBJETIVO**: Neste trabalho objetivou-se padronizar e validar um método de sequenciamento de nova geração (NGS) para análise metagenômica viral utilizando o MinION. **MATERIAL E MÉTODOS**: Dados de treinamento foram utilizados em *benchmarks*, para testar os diferentes *workflows* (*wf*) de bioinformática e otimizar o tempo de processamento das análises de reconhecimento das bases (*basecalling*) e demultiplexação. A principal diferença entre os *wfs* consistiu no *software* utilizado para classificação taxonômica sendo: primariamente, utilizamos o Kraken2 no *wf1*, que classifica vírus e bactérias a partir de um grande banco de dados; e o BLAST no *wf2*, que utiliza um banco local apenas com sequencias de referência virais de interesse (painel); e, alternativamente, utilizamos o Genome Detective no *wf3*, que é uma ferramenta web para análise de sequencias virais; e, por fim, o Epi2ME no *wf4*, para uma análise rápida e automatizada. Um ensaio de diluição seriada utilizando a vacina de poliovírus atenuados (OPV) foi realizado para testar o limiar de detecção dos nossos métodos. Amostras clínicas retrospectivas e recentes de arboviroses, casos suspeitos de meningites virais, infecções agudas do trato respiratório e infecções crônicas causadas por hepatites virais e HIV, além de amostras de isolados virais foram testadas. O RNA viral foi enriquecido pela depleção do DNA por DNAses. O cDNA foi sintetizado por transcrição reversa e amplificado por SISPA antes da preparação das bibliotecas e sequenciamento num dispositivo portátil MinION (ONT). Uma análise de acurácia foi realizada utilizando-se diferentes *cut-off* (0,0% a 5,0%) da abundância relativa em nível de *reads* e *contigs* classificadas para eliminar "ruídos" ou artefatos do sequenciamento. **RESULTADOS**: O *basecalling* para o modelo *fast* foi otimizado, 1,36 horas, e para o modelo *hac*, 30,72 horas, que representa redução de 10% e 81%, respectivamente, no tempo médio de processamento dos dados de treinamento. Entretanto, o modelo *fast* foi preferido por ter melhor relação do custo computacional e acurácia. Foi possível identificar os enterovírus presentes na OPV até a diluição de 10^{-4} tanto pelo *wf1* quanto pelo *wf2*. Ao todo, seis bibliotecas foram preparadas contendo 35 amostras clínicas com suspeita de infecção por vírus RNA. Tanto em nível de *reads* quanto *contigs*, o *wf3* apresentou a melhor acurácia (70%) e (67%), respectivamente, utilizando um *cut-off* ≥ 1,0%. O *wf4* não disponibiliza as *reads* classificadas para montagem e análise em nível de *contigs*. **CONCLUSÕES**: Os resultados deste estudo demonstram que a utilização de uma metodologia de NGS metagenômica possibilita o diagnóstico acurado de patógenos virais de importância clínica. O MinION foi capaz de realizar a análise do viroma e diagnóstico de infecções virais com acurácia. Avanços nessa metodologia podem reduzir custos e poderão possibilitar viabilizar sua utilização na rotina de diagnóstico, com a vantagem de permitir a vigilância e descoberta de agentes emergentes.

Palavras-chave: Metagenômica. Sequenciamento de nova geração. Viroma. Sequenciamento por nanoporos. Protocolo de validação.

NASCIMENTO, Alessandra Gonzalez do. **Metagenomic as a diagnostic tool and identification of emerging pathogens**. 2023. 98 f. il. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) - Fundação Oswaldo Cruz, Centro de Pesquisas Gonçalo Moniz, Salvador, 2023

ABSTRACT

INTRODUCTION: The metagenomic analysis of data from next-generation sequencing technologies (NGS) has provided opportunities for the diagnosis and public health surveillance of known and/or emerging pathogens in clinical samples. **OBJECTIVE**: This work aimed to standardize and validate NGS method for viral metagenomic analysis using MinION. **MATERIAL AND METHODS**: Training data were used in benchmarks, to test the different bioinformatics workflows (wf) and to optimize the processing time of the analysis of bases recognition (basecalling) and demultiplexing. The main difference between the wfs consisted of the software used for taxonomic classification: primarily, we used Kraken2 in wf1, which classifies viruses and bacteria from a large database; and BLAST in wf2, which uses a local bank with only viral reference sequences of interest (panel); and, alternatively, we use Genome Detective in wf3, which is a web tool for viral sequence analysis; and finally, Epi2ME in wf4 for fast automated analysis. A serial dilution assay using the attenuated poliovirus oral vaccine (OPV) was performed to test the detection threshold of our methods. Retrospective and recent clinical samples of arboviruses, suspected cases of viral meningitis, acute respiratory tract infections and chronic infections caused by viral hepatitis and HIV, in addition to samples of viral isolates were tested. Viral RNA was enriched by DNAse depletion. cDNA was synthesized by reverse transcription and amplified by SISPA prior to library preparation and sequencing in a portable MinION device (ONT). An accuracy analysis was performed using different cut-offs (0.0% to 5.0%) of the relative abundance at the level of classified reads and contigs to eliminate "noise" or sequencing artifacts. RESULTS: The basecalling for the fast model was optimized, 1.36 hours, and for the hac model, 30.72 hours, which represents a reduction of 10% and 81%, respectively, in the average processing time of the training data. However, the fast model was preferred because it had a better relation between computational cost and accuracy. It was possible to identify the enteroviruses present in OPV up to a 10^{-4} dilution by both wf1 and wf2. In all, six libraries were prepared containing 35 clinical specimens with suspected RNA virus infection. Both in terms of reads and contigs, wf3 showed the best accuracy (70%) and (67%), respectively, using a cut-off $\geq$ 1.0%. wf4 does not make sorted reads available for contig-level assembly and parsing. **CONCLUSIONS**: The results of this study demonstrate that the use of a metagenomic NGS methodology enables the accurate diagnosis of clinically important viral pathogens. The MinION was able to accurately perform virome analysis and diagnosis of viral infections. Advances in this methodology can reduce costs and enable its use in routine diagnosis, with the advantage of allowing surveillance and discovery of emerging agents.

Keywords: Metagenomics. Next generation sequencing. Virome. Nanopore sequencing. Validation protocol.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1 **INTRODUÇÃO**

Durante uma vigilância em saúde pública prolongada há certas lacunas que permitem a disseminação de surtos locais e que estão relacionadas a desafios durante diagnósticos clínicos primários, como 1) a sobreposição de apresentações clínicas causadas por múltiplos patógenos, 2) patógenos que causam formas novas ou raras de uma doença e 3) o surgimento de patógenos novos ou inesperados não investigados por diagnósticos de rotina. Como pode ser visto, o diagnóstico laboratorial é crucial para identificação dos agentes causadores de doenças infecciosas e a implementação de intervenções efetivas.

Infelizmente, o diagnóstico laboratorial tende a ser limitado pela falta de métodos amplamente padronizados e validados localmente para identificação de alguns agentes circulantes e é inexistente para alguns agentes emergentes. Algumas limitações das atuais metodologias de diagnóstico, como a necessidade de conhecimento prévio do patógeno, a sensibilidade dos métodos e a qualidade da amostra determinam que em 20% a 60% dos casos permanecem sem diagnóstico (FORBES et al., 2017; VAN GAGELDONK-LAFEBER et al., 2005; GLASER et al., 2006; GLASER; BLOCH, 2009; THOMAS et al., 2015).

Os testes microbiológicos convencionais em uso detectam um número limitado de um painel de patógenos e requerem que o microrganismo seja cultivado com sucesso a partir da amostra clínica (CHIU; MILLER, 2019). O diagnóstico de infecções virais é ainda mais limitado pela dificuldade em encontrar sistemas de isolamento viral específicos e disponibilidade de métodos de diagnóstico, sobretudo para agentes emergentes. A análise metagenômica de dados NGS é particularmente atraente para o diagnóstico e a vigilância em saúde pública de doenças infecciosas porque a abordagem pode detectar amplamente vírus, bactérias e parasitas em amostras clínicas a partir de sequências genômicas desses agentes (PALLEN, 2014; MILLER et al., 2013).

Em todos os cenários apresentados, a metagenômica incorporada em um programa clínico de diagnóstico e vigilância pode ajudar a detectar e identificar patógenos causadores de doenças infecciosas graves. Entretanto, a utilização da metagenômica voltada para o estudo da diversidade viral é bastante complexa e a padronização e validação de técnicas que se utilizam dessa tecnologia ainda não estão bem estabelecidos. Em suma, para identificar surtos, prevenir a disseminação de doenças e oferecer o tratamento adequado aos pacientes, faz-se necessário a utilização de técnicas capazes de diagnosticar os agentes virais causadores de doenças infecciosas graves.

2 **REVISÃO DA LITERATURA**

2.1 DIAGNÓSTICO DE DOENÇAS INFECCIOSAS

2.1.1 **Breve histórico do diagnóstico clínico**

Nos últimos 3.000 anos o diagnóstico de doenças se modificou e evoluiu bastante. De acordo com a Organização Mundial da Saúde, o conceito de saúde se define como o estado de completo bem-estar físico, mental e social e não mera ausência de moléstia ou enfermidade (WORLD HEALTH ORGANIZATION, 1948).

Entretanto na antiguidade, acreditava-se que uma doença era determinada por desequilíbrios entre os humores e a medicina era de cunho mágico e/ou religioso, onde os deuses possuíam capacidade de curar ou provocar doenças (QUARESMA, 2011). Foram tempos emblemáticos do horror aos sintomas ao pavor por um sentimento de culpabilidade individual e coletiva" (LE GOFF, 1991). Muitos anos se passaram até que a partir da segunda metade do século XIX é consolidada a teoria que as doenças são geradas por agentes microscópicos espalhados no ar, na água e no próprio ser humano, fundamentando as bases para a saúde pública contemporânea (QUARESMA, 2011).

A ciência do diagnóstico foi avançando a partir do desenvolvimento e aperfeiçoamento de instrumentos de precisão, tais como o estetoscópio, microscópio, termômetro, entre outros equipamentos médicos são responsáveis pela constante evolução no diagnóstico de doenças.

Além da descoberta e implantação de diferentes técnicas de cultivo de microrganismos para fins de estudos. A medicina moderna possui uma dimensão social e coletiva, ao invés de individual (FOUCAULT, 1992). O exame clínico é uma ferramenta diagnóstica poderosa para o médico clínico e é bastante comum ouvirmos de diversos médicos e estudantes de medicina que "a clínica é soberana". Entretanto com o avanço da tecnologia, os diagnósticos deixaram de ser exclusivamente clínicos e ganharam o suporte de testes diagnósticos diversos.

2.1.2 **Técnicas de diagnóstico para infecções virais**

Desde seu surgimento, há mais de 3 bilhões de anos, os microrganismos colonizaram completamente o nosso planeta (ALTERMANN; KAZMIERCZAK, 2003; ALLWOOD et al., 2006). O termo infecção vem do latim *infectio* que primariamente era uma palavra usada para denotar "contaminação" em geral, inclusive no campo das ideias. Contudo, uma doença

infeciosa pode ser definida como uma doença que pode ser provocada por um patógeno (vírus, parasita, bactéria ou fungo) que invade o organismo e causa algum tipo de prejuízo ao hospedeiro (QUARESMA, 2011).

É possível estimar que apenas 0,001% de um trilhão de microrganismos contidos no planeta Terra foram identificados até o momento devido ao foco da maioria dos estudos ser em organismos patogênicos de interesse. (LOCEY; LENNON, 2016; WHITMAN; COLEMAN; WIEBE, 1998). A OMS publicou em 2010 o primeiro relatório sobre doenças negligenciadas no mundo. Sendo 17 doenças na categoria de negligenciadas e pelo menos 12 delas são endêmicas no Brasil (WORLD HEALTH ORGANIZATION, 2010). Diversas doenças infecciosas estão sob controle graças as campanhas de vacinação tais como: poliomielite, varíola, sarampo, difteria, rubéola, entre outras, entretanto, não podemos ignorar a frequência do surgimento de infecções emergentes no mundo e o aumento de dispersão e prevalência em todo o mundo (ZLOJUTRO; REY; GARDNER, 2019).

As doenças infecciosas acompanharam a humanidade ao longo da sua evolução, já que toda doença infecciosa precisa de determinada população humana ou animal para se manter (OSTERHOLM; OLSHAKER, 2020). Mesmo que muitas vezes mal interpretadas, impulsionaram e possibilitaram o avanço e modernização de técnicas para que pudéssemos compreendê-las melhor e desenvolver novas tecnologias. Estima-se que nas últimas duas décadas mais de 15 milhões de mortes no mundo foram associadas a doenças infecciosas (FAUCI; MORENS, 2012), sendo que ainda continuam representando um fardo significativo nos sistemas de saúde de todo o mundo (HAN et al., 2019).

Segundo Osterholm e Olshaker (2020), há apenas quatro eventos que têm de fato o poder de afetar negativamente todo o planeta: 1) uma guerra termonuclear generalizada, 2) o choque de um asteroide com a Terra, 3) mudanças climáticas globais, 4) uma doença infecciosa. As infecções respiratórias ocupam um amplo espaço entre as infecções e merecem destaque devido a facilidade de transmissão por meio das vias aéreas superiores em todas as faixas etárias, principalmente em populações de risco como exemplo: crianças, idosos e imunossuprimidos (TEIXEIRA, 2007).

Os vírus são agentes passivos, portanto, precisam de auxílio de outro ser vivo para se utilizar da maquinaria celular hospedeira para se reproduzir (UJVARI, 2012). É comum ouvir que os vírus são uma ameaça à humanidade (KAZ, 2020). Porém é importante ressaltar que se estima que o DNA do nosso genoma é composto por aproximadamente 8% de material genético de vírus ancestrais. (MACFARLAN et al. 2012). Os vírus de fato são perigosos, contudo, só é possível confrontar e lidar com uma ameaça se a conhecemos. As doenças infecciosas estão

sempre presentes se entendermos melhor sobre o que está por ai, mundo afora, identificando e caracterizando patógenos, conseguimos encarar os desafios da saúde pública de forma que o impensável não se torne o inevitável (OSTERHOLM; OLSHAKER, 2020).

Classicamente, a base da detecção viral por décadas foi o isolamento viral por meio do cultivo de células, limitando a avaliação da variedade de vírus existente (MITCHELL; GLANVILLE, 2018). Este foi um procedimento bastante utilizado por volta de 1970 devido à disponibilidade dos reagentes necessários altamente purificados e linhagens celulares preparadas e vendidas comercialmente (HSIUNG, 1984). Apesar de viabilizar o diagnóstico de infecções virais, a cultura celular sozinha muitas vezes não permite fornecer características confirmatórias sobre a identidade molecular do patógeno (LELAND; GINOCCHIO, 2007).

Há outros dois métodos utilizados, relacionados à cultura celular, para a detecção da presença de vírus, são eles a hemaglutinação e a hemadsorção. Ambas se utilizam de hemácias para a realização dos testes, sendo respectivamente, um teste onde hemácias são colocadas em contato com uma suspensão de viral e outro onde alguns vírus possuem hemaglutininas capazes de fazer hemácias aderirem à superfície das células em cultura. (BERNADELLI, 2022).

As primeiras imagens de vírus foram obtidas após a invenção do microscópio eletrônico em 1931 pelos engenheiros Ernst Ruska e Max Knoll. (WELLE, 2005). A microscopia eletrônica revolucionou o campo da virologia, pois permitiu a observação direta do vírus, facilitando sua diferenciação a partir de sua morfologia (BERNADELLI, 2022). A diferença básica entre os dois tipos de microscópio está na formação da imagem: enquanto o óptico emprega um feixe de luz, o eletrônico emite elétrons. A microscopia eletrônica possui baixa sensibilidade quando comparada a outros métodos diagnósticos e possui custo elevado, todavia é eficiente para a detecção de algumas infecções virais, principalmente em ambientes de pesquisa (WELLE, 2005).

Uma potente ferramenta diagnóstica são os testes sorológicos. A maioria deles buscam detectar anticorpos específicos de algum patógeno no sangue do hospedeiro. Há também a possibilidade de detecção direta do antígeno de interesse. No soro diagnóstico de vírus há diversas técnicas como: ensaio imunoenzimático (ELISA), teste de aglutinação do látex, imunofluorescência, immunoblotting (*Western Blotting*), ensaios de quimioluminescência, teste de neutralização, entre outros (SANTOS et al., 2021). Sendo o ELISA, uma técnica amplamente utilizada em laboratórios clínicos em pesquisa, permitiu a possibilidade de quantificação de anticorpos e antígenos das amostras, diferentemente das técnicas anteriores que em sua maioria eram unicamente qualitativas (KORSMAN, 2014).

A luta dos seres humanos contra vírus é um processo em evolução (LIN; HUI; MAO,

2021). Com a constante busca por aprimoramentos nos exames laboratoriais, foi possível unir essa necessidade com os avanços da tecnologia. Assim surgem as técnicas de biologia molecular voltadas para a identificação e estudo das infecções virais dependendo apenas das sequências únicas de material genético do patógeno (BERNADELLI, 2022).

Não é novidade o quão indispensáveis são os exames laboratoriais, clinicamente, na maioria das vezes, não é possível distinguir infecções virais de bacterianas (OSTERHOLM; OLSHAKER, 2020). A capacidade de detectar potenciais patógenos em uma amostra, sejam eles bactérias, vírus, fungos ou parasitas e correlacionar com a resposta geral do hospedeiro possui uma grande utilidade potencial no diagnóstico de doenças infecciosas (CHIU; MILLER, 2019).

2.1.2.1 Diagnóstico molecular

A genômica, campo da ciência que estuda o código genético de um ser vivo, afetou profundamente o estudo da microbiologia nos levando a revisar nossos conceitos não apenas do que é um microrganismo e o que ele faz, mas também da melhor maneira de abordar o estudo dos mesmos. O século XXI é considerado por muitos autores como “Século do Genoma”, fruto de inúmeros avanços científicos e grande desenvolvimento na área da microbiologia devido as várias descobertas de rastreamento genético de microrganismos (ARAÚJO, 2004; MOORE, 1996).

Na última década, as técnicas moleculares revolucionaram a microbiologia e provavelmente continuarão a fazê-lo durante a próxima década. As informações que estão sendo adicionadas aos bancos de dados de sequências estão aumentando exponencialmente e a cada dia estamos em uma posição melhor para descrever microrganismos por seu genoma e até mesmo ambientes microbianos complexos por seu metagenoma (MEDINI et al., 2008).

A genômica pode ter facilitado ou encorajado o desenvolvimento de abordagens tanto dependentes quanto independentes de cultivo microbiológico, abordagens diagnósticas essas que muitas vezes eram consideradas por seus proponentes como rivais ao invés de estratégias complementares (WARD; FRASER, 2005).

Em 1953 o biólogo americano James Watson e o físico britânico Francis Crick elaboraram o modelo da dupla hélice de DNA “descobrindo-o” (WATSON; CRICK, 1953). Vários estudos se sucederam até a elucidação da estrutura do DNA que conhecemos hoje, entre eles é essencial destacar o bioquímico austríaco Erwin Chargaff. Em seu estudo primeiramente ele afirmou que a composição de nucleotídeos do DNA varia entre as espécies. Ele também

concluiu que a quantidade da base adenina costuma ser semelhante à quantidade de timina e a quantidade de guanina costuma ser semelhante à quantidade de citosina. (CHARGAFF, 1950).

Surge assim a regra de Chargaff que mostra que a quantidade total de purinas (adenina e guanina) é igual a quantidade total de pirimidinas (timina e citosina). A descoberta de Chargaff junto com alguns resultados vitais de cristalografia obtidos pelos pesquisadores ingleses Rosalind Franklin e Maurice Wilkins, estabeleceram uma base sólida para a descoberta do modelo tridimensional, em dupla hélice, da estrutura do DNA proposto por Watson e Crick. (PRAY, 2008).

A partir do conhecimento sobre a estrutura do DNA foi possível iniciar a compreensão de muitos aspectos de sua função, replicação e produção de proteínas celulares. Sem dúvidas, tais descobertas abriram as portas de uma nova era de descobertas na biologia molecular, possibilitando a fundamentação da maioria das pesquisas que envolvem e aplicam biologia molecular em diversas áreas da saúde e biotecnologia.

2.1.2.1.1 *Reação em Cadeia da Polimerase (PCR)*

A Reação em Cadeia da Polimerase (do inglês *Polymerase Chain Reaction* - PCR) é um método que consiste em fazer milhões de copias do material genético e foi desenvolvido por cientistas em 1983 e representa uma das mais significantes descobertas do século XX (MULLIS et al. 1986). Essa técnica envolve a interação recíproca de dois oligonucleotídeos e os produtos de extensão da DNA polimerase cuja síntese eles iniciam, quando são hibridizados com diferentes cadeias de um molde de DNA em uma orientação relativa de modo que seus produtos de extensão se sobreponham (MULLIS et al. 1986). O método consiste em ciclos repetitivos de desnaturação, hibridação e extensão da polimerase via aquecimento e resfriamento da amostra em ciclos feitos manualmente. A cada ciclo percebe-se um aumento exponencial da concentração daquele DNA alvo.

Em 1985, a enzima DNA polimerase da bactéria de fontes termais *Thermus aquaticus* (Taq) foi isolada com sucesso por Susanne Stoffel, tal enzima tem a capacidade de se manter estável em temperaturas de até 117°C (KUBISTA, 2012). Adicionalmente, David Gelfand e Randy Saiki provou que esta enzima poderia ser usada para automatizar o processo (SAIKI et al., 1985). Esta técnica rapidamente se tornou importantíssima nas pesquisas para identificação de espécies de patógenos, em investigações forenses e em testes qualitativos para doenças infecciosas. (KUBISTA, 2012). Em 1989, o DNA *Thermal Cycler* apareceu como o primeiro termociclador automático. Desde então a maioria dos termocicladores funcionam com o mesmo

princípio de ciclagens pode ser dividida em três etapas: desnaturação, anelamento e extensão (SANTOS et al., 2014), tendo o aquecimento por resistências elétricas e refrigeração com ventoinhas e tubulações de serpentina preenchidas com etileno glicol (BONECKER, 2020).
Várias técnicas foram desenvolvidas envolvendo, aperfeiçoando e customizando a PCR. A PCR em tempo real quantitativa (qPCR) representou um importante avanço à técnica de PCR, pois se trata de um método confiável de detecção e quantificação dos produtos amplificados, simultaneamente, podendo ser visualizado em tempo real à amplificação das sequencias de interesse por meio da emissão e captação de fluorescência que ocorre durante todo o processo, diferentemente da PCR convencional (ZAHA; FERREIRA; PASSAGLIA, 2012).

Há também a PCR multiplex, permitem que vários alvos diferentes sejam testados e identificados em uma única reação (MITCHELL; GLANVILLE, 2018), e a RT-qPCR que é a PCR de transcrição reversa quantitativa, onde o material genético base é o cDNA sintetizado a partir do RNA (BONECKER, 2020). Atualmente a PCR é o método diagnóstico de rotina para identificar e amplificar rapidamente sequências específicas de patógeno, e também é o método mais comum para identificar infecções virais (LEFTEROVA et al., 2015; BONECKER, 2020).

Todas as técnicas de PCR possuem uma limitação básica da necessidade de primers específicos para o patógeno-alvo para que seja executada a técnica de acordo com a hipótese formada no diagnóstico clínico primário, limitando ainda mais o diagnóstico de doenças virais onde o agente causador ainda não é conhecido (BERNADELLI, 2022). A introdução de técnicas de detecção viral de base molecular, como a PCR, revolucionou o campo do diagnóstico viral (MITCHELL; GLANVILLE, 2018).

Do princípio da cultura de células como o "padrão-ouro" para diagnóstico de infecções virais, passando por técnicas sorológicas e chegando às técnicas moleculares, é possível afirmar que esses últimos revolucionaram o diagnóstico das infecções virais, permitindo resultados mais rápidos e precisos. Entretanto, o mais "recente" recurso molecular, o sequenciamento genômico, vem ganhando espaço e notoriedade no ramo do diagnóstico (MITCHELL; GLANVILLE, 2018).

2.1.2.1.2 *Sequenciamento de Nova Geração (NGS)*

Saber a sequência exata de nucleotídeos numa molécula de DNA é uma valiosa ferramenta para investigar a informação genética armazenada dentro dos organismos. A partir dessa necessidade de conhecimento, surge as primeiras técnicas de sequenciamento de DNA que são denominadas como "primeira geração", dentre elas uma das mais importantes e

revolucionárias, o sequenciamento Sanger.

Sanger e Coulson em 1977 inventaram o didesoxinucleotideo (ddNTP), o nucleotídeo terminador de cadeia, que a ausência da hidroxila (OH) da pentose no carbono 3' e se for incorporado na cadeia para o crescimento da cadeia parando assim a reação naquela sequência de DNA (SANGER; NICKLEN; COULSON, 1977). O método de Sanger foi aprimorado e é um dos métodos de sequenciamento de DNA mais usado até os dias de hoje. Para a reação é necessário o DNA de interesse a ser sequenciado, primers, enzima DNA-polimerase, 4 desoxinucleotídeos trifosfatos (ddGTP, ddATP, ddTTP, ddCTP) e ddNTP marcado radiotivamente ou por um corante fluorescente. Os ddNTPs se ligam de forma aleatória e no final do processo da PCR de sequenciamento há produção de fragmentos com comprimentos variáveis, que podem ser separados por eletroforese num gel de poliacrilamida ou por eletroforese capilar para revelar a sequência do DNA (figura 1).

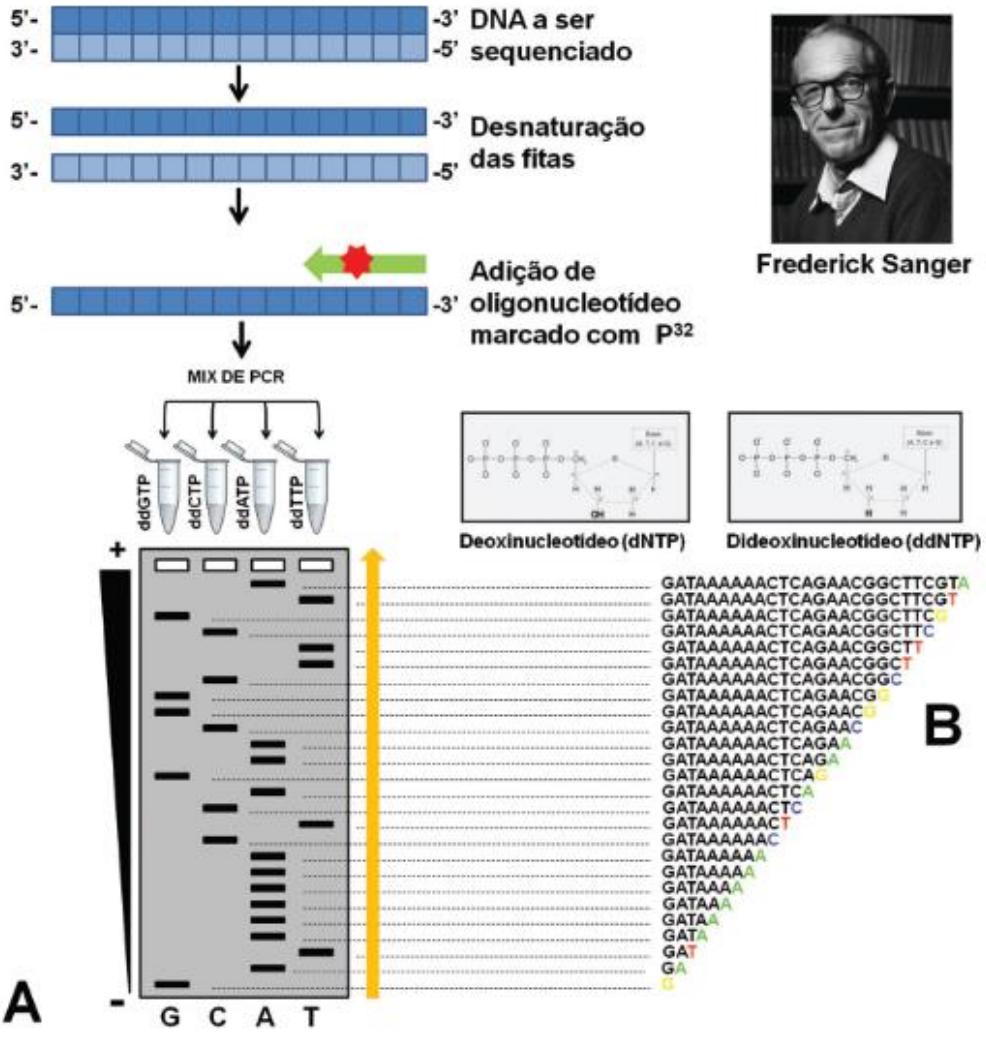


Figura 1 - Sequenciamento Sanger

[1]**Fonte**: (MOREIRA, 2015)

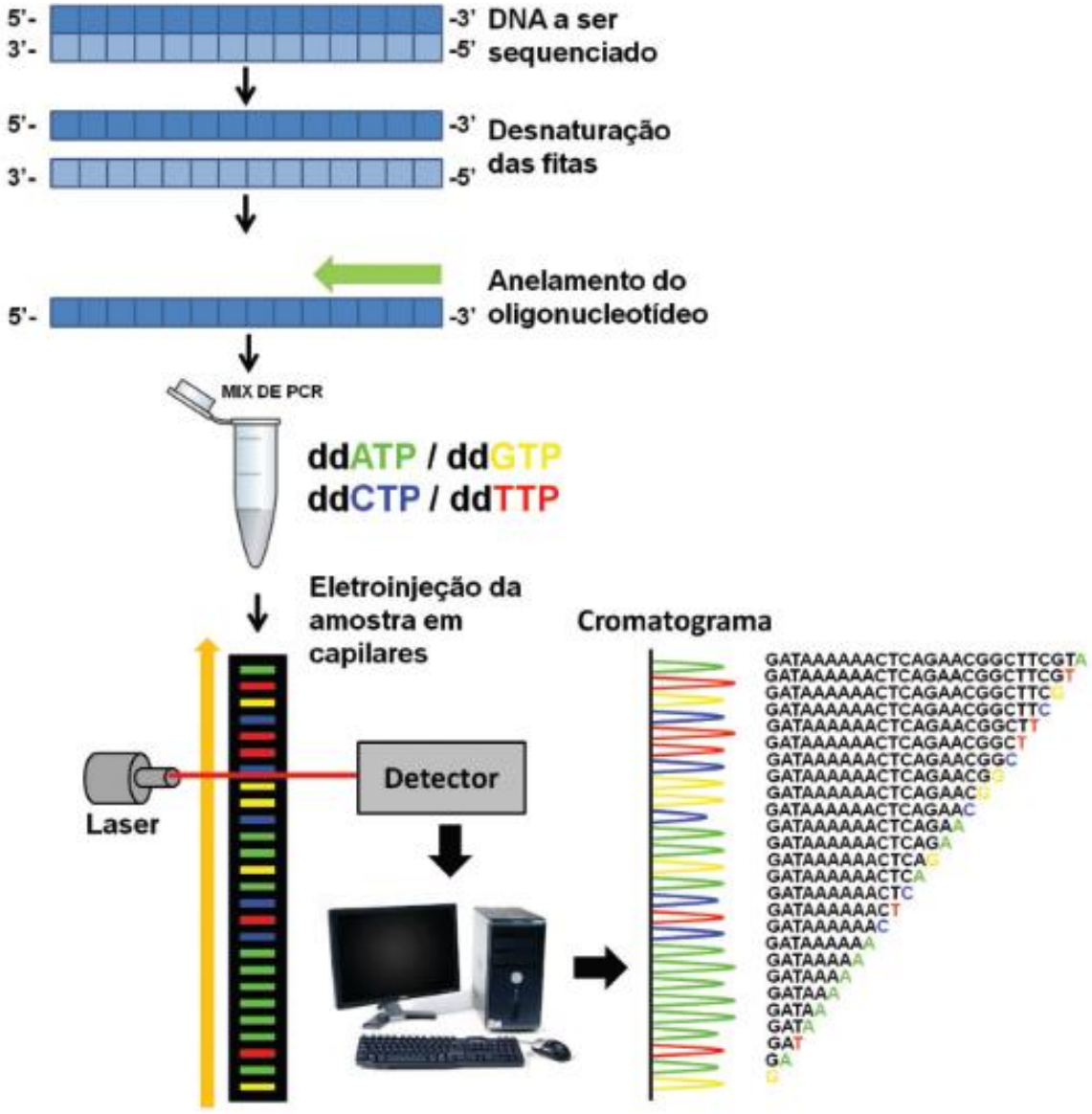


Figura 2 - Automatização do sequenciamento Sanger
[2]**Fonte**: (MOREIRA, 2015)

[1] Livro Ciências Genômicas: fundamentos e aplicações**.** Este método foi gradualmente aprimorado (figura 2) e tornou-se automatizado, a primeira máquina de sequenciamento automático, AB370, foi introduzida em 1987 pela Applied Biosystems (NOWROUSIAN, 2010).

[2] Livro: Ciências Genômicas: fundamentos e aplicações. O sequenciamento Sanger foi o método de utilizado para projetos de sequenciamento em grande escala da época, por exemplo, o Projeto Genoma Humano que teve como objetivo o sequenciamento dos 3,1 bilhões de bases nitrogenadas do genoma humano (GÓES; OLIVEIRA, 2014). Como é possível acompanhar na figura 3 a história dos avanços no sequenciamento do genoma, a partir do sequenciamento Sanger, levou ao desenvolvimento de tecnologias de sequenciamento de próxima geração (HEAHER; CHAIN, 2015; MOREY et al., 2013).

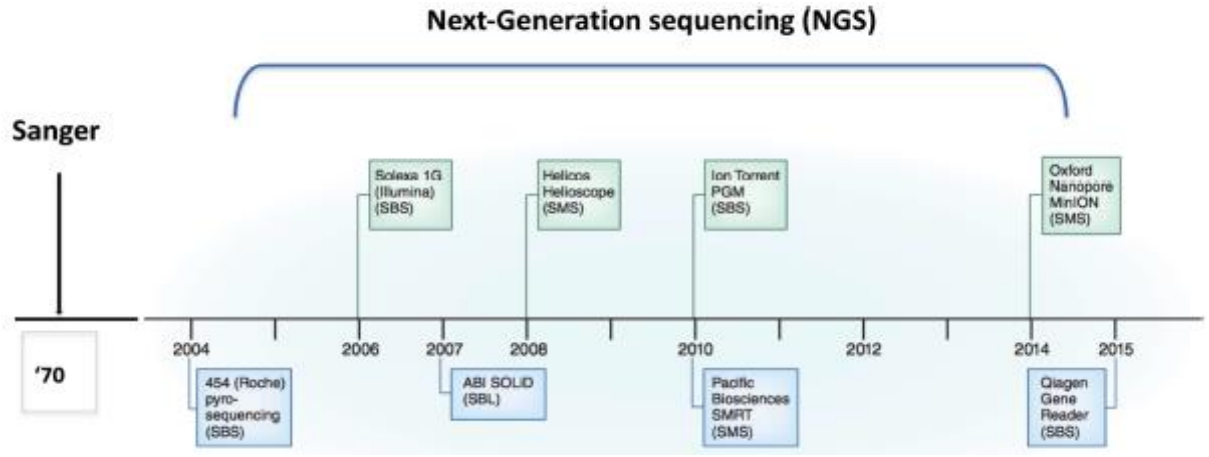


Figura 3 - Linha do tempo do NGS.
Fonte: (DE JESUS et al., 2019).

O sequenciamento Sanger representa a chamada tecnologia sequenciamento de primeira geração (LIN; HUI; MAO, 2021). Já as tecnologias de sequenciamento de segunda e terceira geração, caracterizadas por alto rendimento, também são denominadas como tecnologias de sequenciamento de próxima geração (MARDIS, 2008).

O termo NGS (do inglês *Next Generation Sequencing*) se refere a uma coleção de tecnologias que utilizam abordagens de sequenciamento que produzem milhões de sequências curtas ou longas de leitura em um tempo muito menor, a um custo muito mais barato e com maior rendimento em comparação com o sequenciamento Sanger (KANZI et al., 2020). As tecnologias NGS representam um marco para o avanço do diagnóstico de doenças infecciosas (BERTELLI; GREUB 2013; CHARALAMPOUS et al. 2019; GWINN; MACCANNELL; ARMSTRONG, 2019).

Os sequenciamentos podem ser classificados de maneira geral em sequenciamentos *short-read* (de sequencias pequenas) e *long-read* (de sequencias longas). Os sequenciamentos short-*read* podem ser divididos em duas categorias: por ligação e por síntese. Nesse tipo de sequenciamento por ligação uma sequência de sonda que está ligada com um fluoróforo hibridiza com um fragmento de DNA da amostra e é ligado a um oligonucleotídeo adjacente.

O espectro de emissão do fluoróforo indica a identidade da base ou bases complementares a posições específicas dentro da sonda (GOODWIN; MCPHERSON; MCCOMBIE, 2016). Já no sequenciamento por síntese, uma polimerase é usada e um sinal, como um fluoróforo ou uma mudança na concentração iônica, identifica a incorporação de um nucleotídeo em uma fita alongada. Para classificá-las melhor temos o método de terminação reversível cíclica, a exemplo do Illumina, ou como adição de nucleotídeo único, com no método do Ion Torrent (GOODWIN; MCPHERSON; MCCOMBIE, 2016).

Entretanto ao falar sobre sequenciamentos *long-reads* há um diferencial, pois, tais metodologias oferecem abundância de sequências que possuem várias quilobases (kb), permitindo uma análise complexa de vários elementos que as outras tecnologias de *short-read* não são suficientemente capazes de realizar (GOODWIN; MCPHERSON; MCCOMBIE, 2016). No quadro 1 há o resumo das principais informações sobre essas tecnologias NGS para sequenciamento de DNA.

Quadro 1 - Resumo comparativo das tecnologias de sequenciamento

Método	**Tipo de sequenciamento**	**Tipo de detecção**	**Tamanho das leituras**	**Tipo de abordagem**
Sanger	Por síntese	Por terminadores de cadeia irreversíveis	500-900bp	*Short-read*
Solex/Illumina	Por síntese	dDNTPs terminadores de cadeia reversíveis	150-300bp	*Short-read*
MinION	De molécula única em tempo real	Perturbação na corrente elétrica	900kb	*Long-reads*

Fonte: Elaborado pela autora

Já a tecnologia da Solexa/Illumina usa moléculas terminadoras similares aquelas usadas no sequenciamento de Sanger (GUO et al., 2008; JU et al., 2006). Apesar de também ser um sequenciamento por síntese, este possui uma abordagem de terminação reversíveis cíclica.

Nesse método uma fita de DNA se liga a adaptadores presentes na célula de fluxo (*flowcell)*, sendo amplificada em fase sólida por PCR em ponte, criando clusters que cada um deles irá conter milhões de cópias do mesmo fragmento de DNA adiciona-se os 4 tipos de ddNTPs terminadores reversíveis contendo fluoróforos, junto com a enzima DNA polimerase. A incidência de um feixe de raios laser excita os fluoróforos proporcionando emissão de luz que difere em função da base incorporada. A fluorescência emitida após a incorporação de cada nucleotídeo é registrada como imagem e no final, através de uma decodificação destas imagens, tem se a sequência de interesse (GOODWIN; MCPHERSON; MCCOMBIE, 2016).

A principal inovação dessa plataforma com relação as demais consiste na clonagem dos fragmentos em uma plataforma sólida de vidro, que é a PCR de fase sólida (FEDURCO et al., 2006; TURCATTI et al., 2008). No início de 2010, a Illumina lançou os equipamentos HiSeq 2000 e, em 2011, o MiSeq, um sequenciador Illumina de bancada, foi introduzido,

compartilhando a maioria das tecnologias com o HiSeq.

O dispositivo desenvolvido em 2014 pela Oxford Nanopore Tecnologies (ONT), o MinION, é uma tecnologia de sequenciamento por nanoporos que permite a geração de sequências longa de DNA a partir do sequenciamento de uma única molécula de DNA em tempo real (GOODWIN; MCPHERSON; MCCOMBIE, 2016).

O MinION é um dispositivo de baixo custo, simples preparo de biblioteca e portátil (figura 4). Possui entrada USB que permite o funcionamento do equipamento quando conectado a qualquer computador ou laptop (MIKHEYEV; TIN, 2014). Medindo apenas 10 x 3 x 2 cm e pesando apenas 90g, o MinION é considerado o menor dispositivo de sequenciamento disponível atualmente (LU; GIORDANO; NING, 2016).

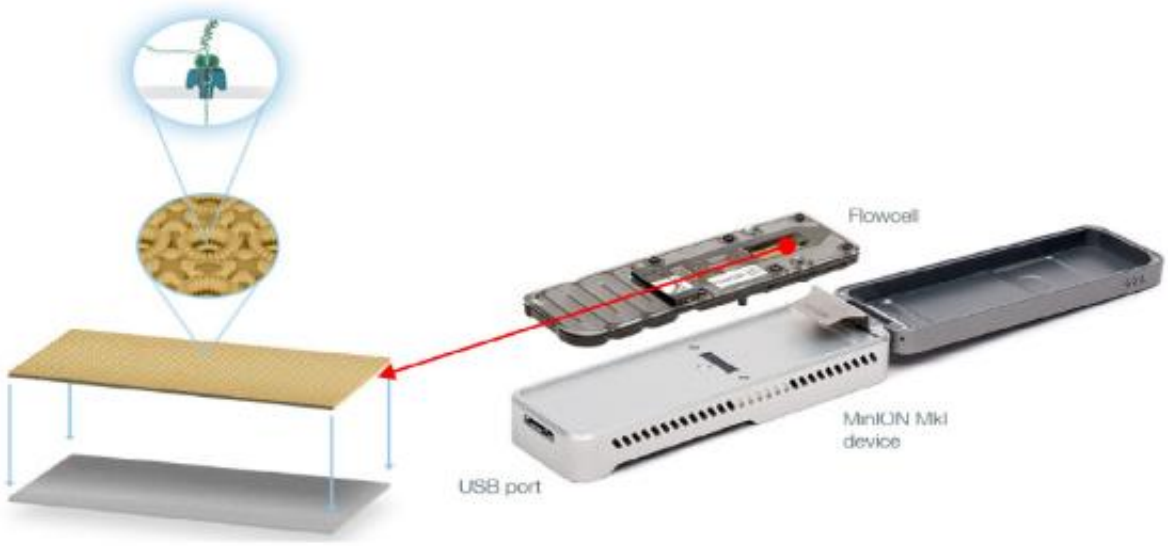


Figura 4 - Dispositivo MinION
Fonte: (LU; GIORDANO; NING, 2016)

O MinION, desde seu lançamento, vem sendo usado para sequenciar diversos microrganismos, como vírus, bactérias e fungos. Este equipamento, em teoria, tem a capacidade de sequenciar fragmentos de DNA de qualquer tamanho, diferentemente de outras plataformas, porém na prática ainda possui certas limitações para fragmentos ultralongos (GOODWIN; WAPPEL; MCCOMBIE, 2017).

A tecnologia disruptiva empregada é a de sequenciamento por nanoporos. Um nanoporo é uma proteína trans-membrana encontrada na natureza que permite a célula trazer moléculas de um lado da célula para outro e funciona como passagem entre dois sistemas. Os buracos podem ser criados por proteínas perfurando membranas (nanoporos biológicos) ou em materiais sólidos (nanoporos de estado sólido) (IP et al., 2015). Na tecnologia empregada pela ONT os

nanoporos utilizados foram desenvolvidos melhorando certas características dessas proteínas que servem como biossensores e são incorporados em uma membrana de polímero eletricamente resistente (WANG et al., 2021) para a utilização dos nanoporos de estado sólido (do inglês, *solid-state nanopore*) presentes na célula de fluxo (*flowcell*) do MinION (GOODWIN; MCPHERSON; MCCOMBIE, 2016). A *flowcell* do MinION possui 512 canais, permitindo que até 512 moléculas de DNA sejam sequenciadas independentemente e simultaneamente (IP et al., 2015). Cada canal está conectado a quatro poços e pode fornecer dados de um dos quatro poços por vez. Atualmente, as *flowcells* do MinION têm 2.048 poços, sendo 4 para cada um dos 512 canais (LU; GIORDANO; NING, 2016).

O nanoporo está localizado dentro da *flowcell* numa membrana de estado sólido, que é um polímero sintético desenvolvido especialmente para permitir a aferição da corrente elétrica ao longo do tempo de funcionamento. Quando aplicado um potencial elétrico ao sistema os íons podem seguir livremente pelo poro e assim é possível medir a corrente pelo fluxo de íons gerada, chamada de corrente aberta (GOODWIN; MCPHERSON; MCCOMBIE, 2016). Cada alteração na voltagem é característica para a molécula que está passando pelo poro (figura 5), portanto os registros dessas alterações na corrente elétrica são identificados por um chip associado (LU; GIORDANO; NING, 2016) e serão analisados por um *software* que faz a identificação da sequência de nucleotídeos do DNA sequenciado (GOODWIN; MCPHERSON; MCCOMBIE, 2016).

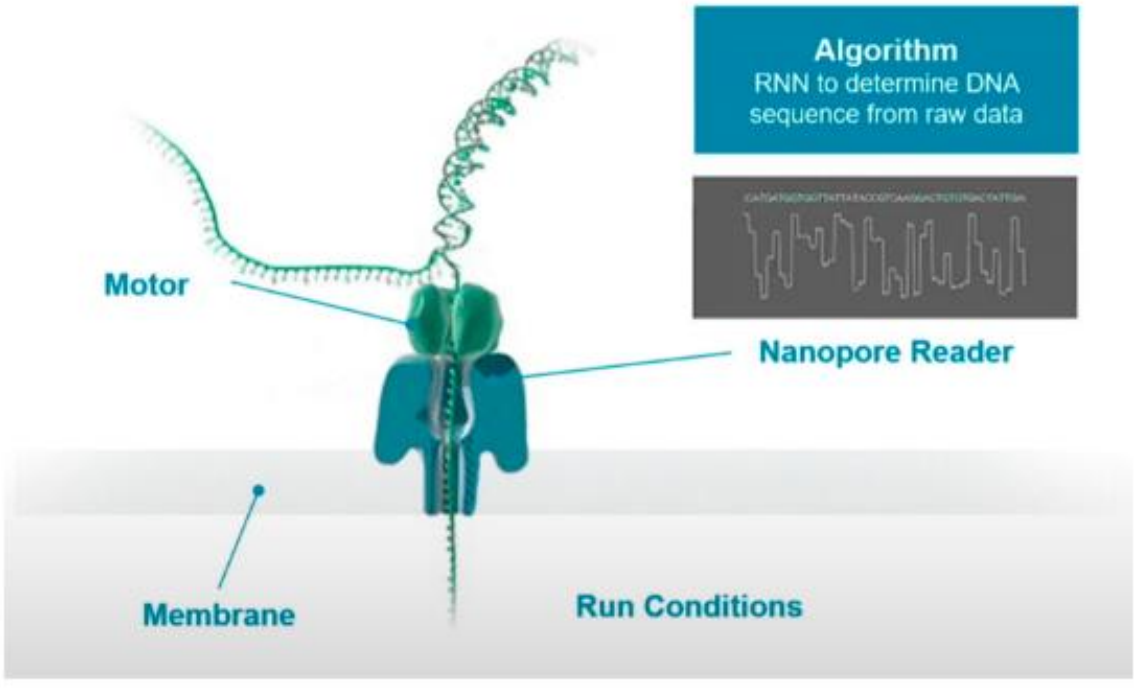


Figura 5 - Funcionamento de um nanoporo.
Fonte: (OXFORD NANOPORE TECHNOLOGIES, 2020)

Uma característica chave do dispositivo MinION é que não há um tempo de execução fixo, portanto, o usuário pode executar o sistema pelo período que o usuário achar necessário, pois os dados são transmitidos em tempo real. É compatível com sistemas operacionais Windows, Mac e Linux e requer, no mínimo, 8 Gb de RAM, um SSD de 512 Gb, um processador quad core i5, portas USB 3 e uma conexão de internet durante a utilização (GOODWIN; WAPPEL; MCCOMBIE, 2017).

O MinION armazena dados biológicos sequenciados na corrida em dois formatos de arquivo, o. fast5 e o *FAST*Q. O *fast*5 é projetado para conter todas as informações necessárias para analisar dados de sequenciamento nanopore e rastreá-los de volta à sua origem. Já o *FAST*Q é um formato de armazenamento de sequência baseado em texto universal, contendo a sequência do ácido nucleico e seus indicadores de qualidade Q (ou qscores). Conforme o experimento avança, por padrão, cerca de 4.000 leituras (do inglês *reads*) são armazenadas em cada arquivo *fast*5 (MinION IT requeriments version 1.0.0).

A ONT constrói e fornece vários tipos de *softwares* envolvidos na aquisição, orquestração e análise: MinKNOW, Guppy, EPI2ME. Este é o *software* principal fornecido pela Oxford Nanopore, sem o qual os dispositivos de sequenciamento não podem ser executados.

O MinKNOW realiza 6 tarefas principais: aquisição de dados, análise e feedback em tempo real, streaming de dados, controle de dispositivo incluindo seleção de parâmetro de execução, identificação e rastreamento de amostras e garantir que a química esteja funcionando corretamente. Além de possuir uma interface gráfica de usuário intuitiva, os *softwares* de análise de dados bioinformática recebem atualizações regularmente da ONT (WANG et al., 2021). Os dados do MinKNOW são compactados em arquivos. fast5 de leitura individual, que são um formato de arquivo personalizado com base no tipo de arquivo. hdf5. A utilização do MinKNOW além de ser obrigatória para o uso do sequenciamento por MinION permite a seleção e o início de experimentos, além de fornecer feedback em tempo real sobre a progressão do experimento (MinION IT requeriments version 1.0.0).

A portabilidade do sistema MinION, dispositivos móveis de extração de DNA e *basecalling* em tempo real com o *software* Guppy, além de outras ferramentas de bioinformática offline, possibilitam a pesquisa de campo até mesmo nos cenários mais desafiadores ou desprovidos de grande aporte tecnológico (RUNTUWENE; TUDA; MONGAN, 2019). O MinION tem sido usado para detecção rápida de patógenos, por exemplo: infecções bacterianas do trato respiratório inferior, (CHARALAMPOUS et al., 2019), endocardite infecciosa (CHENG et al., 2018), pneumonia (GORRIE et al., 2018), meningite bacteriana (MOON et al.,

2019) e infecção em articulações protéticas (SANDERSON et al., 2018).

Apesar de ser um dispositivo ser prático e inovador, ainda há várias limitações para o seu uso, a principal questão é a sua taxa de erro elevada que pode variar de 5% a 30%, além da necessidade de quantidades relativamente altas de material genético (*input*) para que o sequenciamento tenha um bom rendimento (LIN; HUI; MAO, 2021; GOODWIN et al., 2016).

Todavia, mesmo com esta alta taxa de erro, as leituras brutas do MinION ainda sim podem ser usadas para detecção precisa dos vírus presentes nas amostras (BATOVSKA et al., 2017). Superar tais esses desafios exigirão avanços na tecnologia e nos *softwares* de bioinformática para que seja amplamente difundido (KONO; ARAKAWA, 2019; RANG; KLOOSTERMAN; DE RIDDER, 2018).

Com a resolução do problema da alta taxa de erro, otimização do algoritmo do *basecalling* e equiparação dos dispositivos da ONT ao desempenhar no mesmo nível de precisão de outras técnicas de sequenciamento amplamente utilizadas na clínica médica, é possível que o MinION venha a ser usado em larga escala a nível de diagnóstico clínico em breve (GOODWIN; MCPHERSON; MCCOMBIE, 2016). Entretanto é necessário investir em estudos de otimização da tecnologia, ampliação de teste de desempenho diagnóstico para diferentes patógenos e investimento em novos métodos de preparação de bibliotecas de sequenciamento (LIN; HUI; MAO, 2021).

Com a tecnologia do MinION já é possível analisar regiões repetitivas no genoma, simplificar a montagem do genoma de novo e melhorar os genomas de referência existentes, sequenciar micróbios inteiros e em tempo real, explorar modificações epigenéticas usando sequenciamento de DNA direto de longa leitura, estudar composição ambiental, de amostras clínicas ou de microbiomas aumentando a identificação metagenômica de organismos diversos.

2.2 METAGENÔMICA

Metagenômica é o estudo do material genético recuperado diretamente a partir de amostras ambientais (ZEPEDA MENDOZA; SICHERITZ-PONTÉN; GILBERT, 2015). Essa área de estudo visa compreender melhor a composição genética de ambientes e amostras complexas sem a necessidade de possuir um alvo específico para investigação, sem criar clones de interesse e até mesmo sem cultivar previamente em meios de cultura. Assim como as outras ciências “ômicas” a metagenômica compreende em um conjunto de técnicas e métodos relacionados a um campo de pesquisa, neste caso a utilização do sequenciamento de nova geração e a análise de bioinformática dos dados sequenciados se torna necessária para fornecer

uma interpretação biologicamente significativa dos dados obtidos (FORBES et al., 2018).

A primeira utilização do termo metagenômica foi para descrever um recurso de exploração de todos os genes em uma determinada comunidade avaliando assim as funções bioquímicas de clones produzidos, hoje nomeada como metagenômica funcional (RONDON et al., 2000). Palácios et al (2008) implementaram pela primeira vez a metagenômica para fins de identificação de patógenos, com esse estudo foi possível identificara um novo arenavírus em uma série de doenças fatais associadas a transplantes. A aplicação da metagenômica na clínica deriva do uso de microarrays por volta do início dos anos 2000. (STREIT; SCHMITZ, 2004; MILLER; TANG, 2009). Entretanto, com o advento do uso de tecnologias de sequenciamento NGS, essa metodologia foi um dos principais impulsos no campo da metagenômica (VOELKERDING; DAMES; DURTSCHI, 2009).

A detecção não direcionada e imparcial de patógenos pode promover bons resultados em prazos relevantes para o diagnóstico clínico do paciente (BROWN et al., 2015; NACCACHE et al., 2015; PALACIOS et al., 2008; WILSON et al., 2014) e informações chave para a saúde pública (BRIESE et al., 2009; GIRE et al., 2014). É possível também encontrar na literatura que os ensaios de metagenômica clínica têm sido mais frequentemente utilizados para fins diagnósticos em infecções respiratórias (29,2%), neurológicas (26,2%), cardíacas e sanguíneas (13,9%) e gastrointestinais (10,8%) (FORBES et al., 2018).

Atualmente, o uso do sequenciamento metagenômico se torna extremamente útil à clínica quando há falha na identificação de agentes causadores de doenças infeciosas (WILSON et al. 2014). Patógenos difíceis de detectar por metodologias de rotina são comumente observados na clínica (HAN et al., 2019). Diversos autores afirmam que a causa das doenças infecciosas permanece desconhecida em aproximadamente 50% das infecções da corrente sanguínea (FENOLLAR; RAOULT, 2007), 40% das infecções entéricas (THOMAS et al. 2015) e em mais de 50% de infecções do sistema nervoso central (GLASER et al. 2006).

O fluxo do diagnóstico tradicional no caso de doenças infecciosas envolve a formulação de hipóteses de diagnóstico diferencial por um médico que a partir disso solicitará testes que achar adequado a hipótese na tentativa de identificar o patógeno causador do quadro clínico. Essa metodologia é aplicada a mais de um século e vem sendo aperfeiçoada graças ao avanço dos testes microbiológicos ao longo dos anos (FORBES et al., 2018). As metodologias de PCR e ELISA possuem um número limitado dentro de um painel de patógenos e geralmente necessitam que o microrganismo de interesse seja cultivado adequadamente a partir da amostra clínica, porém para especificações de cultivo é necessário ter uma hipótese ou suspeita inicial para direcionar os testes (CHIU; MILLER, 2019).

2.2.1 **Metagenômica e o estudo do viroma**

A PCR é um ensaio molecular convencional usado para detecção de material genético de um patógeno. Entretanto, sozinha, essa técnica não pode fornecer caracterização do viroma (KO et al., 2019). Atualmente o sequenciamento metagenômico por NGS é bastante utilizado nesses casos e possui grande potencial de ser aplicado ao diagnóstico clínico (PARKER; CHEN, 2018; HOULDCROFT; BEALE; BREUER, 2017).

Ainda pouco se sabe sobre papel dos vírus no microbioma humano (MITCHELL; OLIVER; GLANVILLE, 2016). O metagenoma viral (viroma) é uma coleção das sequências de DNA e RNA derivadas de uma comunidade (MITCHELL; GLANVILLE, 2018). O viroma humano pode ser considerado um subconjunto do microbioma humano geral, que pode incluir vírus patogênicos, vírus residentes associados de forma estável a tecidos hospedeiros saudáveis, bacteriófagos e outros elementos (VIRGIN; WHERRY; AHMED, 2009; DUERKOP; HOOPER, 2013).

Para infecções mais comuns na rotina clínica o custo-benefício de um diagnóstico molecular é que ele oferece um resultado rápido e confiável, porém para outros patógenos não esperados nesse diagnóstico de rotina ou até mesmo patógenos novos circulantes, essa identificação se torna mais complexa e demorada. A NGS metagenômica permite a permite que uma gama maior de patógenos sejam identificados via cultivo ou diretamente de amostras clínicas (LEFTEROVA et al., 2015).

De um ponto de vista crítico, a implementação da metagenômica clínica parece ser bastante promissora para diagnóstico de doenças infecciosas (FORBES et al., 2018; RUPPÉ; SCHRENZEL, 2018; MIAO et al. 2018). O principal desafio para esta tecnologia ser utilizada para o sequenciamento do genoma viral é a necessidade de enriquecer pequenas quantidades de partículas virais de grandes hospedeiros, como o ser humano (YANG et al., 2011; HOULDCROFT; BEALE; BREUER, 2017). O alto custo também é uma grande preocupação em relação ao uso da metagenômica no diagnóstico de doenças infecciosas (HAN et al., 2019).

2.2.2 **Bioinformática e sequenciamento NGS**

Para que haja diagnóstico, são necessárias etapas importantes (figura 6) do fluxo de trabalho: seleção e preparação da amostra, escolha do método de sequenciamento, compra de reagentes adequados, padronização e/ou validação de um protocolo de sequenciamento e análise de bioinformática (HAN et al., 2019; SODING, 2005).

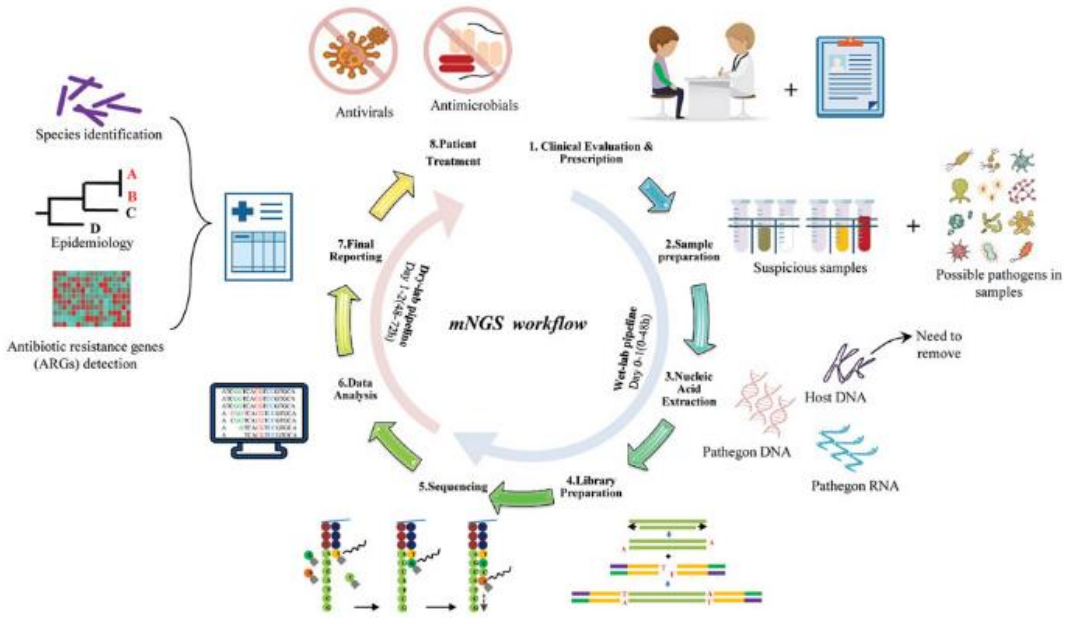


Figura 6 - Resumo do fluxo de trabalho com tecnologia NGS para metagenômica.
Fonte: HAN et al., 2019

A falta de padronização de métodos e validação desses fluxos de trabalho dificulta a capacidade de avaliar a aplicação clínica por diferentes laboratórios (DEURENBERG et al., 2017). Sendo assim, antes ser implementada em um laboratório é necessário uma extensa validação afim de garantir resultados precisos, reprodutíveis e confiáveis (HAN et al., 2019).

Dependendo da metodologia NGS utilizada e da estratégia de sequenciamento, medidas de controle de qualidade específicas podem ser recomendadas pelo fabricante e devem ser seguidas com base no fluxo de trabalho validado (SIMNER; MILLER; CARROLL, 2018).

O maior desafio referente à introdução de NGS no laboratório de microbiologia clínica é a análise dos dados (DEURENBERG et al., 2017). A análise de grandes conjuntos de dados complexos como os de tecnologias NGS metagenômica requer profissionais altamente treinados e cuidado extremo no manuseio das amostras para evitar erros e contaminação cruzada (SALTER et al., 2014; STRONG et al., 2014).

Se torna necessário também a combinação da clínica com a bioinformática com bons recursos computacionais, necessitando assim a colaboração de uma equipe multidisciplinar e multissetorial para o desenvolvimento de possíveis fluxos de trabalho adequados e análises de dados (GARGIS; KALMAN; LUBIN, 2016), a fim de mitigar a interpretação errônea dos resultados a falso positivo e falso negativo (SIMNER; MILLER; CARROLL, 2018). Além disso, são necessários mais estudos para melhorar o fluxo de trabalho visando reduzir o tempo de diagnóstico e de preparação da biblioteca e, consequentemente, popularizar a técnica resultando reduzir ainda mais os custos associados (DEURENBERG et al., 2017).

3 **OBJETIVOS**

3.1 OBJETIVO GERAL

Padronizar e validar um método de sequenciamento de nova geração (NGS) para análise metagenômica viral utilizando o MinION.

3.2 OBJETIVOS ESPECÍFICOS

- Testar amostras clínicas de casos: suspeitos de meningites virais, arboviroses, infecções do trato respiratório, infecções crônicas causadas por hepatites virais e infecção por HIV, além de amostras de isolados virais;
- Otimizar o tempo de processamento dos dados de sequenciamento metagenômico;
- Testar diferentes pipelines visando obter melhores resultados em especificidade, sensibilidade e acurácia no diagnóstico por metagenômica;
- Classificar taxonomicamente sequências de agentes infecciosos que possam ter relevância clínica.

4 **METODOLOGIA**

4.1 DESENHO DO ESTUDO

Trata-se de um estudo de validação de uma metodologia utilizando o sequenciador MinION, que incluiu a avaliação de casos representativos de diferentes doenças infecciosas virais de interesse médico. A análise molecular e de bioinformática foi realizada pela equipe do Laboratório de Patologia e Biologia Molecular (LPBM), responsável pelo projeto, no Instituto Gonçalo Moniz – Fiocruz-BA.

4.2 CASUÍSTICA

4.2.1 **Aspectos éticos**

O projeto está aprovado no Comitê de Ética em Pesquisa (CEP) da Fiocruz-BA sob CAAE n° 23864719.0.0000.0040. Este estudo deriva do projeto de pesquisa intitulado: Detecção Avançada de Patógenos Emergentes de Impacto Global em um Foco Urbano; sob a coordenação do Prof. Dr. Mitermayer Galvão dos Reis e do Prof. Dr. Federico Costa.

4.2.2 **Seleção de amostras**

Amostras clínicas retrospectivas e recentes de casos de arboviroses, casos suspeitos de meningites virais, infecções do trato respiratório, infecções crônicas causadas por hepatites virais e casos de infecção por HIV, além de amostras de isolados virais, foram providas de forma anônima e voluntária por diferentes grupos de pesquisa da Fiocruz-BA e pelo Laboratório Central (LACEN-BA), para servirem como controle na padronização e validação dos métodos, já que seus resultados já eram conhecidos e confirmados.

4.2.3 **Definição de caso**

Diagnóstico presumido: Todas as amostras utilizadas possuíam um diagnóstico viral prévio confirmado por algum método de biologia molecular, ou seja, com ácido nucléico detectável.

Diagnóstico obtido: O diagnóstico obtido baseou-se na análise de abundância relativa

em nível de *reads* ou *contigs* classificadas e do *cut-off* (0,0% a 5,0%).

a) Vírus detectado, para fins diagnósticos foram analisamos apenas os vírus do painel reduzido (EV, DENV, ZIKV, CHIKV, HCV, HIV e SARS-CoV2) Outros vírus podem ter sido identificados, mas não foram considerados para análise de acurácia neste trabalho (outros vírus);
b) Indetectável: O número de *reads* ou *contigs* do viroma foi abaixo do *cut-off* ou o viroma foi “zero”;
c) Indeterminado: Nenhum dos vírus do painel apresentou um número de *reads* ou *contigs* acima do *cut-off*.

Diagnóstico acurado: Diagnóstico obtido em nível de *reads* ou *contigs* sensível e específico.

a) Sensível: quando o diagnóstico obtido em nível de *reads* ou *contigs* incluiu o vírus descrito no diagnóstico presumido;
b) Específico: quando o diagnóstico obtido em nível de *reads* ou *contigs* incluiu exclusivamente o vírus do diagnóstico presumido.

Amostras com resultado indetectável ou indeterminado foram considerados falha de sensibilidade e especificidade. As bibliotecas consideradas falhas apresentaram um dos três critérios: falha ocorrida nas primeiras bibliotecas devido ao processo de padronização do método e na análise de amostras retrospectivas que se mostraram degradadas; bibliotecas onde não foram capazes de realizar o processo de demultiplexação; e bibliotecas que não possuem resultados em nenhuma ou na maioria das amostras. Todas as bibliotecas consideradas falhas foram excluídas das análises desse estudo.

4.3 DIAGNÓSTICO POR METAGENÔMICA

4.3.1 **Extração do RNA viral**

O RNA viral foi extraído de 140µL das amostras de soro usando o QIAmp Viral RNA Mini Kit da Quiagen de acordo com as instruções do fabricante e quantificado utilizando o kit Qubit RNA High Sensivity no Qubit 3.0 fluorimeter (ThermoFisher).

4.3.2 **Preparação da biblioteca de sequenciamento**

A preparação da biblioteca de sequenciamento foi realizada em cinco etapas:

tratamento com DNAse, síntese de cDNA, amplificação independente de sequência com iniciador único (do inglês, *Sequence-Independent, Single-Primer Amplification* ou SISPA), ligação dos adaptadores e *barcodes*, verifcação da *flowcell* e aplicação da biblioteca no MinION para sequenciamento. Etapas de purificação foram realizadas, segundo o nosso protocolo, utilizando *beads* magnéticas AMPure XP beads (Beckman Coulter).

4.3.2.1 Tratamento da amostra com DNAse

O RNA extraído foi tratado com DNase utilizando o 10x Qiagen DNase Buffer, Qiagen RNase-free DNase I, SUPERase-In RNase Inhibitor e Linear acrylamide. A reação foi incubada por 30 minutos a 37°C, depois do tempo determinado foi finalizada com EDTA a 0.5M e a primeira *cleanup* com *beads* magnéticas foi realizada numa proporção de 0.8:1 de *beads* para a amostra.

4.3.2.2 Síntese de cDNA e SISPA

A transcrição reversa e síntese de cDNA foi realizada utilizando-se o kit da SuperScript IV First-Strand Synthesis System (ThermoFisher) com o primer A (5′-GTTTCCCACTGGAGGATA-NNN-NNN-NNN-3′). Na SISPA (REYES; KIM, 1991), são utilizados Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs - NEB) e Sequenase Version 2.0 DNA Polymerase (ThermoFisher) e o primer B (5′-GTTTCCCACTGGAGGATA-3′). O cDNA foi purificado pelo método de separação com esperas (do inglês, *beads*) magnéticas na proporção: 0.8:1 beads:amostra. As concentrações do DNA amplificado e purificado foram medidas usando o kit Qubit dsDNA High Sensivity no Qubit 3.0 fluorimeter (ThermoFisher).

4.3.2.3 Ligação dos adaptadores e dos *barcodes*

O DNA amplificado seguiu o 1D Native Barcoding Protocol (Comunidade ONT online, disponível em http://community.nanoporetech.com). Resumidamente, adaptadores foram ligados aos fragmentos de DNA para permitir o sequenciamento direcional único (1D). Após purificado, o cDNA seguiu para a etapa de reparação das extremidades usando o kit NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs – NEB), e ligação dos *barcodes* com o kit NEB Blunt/TA Ligase Mix (New England Biolabs – NEB) utilizando o kit Native

Barcoding (ONT) EXP NBD 104/EXP NBD 114 (ONT) e ligação dos adaptadores o kit NEBNext Quick Ligation Module (New England Biolabs – NEB) e o Ligation Sequencing Kit (ONT). As amostras foram multiplexadas e foi aplicado na *flowcell* a concentração final de 100ng de biblioteca, a fim de aumentar o rendimento da corrida.

4.3.2.4 Verificação das *flowcells* (MinION, ONT)

Antes da utilização das *flowcells* para sequenciamento, foram realizados testes para avaliar a quantidade de poros ativos disponíveis para o sequenciamento. Testes prévios foram feitos em diferentes períodos do ano e novamente no dia que a biblioteca estava pronta para sequenciamento. Em nossos ensaios foram utilizadas ao todo 22 *flowcells*, os dados de checagem das *flowcells* foram sumarizados no APÊNDICE A.

4.3.2.5 Aplicação da biblioteca no MinION e sequenciamento

Cerca de 70uL de biblioteca final diluída (amostras multiplexadas) foi aplicada na *flowcell* com expectativa de corrida por no mínimo de 24h e máximo de 72h. A aquisição primária dos dados foi realizada pelo *software* MinKNOW (ONT) e os resultados brutos da corrida de cada biblioteca foram armazenados em um SSD no formato. fast5 para análise posterior. As corridas com baixa número de leituras (*reads)* foram interrompidas precocemente.

4.4 ANÁLISE DOS DADOS

4.4.1 **Hardware, sistema operacional e *softwares***

A análise dos dados do sequenciamento foi realizada em um computador dedicado ao projeto com as seguintes especificações de hardware e sistema operacional (quadro 2).

Quadro 2 - Dados sobre o hardware e sistema operacional.

Hardware	*Software*
CPU: 12-Core Intel i7 (24 th*reads*) RAM: 16Gb RAM GPU: NVIDIA GeForce GTX 1650 com 1024 CUDA Cores VRAM: 4Gb	Sistema Operacional: Linux Ubuntu 20.04 LTS 64 bits Nvidia drivers: 465.19.01 CUDA Version: 11.3

Fonte: Elaborado pela autora

Além do sistema operacional Linux, vários *softwares* de bioinformática de código aberto foram instalados para execução de um *workflow* de bioinformática para análise metagenômica de dados de sequenciamento do MinION.

4.4.2 ***Workflows* de bioinformática**

Resumidamente, o *workflow* de bioinformática proposto foi constituído das seguintes etapas: reconhecimento de bases (do inglês, *basecalling*) e remoção dos adaptadores, demultiplexação e remoção dos *barcodes*, remoção dos *primers*, análise e filtros de qualidade e complexidade, filtro das sequências humanas e classificação taxonômica das *reads* (análise em nível de *reads*). As *reads* foram montadas com genomas de referência para classificação taxonômica das *contigs* (análise em nível de *contigs*) e montagem de genomas. Diferentes *softwares* e parâmetros foram testados para cada etapa em quatro *workflows* de bioinformática alternativos (quadro 3). Tivemos como base para a criação das etapas dos nossos *workflows* (https://github.com/lkalabric/ngs-scripts) o *workflow* do nosso colaborador internacional descrito no APÊNDICE B.

O *workflow* 1 apresenta uma análise completa dos dados e utiliza como classificador taxonômico o *software* Kraken2. O *workflow* 2 utiliza as mesmas estratégias, porém suprime completamente o filtro do genoma humano e o polimento e utiliza apenas um banco de dados local para classificação taxonômicas das *reads* e *contigs* por meio do BLAST. Já o *workflow* 3 simplifica o *workflow* 1, uma vez que o filtro do genoma humano e a etapa de "polimento" não são realizados e utiliza como classificador taxonômico a ferramenta online *Genome Detective* (VILSKER et al., 2019), que é capaz realizar estas tarefas, montar *contigs* e identificar os vírus a partir do banco de dados do *software*. O *workflow* 4, por sua vez, é simplificado, assim como o *workflow* 3, utiliza o *software* Epi2ME e possui uma interface mais "amigável" para o usuário.

Quadro 3 – Lista de *softwares* utilizados nos *workflows* de bioinformática.

Etapa	*Workflow* 1 (Kraken2)	*Workflow* 2 (BLAST)	*Workflow* 3 (Genome Detective)	*Workflow* 4 (Epi2ME)
Basecalling e remoção do adaptador	Guppy *Basecaller* v. 5.0.11+2b6dbff (ONT)			

<table>
<tr><td>Demultiplexação</td><td colspan="4">Guppy Barcoder
v5.0.11+2b6dbff (ONT)</td></tr>
<tr><td>Remoção do primer</td><td>Guppy Barcoder
v5.0.11+2b6dbff
(ONT)</td><td>cutdapt v.2.8
(MARTIN,
2011)</td><td colspan="2">Guppy Barcoder
v5.0.11+2b6dbff (ONT)</td></tr>
<tr><td>Análise de qualidade</td><td colspan="2">pycoQC v.2.5.2 (LEGER;
LEONARDI, 2019)</td><td rowspan="6">Genome
Detective
(VILSKER et
al., 2019)</td><td rowspan="5">Epi2ME</td></tr>
<tr><td>Filtro de qualidade e complexidade</td><td colspan="2">NanoFilt 2.8.0 & PRINSEQ 0.20.4 |
0.13</td></tr>
<tr><td>Remoção de sequencias humanas</td><td>GRCh38
minimap2-2.22 (LI,
2018), GMAP (WU;
WATANAE, 2005)
& samtools v.1.7</td><td rowspan="3">BLAST
2.12.0+
(ALTSCHUL
et al., 1990)</td></tr>
<tr><td>Processamento e correção erros de reads</td><td>minimap2-2.22 (LI,
2018), racon-1.4.20
(VASER et al., 2017)
& nanopolish
(SIMPSON et al.,
2017)</td></tr>
<tr><td>Classificação taxonômica</td><td>Kraken2 (WOOD;
LU; LANGMEAD,
2019)</td></tr>
<tr><td>Montagem de contigs</td><td>minimap2-2.22 (LI,
2018) & samtools
v.1.7</td><td>minimap2-
2.22 (LI,
2018) &
samtools v.1.7</td><td>-</td></tr>
</table>

Fonte: Elaborado pela autora

4.4.3 **Dados para treinamento e sequências referência**

O conjunto de dados de treinamento (*training dataset*) DENV_FTA_1 foi utilizado para testar os *softwares* dos *workflows* de bioinformática. Este *dataset* contém aproximadamente 3 milhões de leituras e consiste no diagnóstico de 12 amostras de plasma eluidos de cartões FTA e amplificados por RT-PCR para diagnóstico de dengue. Um *dataset* randômico de

aproximadamente 3 milhões de sequências geradas artificialmente, aleatoriamente e sem significado clínico também foi utilizado como controle. Para realizar a classificação taxonômica proposta no *workflow* 2, um banco de dados local contendo sequencias referências de vírus de interesse médico para esse trabalho, foi montado a partir dos números de acesso do *GenBank* listadas no ANEXO B.

4.4.4 *Benchmark* Guppy *Basecalling* e Guppy *Barcode*r

A etapa de *basecalling* decodifica o sinal da sequência de nucleotídeos (WANG et al., 2021) e foi realizada utilizando-se o *software* Guppy *Basecaller* v. 5.0.11+2b6dbff (ONT). Esta etapa corresponde ao maior custo computacional da análise e pode ser realizada conforme três modelos: *fast*, *hac* e *sup*. Para otimizar o tempo de processamento (horas), o *software* foi executado na GPU da placa de vídeo que é compatível com a "*Compute Unified Device Architecure*" (CUDA GPU). Além disso, uma série de testes virtuais analíticos (*benchmark)* usando o *dataset* DENV_FTA_1 permitiu testar diferentes combinações dos parâmetros no modelo *fast* e *hac*. O *basecalling* pelo modelo *fast* com os parâmetros otimizados pelo nosso *benchmark* foi adotado por possuir uma melhor relação entre o tempo de processamento e a acurácia no hardware disponível para análise dos dados de sequenciamento.

O *basecalling* foi executado utilizando o valor mínimo de qualidade (qcore) igual a 9 (qc9), mas também realizamos alguns testes com o qc7. A tabela 1 descreve os demais parâmetros padrões dos arquivos de configuração dna_r9.4.1_450bps para análise pelos modelos *fast*, *hac* e *sup*, respectivamente (ANEXO A). Dois *benchmarks* foram realizados variando estes parâmetros para os modelos *fast* e *hac*. Devido o hardware disponível, o modelo *sup* não foi testado.

Tabela 1 - Parâmetros padrão de GPU padrão descritos para cada tipo de análise.

Modelo	Parâmetros de configuração dna_r9.4.1_450bps				Acurácia
	gpu_runners_per_device	chunk_size	chunks_per_runner	num_callers	presumida (%)
fast	8	2000	160	4	~95
hac	4	2000	256	4	~97
sup	12	1000	256	4	~98

Fonte: Elaborado pela autora

Logo após o *basecalling* foi executada a etapa de demultiplexação, realizada pelo Guppy

*Barcode*r v5.0.11+2b6dbff (ONT). As *reads* de cada amostra foram separadas em pastas diferentes através da identificação dos códigos de barras (*barcodes*) utilizados na preparação da biblioteca. O parâmetro –trim_*barcodes* "apara" as sequencias do *barcode* enquanto o parâmetro --required_*barcodes*_both_ends filtrar as *reads* que ligaram *barcodes* no início e no final das sequencias (BC). Outros parâmetros foram testados no *benchmark* do *Barcode*r.

4.4.5 **Teste do limiar de detecção do método**

Realizamos um ensaio de diluição seriada utilizando a vacina oral contra a poliomielite produzida por Bio-Manguinhos. No concentrado viral são utilizadas as cepas de *Poliovírus* atenuados do tipo Sabin 1 e 3, propagadas em cultivo de célula diploides humanas, segundo as normas da Organização Mundial de Saúde (BIO-MANGUINHOS, 2022). Do mesmo modo que as amostras, o RNA total foi extraído a partir de 140µL da vacina (diluição 10^0 ou E0) e realizamos 6 diluições de base 10 (diluições 10^{-1} a 10^{-6} ou E-1 a E-6).

Este teste teve como objetivo avaliar o limiar de detecção viral do MinION utilizando apenas uma amostra em diferentes concentrações, permitindo inferir sobre possíveis limitações do protocolo ou do método analítico, além de simular diferentes concentrações virais que podem estar presentes nos indivíduos na realidade clínica.

5 **RESULTADOS**

5.1 *BENCHMARKS*

Os nossos primeiros resultados referem-se aos *benchmarks* para otimização da etapa de *basecalling* realizado pelo *software* Guppy *Basecaller* v. 5.0.11+2b6dbff (ONT) executado em GPU CUDA. Os *benchmarks* foram realizados no *dateset* DENV_FTA_1 em triplicata, variando-se os parâmetros de configuração: gpu_runners_per_device, chunk_size, chunks_per_runner e num_callers. O uso da memória VRAM (Mb) foi monitorado pelo comando Linux nvidia-smi. Nos casos com erro "*Out of memory*" a análise foi considerada falha. Os valores de gpu_runners_per_device = 4, chunk_size = 2.000, chunks_per_runner = 50 e num_callers = 4, produziram menor tempo de processamento médio (1,36 horas) pelo modelo de *basecalling fast* (tabela 2).

Tabela 2 - Benchmark de parâmetros do Guppy *Basecaller* no modelo *fast*.

BM	gpu_runners_per_device	chunk_size	chunks_per_runner	num_callers	Tempo médio (horas)	Uso médio VRAM (Mb)
1	1	2000	160	4	1,55	1099,7
2	4	2000	160	4	1,48	2666,3
3	8	2000	160	4	1,52	3685,3
4	12	2000	160	4	1,53	3260,7
5	24	2000	160	4	1,52	3331,7
6	4	500	160	4	1,50	1038,7
7	4	1000	160	4	1,42	1558,0
8	4	1000	50	4	1,37	829,0
9	4	1000	256	4	1,43	2209,7
10	4	1000	512	4	1,55	3378,3
11	4	1000	1024	4	1,65	2805,7
12	4	1000	50	12	1,37	844,0
13	4	1000	50	24	1,37	835,3

Nota: As células marcadas em azul claro representam os valores testados de cada parâmetro de configuração. Em negrito, os valores fixados que produziram os menores tempos de processamento médios. Em vermelho, os parâmetros *default*. A célula marcada em roxo representa a menor média de tempo de análise.
Fonte: Elaborado pela autora

Os valores de gpu_runners_per_device = 12, chunk_size = 2.000, chunks_per_runner = 256 e num_callers = 12, produziram menor tempo de processamento médio (30,72 horas) pelo modelo de *basecalling hac* (tabela 3).

Tabela 3 - *Benchmark* de parâmetros do Guppy *Basecaller* no modelo *hac*.

BM	gpu_runners_per_device	chunk_size	chunks_per_runner	num_callers	Tempo médio (horas)	Uso médio VRAM (Mb)
1	1	2000	256	4	30,93	3603,3
2	4	2000	256	4	162,96	3817,3
3	8	2000	256	4	31,47	3355,0
4	12	2000	256	4	30,92	3749,0
5	24	2000	256	4	32,24	2453,3
6	12	500	256	4	32,99	3704,7
7	12	1000	256	4	31,55	3832,0
8	12	2000	50	4	37,39	3834,3
9	12	2000	160	4	35,17	3662,7
10	12	2000	512	4	30,91	3766,0
11	12	2000	1024	4	-	*memory error*
12	12	2000	256	12	30,72	3738,7
13	12	2000	256	24	32,34	3719,0

Nota: As células marcadas em azul claro representam os valores testados de cada parâmetro de configuração. Em negrito, os valores fixados que produziram os menores tempos de processamento médios. Em vermelho, os parâmetros *default*. A célula marcada em roxo representa a menor média de tempo de análise.
Fonte: Elaborado pela autora

Nossos resultados apontam para uma redução de 10% do tempo de análise utilizando a melhor combinação dos parâmetros no modelo *fast* em relação ao tempo gasto com os parâmetros padrão. Já no modelo *hac* a redução foi ainda mais significativa, foi possível reduzir 81% o tempo de processamento quando comparado ao padrão, representando uma mudança impactante de seis dias de análise para aproximadamente um dia e meio. Entretanto, o modelo *fast* foi adotado por possuir uma melhor relação entre o tempo de processamento no hardware disponível e a acurácia das sequências quando comparado com o modelo *hac*.

A fim de otimizar ainda mais o processamento dos resultados, fizemos um *benchmark* da etapa de demultiplexação realizada pelo *software* Guppy *Barcoder* v5.0.11+2b6dbff (ONT) executado em GPU CUDA. O *benchmark* também foi realizado no *dateset* DENV_FTA_1

testando diferentes combinações dos parâmetros disponíveis pelo Guppy *Barcoder*. Avaliou-se o total de *reads*, o número de *reads* não classificadas pelo *Barcoder* (*Unclassified*), o número de *reads* identificadas com BC e a porcentagem das *reads* com BC (BC%), que seguem para as próximas etapas subsequentes da análise (tabela 4).

Tabela 4 - *Benchmark* de parâmetros do Guppy *Barcode*r utilizando o modelo *fast*.

BM	Parâmetros do *guppy_barcoder*	Total	Unclassified	BC	BC%
1	--trim_*barcodes*	2.946.832	181.022	2.765.810	94%
2	--detect_mid_strand_adapter --trim_*barcodes*	2.946.832	181.301	2.765.531	94%
3	--require_*barcodes*_both_ends --trim_*barcodes*	2.946.832	2.224.273	722.559	25%
4	**--require_*barcodes*_both_ends --detect_mid_strand_*barcodes* --trim_*barcodes***	**2.946.832**	**2.226.255**	**720.577**	**24%**
5	--require_*barcodes*_both_ends --detect_mid_strand_*barcodes* --detect_mid_strand_adapter --trim_*barcodes*	2.946.832	2.226.255	720.577	24%
6	--*barcode*_kits EXP-NBD104 --require_*barcodes*_both_ends --detect_mid_strand_*barcodes* --trim_*barcodes*	2.946.832	2.226.255	720.577	24%

Nota: As células marcadas em verde claro apresentam parâmetros mais liberais que resultaram menor perda percentual do número de *reads*. Em negrito, estão os parâmetros mais restritivos, que são recomendados pela ONT. BM = *benchmark*. BC = *barcode.*
Fonte: Elaborado pela autora

Em termos de tempo de processamento, não observamos diferenças significativas. Entretanto, optamos por utilizar os parâmetros do BM4 do *barcoder* (ou bc4), que são recomendados pela ONT, e que primam por um maior rigor na identificação das *reads* pelo próprio processo de demultiplex.

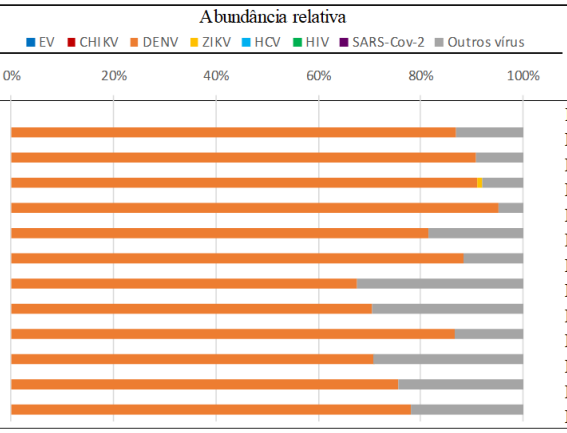
5.2 DIAGNÓSTICO POR CLASSIFICAÇÃO TAXONÔMICA A NÍVEL DE *READS*

Todos os *workflows* utilizaram os parâmetros de *basecalling* e *barcoder* acima. Entretanto, na etapa de classificação taxonômica testamos diferentes *cut-offs* para eliminar *reads* pouco abundantes que poderiam representar "ruídos" ou artefatos do sequenciamento. O Gráfico 1 apresenta os diagnósticos obtidos dos dados de treinamento com os *cut-offs* maiores ou iguais a 0%, 0,5%, 1% e 5%, respectivamente de A à D (gráfico 1). Estes resultados demonstram graficamente que, para os dados de treinamento, quanto maior e mais restritivo o *cut-off* menos "ruido". Todavia, as amostras que apresentaram menor rendimento no sequenciamento passaram ao diagnóstico de indetectável. Por estes resultados, o *cut-off* ≥ 0,5% foram capazes de eliminar "ruídos" de sequenciamento, diagnosticando corretamente todas as amostras do *dataset* DENV_FTA_1 pelo classificador taxonômico Kraken2.

Gráfico 1 - Classificação taxonômica das *reads* pelo Kraken2 – dados de treinamento

A

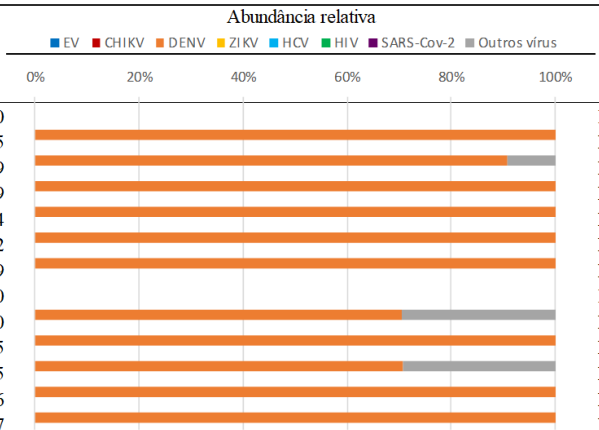
Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Diag. Obtido
Random	-	Randômica	-	0	Indetectável
DS_P1	1	Produto PCR	DENV	2319	DENV
DS_P2	2	Produto PCR	DENV	5509	DENV
DS_P3	3	Produto PCR	DENV	1370	DENV+ZIKV
DS_P4	4	Produto PCR	DENV	1886	DENV
DS_P5	5	Produto PCR	DENV	629	DENV
DS_P6	6	Produto PCR	DENV	2366	DENV
DS_P7	7	Produto PCR	DENV	412	DENV
DS_P8	8	Produto PCR	DENV	3260	DENV
DS_P9	9	Produto PCR	DENV	1159	DENV
DS_P10	10	Produto PCR	DENV	1788	DENV+ZIKV
DS_P11	11	Produto PCR	DENV	787	DENV
DS_P12	12	Produto PCR	DENV	137	DENV



* Classified as virus (virome). Cut-off = 0,0%.

B

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Diag. Obtido
Random	-	Randômica	-	0	Indetectável
DS_P1	1	Produto PCR	DENV	2015	DENV
DS_P2	2	Produto PCR	DENV	5509	DENV
DS_P3	3	Produto PCR	DENV	1249	DENV
DS_P4	4	Produto PCR	DENV	1794	DENV
DS_P5	5	Produto PCR	DENV	512	DENV
DS_P6	6	Produto PCR	DENV	2089	DENV
DS_P7	7	Produto PCR	DENV	0	Indeterminado
DS_P8	8	Produto PCR	DENV	3260	DENV
DS_P9	9	Produto PCR	DENV	1005	DENV
DS_P10	10	Produto PCR	DENV	1785	DENV
DS_P11	11	Produto PCR	DENV	596	DENV
DS_P12	12	Produto PCR	DENV	107	DENV



* Classified as virus (virome). Cut-off = 0,5%.

C

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	2015		DENV
DS_P2	2	Produto PCR	DENV	5006		DENV
DS_P3	3	Produto PCR	DENV	1249		DENV
DS_P4	4	Produto PCR	DENV	1794		DENV
DS_P5	5	Produto PCR	DENV	512		DENV
DS_P6	6	Produto PCR	DENV	2089		DENV
DS_P7	7	Produto PCR	DENV	0		Indetectável
DS_P8	8	Produto PCR	DENV	3260		DENV
DS_P9	9	Produto PCR	DENV	1005		DENV
DS_P10	10	Produto PCR	DENV	1785		DENV
DS_P11	11	Produto PCR	DENV	596		DENV
DS_P12	12	Produto PCR	DENV	0		Indeterminado

* Classified as virus (virome). Cut-off = 1,0%.

D

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	0		Indetectável
DS_P2	2	Produto PCR	DENV	5006		DENV
DS_P3	3	Produto PCR	DENV	0		Indetectável
DS_P4	4	Produto PCR	DENV	0		Indetectável
DS_P5	5	Produto PCR	DENV	0		Indetectável
DS_P6	6	Produto PCR	DENV	0		Indetectável
DS_P7	7	Produto PCR	DENV	0		Indetectável
DS_P8	8	Produto PCR	DENV	0		Indetectável
DS_P9	9	Produto PCR	DENV	0		Indetectável
DS_P10	10	Produto PCR	DENV	0		Indetectável
DS_P11	11	Produto PCR	DENV	0		Indetectável
DS_P12	12	Produto PCR	DENV	0		Indetectável

* Classified as virus (virome). Cut-off = 5,0%.

Fonte: Elaborado pela autora

Os diferentes *cut-offs* analisados com suas respectivas sensibilidades, especificidades e acurácias são descritos na tabela 5 para o classificador taxonômico Kraken2 na análise das 12 amostras do *dataset* DENV_FTA_1. É importante destacar que os dados randômicos não identificaram nenhum vírus em nenhum *cut-off* configurado.

Tabela 5 - Cálculos de sensibilidade, especificidade e acurácia em diferentes *cut-offs* para os *datasets* utilizando o classificador taxonômico Kraken2.

Kraken2 - controles

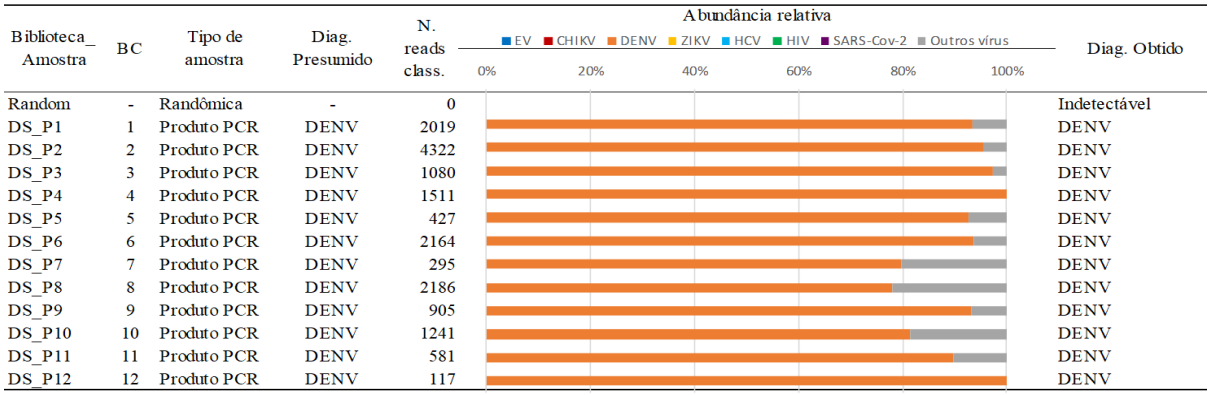
Sumário	Cut-off = 0.0%		**Cut-off ≥ 0,5%**		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	12	100%	**12**	**100%**	12	100%	12	100%
Sensibilidade	12	100%	**11**	**92%**	10	83%	1	8%
Especificidade	10	83%	**11**	**92%**	10	83%	1	8%
Acurácia	22	92%	**22**	**92%**	20	83%	2	8%

Fonte: Elaborado pela autora

No gráfico 2, são representados os diagnósticos obtidos dos dados de treinamento com os *cut-offs* 0%, 0,5%, 1% e 5%, respectivamente de A à D. Estes resultados demonstram graficamente que, para os dados de treinamento, quando analisados pelo BLAST, o *cut-off* de 0% conseguiu diagnosticar corretamente as 12 amostras do *dataset* DENV_FTA_1, sem "ruídos" de sequenciamento. Com valores mais altos de *cut-off* há um aumento de diagnósticos indetectáveis.

Gráfico 2 - Classificação taxonômica das *reads* pelo BLAST – dados de treinamento

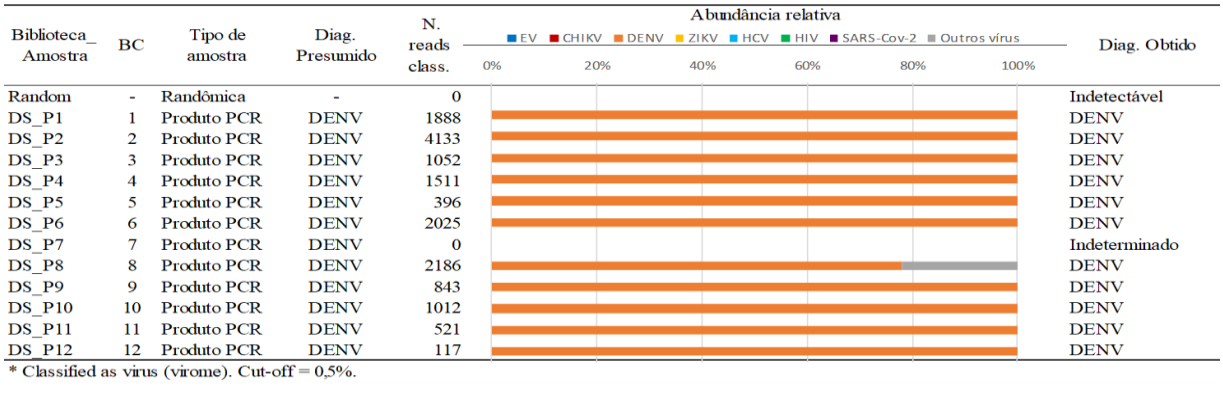
A



Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Diag. Obtido
Random	-	Randômica	-	0	Indetectável
DS_P1	1	Produto PCR	DENV	2019	DENV
DS_P2	2	Produto PCR	DENV	4322	DENV
DS_P3	3	Produto PCR	DENV	1080	DENV
DS_P4	4	Produto PCR	DENV	1511	DENV
DS_P5	5	Produto PCR	DENV	427	DENV
DS_P6	6	Produto PCR	DENV	2164	DENV
DS_P7	7	Produto PCR	DENV	295	DENV
DS_P8	8	Produto PCR	DENV	2186	DENV
DS_P9	9	Produto PCR	DENV	905	DENV
DS_P10	10	Produto PCR	DENV	1241	DENV
DS_P11	11	Produto PCR	DENV	581	DENV
DS_P12	12	Produto PCR	DENV	117	DENV

* Classified as virus (virome). Cut-off = 0,0%.

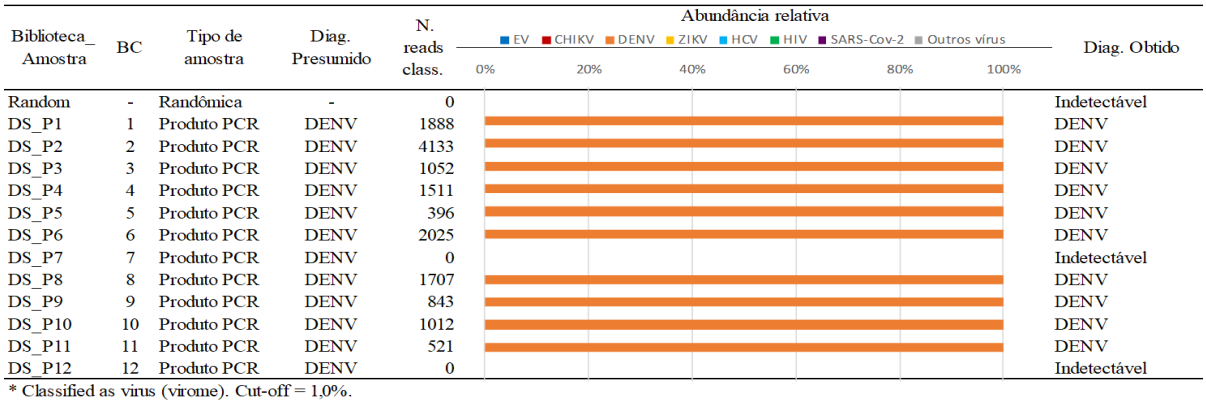
B



Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Diag. Obtido
Random	-	Randômica	-	0	Indetectável
DS_P1	1	Produto PCR	DENV	1888	DENV
DS_P2	2	Produto PCR	DENV	4133	DENV
DS_P3	3	Produto PCR	DENV	1052	DENV
DS_P4	4	Produto PCR	DENV	1511	DENV
DS_P5	5	Produto PCR	DENV	396	DENV
DS_P6	6	Produto PCR	DENV	2025	DENV
DS_P7	7	Produto PCR	DENV	0	Indeterminado
DS_P8	8	Produto PCR	DENV	2186	DENV
DS_P9	9	Produto PCR	DENV	843	DENV
DS_P10	10	Produto PCR	DENV	1012	DENV
DS_P11	11	Produto PCR	DENV	521	DENV
DS_P12	12	Produto PCR	DENV	117	DENV

* Classified as virus (virome). Cut-off = 0,5%.

C



Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Diag. Obtido
Random	-	Randômica	-	0	Indetectável
DS_P1	1	Produto PCR	DENV	1888	DENV
DS_P2	2	Produto PCR	DENV	4133	DENV
DS_P3	3	Produto PCR	DENV	1052	DENV
DS_P4	4	Produto PCR	DENV	1511	DENV
DS_P5	5	Produto PCR	DENV	396	DENV
DS_P6	6	Produto PCR	DENV	2025	DENV
DS_P7	7	Produto PCR	DENV	0	Indetectável
DS_P8	8	Produto PCR	DENV	1707	DENV
DS_P9	9	Produto PCR	DENV	843	DENV
DS_P10	10	Produto PCR	DENV	1012	DENV
DS_P11	11	Produto PCR	DENV	521	DENV
DS_P12	12	Produto PCR	DENV	0	Indetectável

* Classified as virus (virome). Cut-off = 1,0%.

D

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	0		Indeterminado
DS_P2	2	Produto PCR	DENV	4133		DENV
DS_P3	3	Produto PCR	DENV	0		Indetectável
DS_P4	4	Produto PCR	DENV	0		Indetectável
DS_P5	5	Produto PCR	DENV	0		Indetectável
DS_P6	6	Produto PCR	DENV	0		Indetectável
DS_P7	7	Produto PCR	DENV	0		Indetectável
DS_P8	8	Produto PCR	DENV	0		Indetectável
DS_P9	9	Produto PCR	DENV	0		Indetectável
DS_P10	10	Produto PCR	DENV	0		Indetectável
DS_P11	11	Produto PCR	DENV	0		Indetectável
DS_P12	12	Produto PCR	DENV	0		Indetectável

* Classified as virus (virome). Cut-off = 5,0%.

Fonte: Elaborado pela autora

Os diferentes *cut-offs* analisados com suas respectivas sensibilidades, especificidades e acurácias são descritos na tabela 6 para o classificador taxonômico BLAST na análise das 12 amostras do *dataset* DENV_FTA_1.

Tabela 6 - Cálculos de sensibilidade, especificidade e acurácia em diferentes *cut-offs* para os *datasets* utilizando o classificador taxonômico BLAST.

BLAST - controles

Sumário	**Cut-off = 0.0%**		Cut-off ≥ 0.5%		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	**12**	**100%**	12	100%	12	100%	12	100%
Sensibilidade	**12**	**100%**	12	100%	10	83%	1	8%
Especificidade	**12**	**100%**	12	100%	10	83%	1	8%
Acurácia	**24**	**100%**	24	100%	20	83%	2	8%

Fonte: Elaborado pela autora

Uma biblioteca foi preparada com diluições seriadas de 10^0 a 10^{-6} da OPV, que é distribuída em território nacional pelo Sistema Único de Saúde. A vacina contém os poliovírus atenuados 1, 2 e 3 em sua composição. Essa corrida foi realizada em uma *flowcell* que já havia sido usada e lavada previamente. As demais amostras das bibliotecas analisadas foram testadas em *flowcells* novas. Na tabela 7 consta as informações da corrida como: duração da corrida, quantidade de *reads* que passaram e falharam de acordo com o filtro de qualidade ≥ 9 (qc ≥ 9), bem como, a mediana do tamanho das *reads* que passaram e quantas foram identificadas com *barcodes* ou não.

Tabela 7 - Dados da corrida, *basecalling* e demultiplex da vacina de poliovírus atenuados (OPV)

Descrição	N. de amostras	N. de corridas	Tempo da corrida (h)	Total reads	Pass reads (qscore≥9)	Fail reads (qscore≥9)	Mediana pass reads (pb)	BC*	No-BC **
Vacina contra a poliomielite	7	Segunda	25,05	679.407	124.478	554.929	698	35.960	88.518

*BC = com barcode
**No-BC = sem barcode

Fonte: Elaborado pela autora

No gráfico 3, são representados os diagnósticos obtidos da biblioteca de diluição seriada da vacina anti pólio oral a partir da classificação taxonômica pelo Kraken2 e BLAST, A e B, respetivamente. Foi possível notar a presença de contaminação com vírus, possivelmente derivados de um sequenciamento anterior, já que usamos a *flowcell* lavada para essa biblioteca ao invés de uma nova. Apesar da contaminação foi possível identificar os enterovírus presentes na vacina oral até a diluição de 10^{-4} tanto com *wf 1* como pelo *wf* 2. Caso não houvesse tal contaminação talvez fosse possível detectar enterovírus em diluições menores.

Gráfico 3 - Classificação taxonômica das *reads* – vacina anti pólio ora (OPV)

A

Classificação taxonômica das *reads* do viroma em nível de espécie pelo Kraken2 - vacina anti pólio

Amostras	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV CHIKV DENV ZIKV HCV HIV SARS-Cov-2 Outros vírus; 0% 20% 40% 60% 80% 100%)	Diag. Obtido
1	1	Vacina (1E0)	EV	5817		EV+CHIKV+DENV
2	2	Vacina (1E-1)	EV	5603		EV+CHIKV+DENV+ZIKV+HCV+SARS-Cov-2
3	3	Vacina (1E-2)	EV	1912		EV+CHIKV+DENV+HCV+SARS-Cov-2
4	4	Vacina (1E-3)	EV	711		EV+CHIKV+DENV+ZIKV+HCV+SARS-Cov-2
5	5	Vacina (1E-4)	EV	244		EV+CHIKV+DENV+ZIKV+SARS-Cov-2
6	6	Vacina (1E-5)	EV	693		CHIKV+DENV+ZIKV+HCV+SARS-Cov-2
7	7	Vacina (1E-6)	EV	312		CHIKV+DENV+ZIKV+SARS-Cov-2

* Classified as virus (virome). Cut-off ≥ 00%.

B

Classificação taxonômica das *reads* do viroma em nível de espécie pelo BLAST - vacina anti pólio

Amostras	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV CHIKV DENV ZIKV HCV HIV SARS-Cov-2 Outros vírus; 0% 20% 40% 60% 80% 100%)	Diag. Obtido
1	1	Vacina (1E0)	EV	7996		EV+CHIKV+DENV+SARS-Cov-2
2	2	Vacina (1E-1)	EV	8398		EV+CHIKV+DENV+ZIKV+HCV+SARS-Cov-2
3	3	Vacina (1E-2)	EV	3301		EV+CHIKV+DENV+ZIKV+HCV+SARS-Cov-2
4	4	Vacina (1E-3)	EV	1127		EV+CHIKV+DENV+ZIKV+SARS-Cov-2
5	5	Vacina (1E-4)	EV	407		EV+CHIKV+DENV+ZIKV+SARS-Cov-2
6	6	Vacina (1E-5)	EV	984		EV+CHIKV+DENV+ZIKV+HCV+SARS-Cov-2
7	7	Vacina (1E-6)	EV	517		CHIKV+DENV+ZIKV+SARS-Cov-2

* Classified as virus (virome). Cut-off ≥ 0.0%.

Fonte: Elaborado pela autora

Entretanto, realizamos as análises de diferentes *cut-offs* em ambos os classificadores taxonômicos Kraken2 e BLAST para verificar a possibilidade de eliminar o erro de diagnóstico devido a contaminação e obtivemos melhor acurácia (50%) com o *cut-off* de 5% (dados não mostrados).

Após a extensa padronização dos métodos otimizados, seis bibliotecas foram preparadas e permitiram testar diferentes condições do experimento para realizar corridas no aparelho, cada biblioteca representa uma corrida (tabela 8).

A biblioteca 1 corresponde ao teste de 3 amostras distintas de isolados de culturas virais dos arbovírus dengue, zika e chikungunya. A biblioteca 2 corresponde ao teste de 3 amostras clínicas retrospectivas de líquor providas de forma anônima e voluntária por diferentes grupos de pesquisa da Fiocruz-BA e 2 amostras recentes de HIV e hepatite C com alta carga viral confirmada. A biblioteca 3 corresponde a amostras clínicas virais de chikungunya, dengue, hepatite C, hepatite B, HIV, SARS-CoV-2 e uma amostra de um isolado viral de zika vírus. As bibliotecas 4 e 5 correspondem ao teste de amostras clínicas virais de chikungunya, dengue, hepatite C, HIV e SARS-CoV-2 a fim de aumentar o tamanho amostral das análises.

Tabela 8 - Dados da corrida, *basecalling* e demultiplex

N. biblioteca	Descrição	N. de amostras	Tempo da corrida (h)	Total *reads*	Pass *reads* (qscore≥9)	Fail *reads* (qscore≥9)	Mediana Pass *reads* (pb)	BC*	No-BC **
1	Cultura	3	16,61	2.212.000	987.974	1.224.026	434	29.761	958.213
2	Clínica	5	38,8	1.817.886	256.916	1.560.970	512	73.591	183.325
3	Clínica	8	40.26	4.996.921	1.345.793	3.651.128	534	256.976	1.088.817
4	Clínica	11	16.69	392.258	96.373	295.885	469	18.478	77.895
5	Clínica	12	24,23	257.420	28.857	228.563	496	5.473	23.384

*BC = com barcode

**No-BC = sem barcode

Fonte: Elaborado pela autora

Nossos resultados com parâmetros mais liberais do Guppy *Barcode*r, BM2 (bc2), combinados com um filtro de qualidade 7 (qc7) mostram que *wf1* foi sensível para diagnosticar 21 das 35 amostras analisadas com o *cut-off* de 1,0% (gráfico 4).

Gráfico 4 - Classificação taxonômica em nível de espécie das *reads* analisadas pelo Kraken2 (qc7_bc2)

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
L1_C1	1	Cultura	CHIKV	316		CHIKV
L1_C2	2	Cultura	DENV	11491		DENV
L1_C3	3	Cultura	ZIKV	9524		ZIKV
L2_A1	4	Clínica	HIV	893		HIV
L2_A2	5	Clínica	HCV	1549		HCV
L3_A3	2	Clínica	CHIKV	3448		CHIKV
L3_A4	3	Clínica	DENV	5612		DENV
L3_A5	4	Clínica	HCV	0		Indetectável
L3_A6	5	Clínica	HIV	0		Indetectável
L3_A7	6	Clínica	SARS-CoV-2	15679		SARS-Cov-2
L3_A8	7	Clínica	HCV	2271		HCV
L3_C4	8	Cultura	ZIKV	0		Indetectável
L4_A9	1	Clínica	CHIKV	790		CHIKV
L4_A10	2	Clínica	CHIKV	3426		CHIKV
L4_A11	3	Clínica	DENV	92		DENV+ZIKV
L4_A12	4	Clínica	DENV	1895		DENV
L4_A13	5	Clínica	HCV	0		Indetectável
L4_A14	6	Clínica	HCV	0		Indetectável
L4_A15	7	Clínica	HIV	29		CHIKV+DENV
L4_A16	8	Clínica	HIV	19		ZIKV
L4_A17	9	Clínica	SARS-CoV-2	37		SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	89		SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	121		SARS-Cov-2
L5_A20	1	Clínica	CHIKV	224		CHIKV
L5_A21	2	Clínica	CHIKV	747		CHIKV
L5_A22	3	Clínica	DENV	384		DENV
L5_A23	4	Clínica	DENV	0		Indeterminado
L5_A24	5	Clínica	HCV	0		Indeterminado
L5_A25	6	Clínica	HCV	0		Indetectável
L5_A26	7	Clínica	HIV	0		Indetectável
L5_A27	8	Clínica	HIV	12		CHIKV+HIV
L5_A28	9	Clínica	HIV	8		HIV
L5_A29	10	Clínica	HIV	0		Indetectável
L5_A30	11	Clínica	SARS-CoV-2	0		Indetectável
L5_A31	12	Clínica	SARS-CoV-2	21		DENV+ZIKV+SARS-Cov-2

* Classified as virus (virome). Cut-off ≥ 1,0%.

Fonte: Elaborado pela autora

A acurácia, sensibilidade e especificidade do *wf1* foi de 56%, 60% e 51%, respectivamente para o *cut-off* = 1,0% com esses parâmetros mais liberais (tabela 9).

Tabela 9 - Cálculos de sensibilidade, especificidade e acurácia em diferentes *cut-offs* utilizando o classificador taxonômico Kraken2 (qc7_bc2).

Kraken2 - qc7_bc2

Sumário	Cut-off = 0.0%		Cut-off ≥ 0,5%		**Cut-off ≥ 1,0%**		Cut-off ≥ 5,0%	
Total	35	100%	35	100%	**35**	**100%**	35	100%
Sensibilidade	28	80%	21	60%	**21**	**60%**	16	46%
Especificidade	6	17%	17	49%	**18**	**51%**	16	46%
Acurácia	34	49%	38	54%	**39**	**56%**	32	46%

Fonte: Elaborado pela autora

Já na análise com qc9_bc4, obtivemos 23 resultados corretos em um universo de 35 amostras. Sendo uma amostra a mais que nos resultados apresentados anteriormente (gráfico 5).

Gráfico 5 - Classificação taxonômica em nível de espécie das *reads* analisadas pelo Kraken2 (qc9_bc4)

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
L1_C1	1	Cultura	CHIKV	249		CHIKV
L1_C2	2	Cultura	DENV	8625		DENV
L1_C3	3	Cultura	ZIKV	7479		ZIKV
L2_A1	4	Clínica	HIV	552		HIV
L2_A2	5	Clínica	HCV	777		HCV
L3_A3	2	Clínica	CHIKV	2112		CHIKV
L3_A4	3	Clínica	DENV	3904		DENV
L3_A5	4	Clínica	HCV	0		Indetectável
L3_A6	5	Clínica	HIV	0		Indetectável
L3_A7	6	Clínica	SARS-CoV-2	11416		SARS-Cov-2
L3_A8	7	Clínica	HCV	1475		HCV
L3_C4	8	Cultura	ZIKV	31827		ZIKV
L4_A9	1	Clínica	CHIKV	461		CHIKV
L4_A10	2	Clínica	CHIKV	2001		CHIKV
L4_A11	3	Clínica	DENV	61		DENV
L4_A12	4	Clínica	DENV	1238		DENV
L4_A13	5	Clínica	HCV	0		Indetectável
L4_A14	6	Clínica	HCV	0		Indetectável
L4_A15	7	Clínica	HIV	0		Indeterminado
L4_A16	8	Clínica	HIV	0		Indeterminado
L4_A17	9	Clínica	SARS-CoV-2	18		SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	53		SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	63		SARS-Cov-2
L5_A20	1	Clínica	CHIKV	214		CHIKV
L5_A21	2	Clínica	CHIKV	326		CHIKV
L5_A22	3	Clínica	DENV	204		DENV
L5_A23	4	Clínica	DENV	815		DENV
L5_A24	5	Clínica	HCV	0		Indetectável
L5_A25	6	Clínica	HCV	0		Indetectável
L5_A26	7	Clínica	HIV	0		Indetectável
L5_A27	8	Clínica	HIV	5		CHIKV+HIV
L5_A28	9	Clínica	HIV	3		HIV
L5_A29	10	Clínica	HIV	0		Indetectável
L5_A30	11	Clínica	SARS-CoV-2	0		Indetectável
L5_A31	12	Clínica	SARS-CoV-2	7		DENV

* Classified as virus (virome). Cut-off = 1,0%.

Fonte: Elaborado pela autora

O melhor *cut-off* analisado para esse classificador taxonômico foi o de 1%, apresentando sensibilidade de 66%, especificidade de 63% e acurácia de 64%. Na tabela 10 há a análise dos valores dos diferentes *cut-offs* analisados com suas respectivas sensibilidades, especificidades e acurácias para análises pelo Kraken2 com esses parâmetros mais liberais.

Tabela 10 - Cálculos de sensibilidade, especificidade e acurácia em diferentes *cut-offs* utilizando o classificador taxonômico Kraken2 (qc9_bc4).

Kraken2 - qc9

Sumário	Cut-off = 0.0%		Cut-off ≥ 0,5%		**Cut-off ≥ 1,0%**		Cut-off ≥ 5,0%	
Total	35	100%	35	100%	**35**	**100%**	35	100%
Sensibilidade	31	89%	24	69%	**23**	**66%**	16	46%
Especificidade	9	26%	20	57%	**22**	**63%**	16	46%
Acurácia	40	57%	44	63%	**45**	**64%**	32	46%

Fonte: Elaborado pela autora

O segundo *workflow* com parâmetros mais liberais do Guppy *Barcode*r combinados com um qc7 e BM2 do *barcoder* (bc2) mostram que *wf2* foi sensível para diagnosticar 23 amostras de 35 totais (gráfico 6).

Gráfico 6 - Classificação taxonômica em nível de espécie das *reads* analisadas pelo BLAST (qc7_bc2)

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
L1_C1	1	Cultura	CHIKV	212		CHIKV
L1_C2	2	Cultura	DENV	7350		DENV
L1_C3	3	Cultura	ZIKV	6319		ZIKV
L2_A1	4	Clínica	HIV	406		HIV
L2_A2	5	Clínica	HCV	506		HCV
L3_A3	2	Clínica	CHIKV	2238		CHIKV
L3_A4	3	Clínica	DENV	4753		DENV
L3_A5	4	Clínica	HCV	0		Indetectável
L3_A6	5	Clínica	HIV	0		Indetectável
L3_A7	6	Clínica	SARS-CoV-2	36573		SARS-Cov-2
L3_A8	7	Clínica	HCV	1350		HCV
L3_C4	8	Cultura	ZIKV	34308		ZIKV
L4_A9	1	Clínica	CHIKV	578		CHIKV
L4_A10	2	Clínica	CHIKV	2365		CHIKV+SARS-Cov-2
L4_A11	3	Clínica	DENV	131		CHIKV+DENV+ZIKV+SARS-Cov-2
L4_A12	4	Clínica	DENV	1576		DENV
L4_A13	5	Clínica	HCV	0		Indeterminado
L4_A14	6	Clínica	HCV	0		Indeterminado
L4_A15	7	Clínica	HIV	56		CHIKV+DENV+HCV+SARS-Cov-2
L4_A16	8	Clínica	HIV	88		CHIKV+DENV+ZIKV+SARS-Cov-2
L4_A17	9	Clínica	SARS-CoV-2	84		SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	203		SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	237		SARS-Cov-2
L5_A20	1	Clínica	CHIKV	1926		CHIKV
L5_A21	2	Clínica	CHIKV	2251		CHIKV
L5_A22	3	Clínica	DENV	1673		DENV
L5_A23	4	Clínica	DENV	3934		DENV
L5_A24	5	Clínica	HCV	207		CHIKV+DENV+SARS-Cov-2
L5_A25	6	Clínica	HCV	0		Indetectável
L5_A26	7	Clínica	HIV	0		Indeterminado
L5_A27	8	Clínica	HIV	83		CHIKV+DENV+HIV
L5_A28	9	Clínica	HIV	51		DENV+HIV
L5_A29	10	Clínica	HIV	0		Indeterminado
L5_A30	11	Clínica	SARS-CoV-2	0		Indeterminado
L5_A31	12	Clínica	SARS-CoV-2	185		CHIKV+DENV+ZIKV+SARS-Cov-2

* Classified as virus (virome). Cut-off ≥ 0,5%.

Fonte: Elaborado pela autora

Diferentemente dos *workflows* anteriores, para o BLAST, o melhor *cut-off* analisado foi o de 0,5%. A acuraria, sensibilidade e especificidade do *wf1* foi de 59%, 63% e 54%, respectivamente para o *cut-off* = 1,0% com esses parâmetros mais liberais (tabela 11).

Tabela 11 - Cálculos de sensibilidade, especificidade e acurácia em diferentes *cut-offs* utilizando o classificador taxonômico BLAST (qc7_bc2)

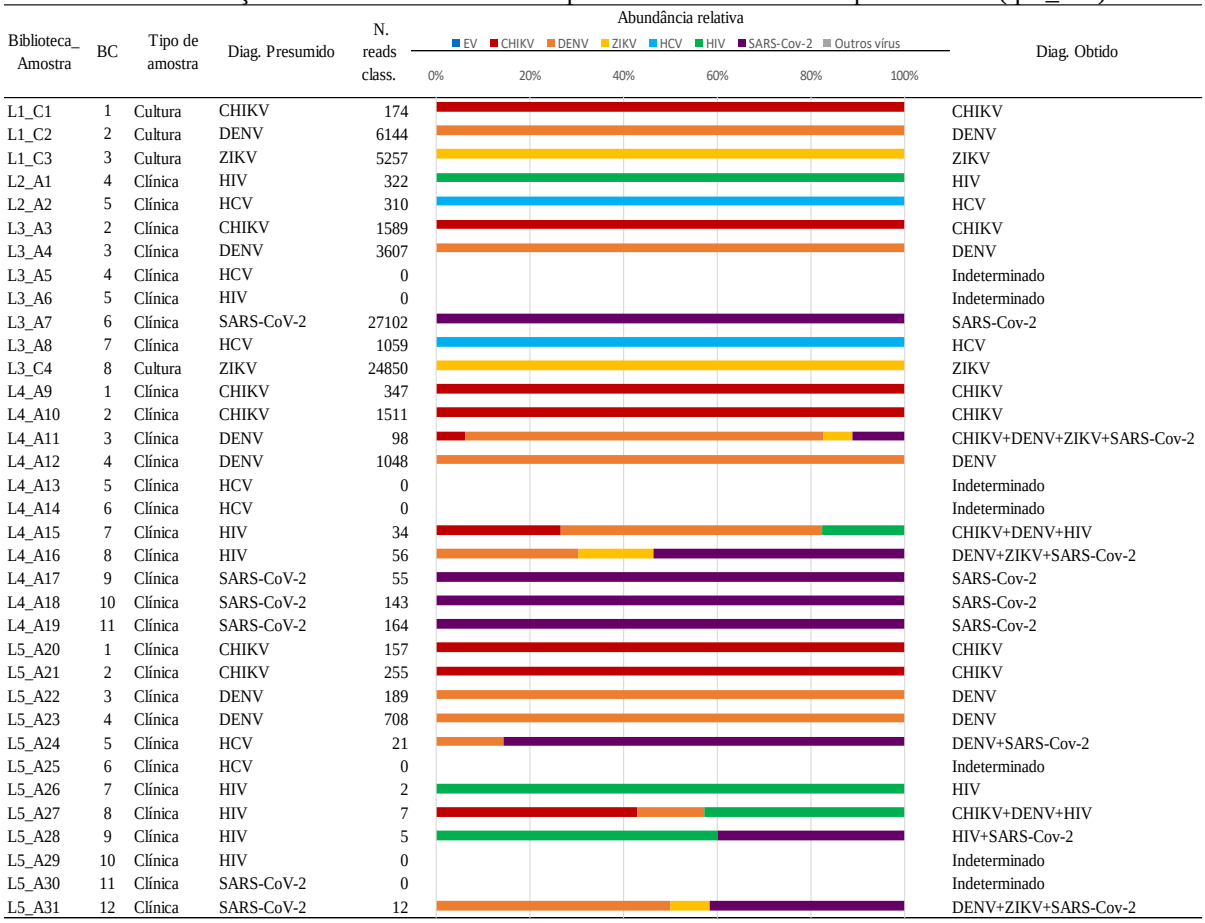
BLAST - qc7_bc2

Sumário	Cut-off = 0.0%		**Cut-off ≥ 0,5%**		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	35	100%	**35**	**100%**	35	100%	35	100%
Sensibilidade	32	91%	**23**	**66%**	22	63%	18	51%
Especificidade	5	14%	**18**	**51%**	19	54%	17	49%
Acurácia	37	53%	**41**	**59%**	41	59%	35	50%

Fonte: Elaborado pela autora

Já na análise com qc9_bc4, obtivemos 26 resultados corretos em um universo de 35 amostras. Sendo uma amostra a mais que nos resultados apresentados anteriormente (gráfico 7).

Gráfico 7 - Classificação taxonômica em nível de espécie das *reads* analisadas pelo BLAST (qc9_bc4)



Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Diag. Obtido
L1_C1	1	Cultura	CHIKV	174	CHIKV
L1_C2	2	Cultura	DENV	6144	DENV
L1_C3	3	Cultura	ZIKV	5257	ZIKV
L2_A1	4	Clínica	HIV	322	HIV
L2_A2	5	Clínica	HCV	310	HCV
L3_A3	2	Clínica	CHIKV	1589	CHIKV
L3_A4	3	Clínica	DENV	3607	DENV
L3_A5	4	Clínica	HCV	0	Indeterminado
L3_A6	5	Clínica	HIV	0	Indeterminado
L3_A7	6	Clínica	SARS-CoV-2	27102	SARS-Cov-2
L3_A8	7	Clínica	HCV	1059	HCV
L3_C4	8	Cultura	ZIKV	24850	ZIKV
L4_A9	1	Clínica	CHIKV	347	CHIKV
L4_A10	2	Clínica	CHIKV	1511	CHIKV
L4_A11	3	Clínica	DENV	98	CHIKV+DENV+ZIKV+SARS-Cov-2
L4_A12	4	Clínica	DENV	1048	DENV
L4_A13	5	Clínica	HCV	0	Indeterminado
L4_A14	6	Clínica	HCV	0	Indeterminado
L4_A15	7	Clínica	HIV	34	CHIKV+DENV+HIV
L4_A16	8	Clínica	HIV	56	DENV+ZIKV+SARS-Cov-2
L4_A17	9	Clínica	SARS-CoV-2	55	SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	143	SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	164	SARS-Cov-2
L5_A20	1	Clínica	CHIKV	157	CHIKV
L5_A21	2	Clínica	CHIKV	255	CHIKV
L5_A22	3	Clínica	DENV	189	DENV
L5_A23	4	Clínica	DENV	708	DENV
L5_A24	5	Clínica	HCV	21	DENV+SARS-Cov-2
L5_A25	6	Clínica	HCV	0	Indeterminado
L5_A26	7	Clínica	HIV	2	HIV
L5_A27	8	Clínica	HIV	7	CHIKV+DENV+HIV
L5_A28	9	Clínica	HIV	5	HIV+SARS-Cov-2
L5_A29	10	Clínica	HIV	0	Indeterminado
L5_A30	11	Clínica	SARS-CoV-2	0	Indeterminado
L5_A31	12	Clínica	SARS-CoV-2	12	DENV+ZIKV+SARS-Cov-2

* Classified as virus (virome). Cut-off = 0,5%.

Fonte: Elaborado pela autora

O melhor *cut-off* analisado foi o de 0,5%, apresentando sensibilidade de 74%, especificidade de 60% e acurácia de 67%. Na tabela 12 há a análise dos valores dos diferentes *cut-offs* analisados com suas respectivas sensibilidades, especificidades e acurácias para análises pelo BLAST.

Tabela 12 - Cálculos de sensibilidade, especificidade e acurácia em diferentes *cut-offs* utilizando o classificador taxonômico BLAST (qc9_bc4)

BLAST - qc9

Sumário	Cut-off = 0.0%		**Cut-off ≥ 0,5%**		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	35	100%	**35**	**100%**	35	100%	35	100%
Sensibilidade	32	91%	**26**	**74%**	24	69%	18	51%
Especificidade	10	29%	**21**	**60%**	22	63%	18	51%
Acurácia	42	60%	**47**	**67%**	46	66%	36	51%

Fonte: Elaborado pela autora

Devido ao desempenho do qc9_bc4 em relação ao qc7_bc2, definimos que todas as análises a seguir seriam realizadas com os parâmetros qc9_bc4 (esta informação não aparece mais nas ilustrações).

Além dos *wf1* e *wf2*, testamos outros *wf* alternativos: Genome Detective (*wf3*) e Epi2Me (*wf4*). A ferramenta web Genome Detective performou melhor que o *wf1* e *wf2* com acurácia com o *cut-off* = 0%, entretanto quando aplicado o *cut-off* = 1% foi possível reduzir "ruídos" de sequenciamento e obter o diagnóstico correto de 26 das nossas 35 amostras (gráfico 8).

Gráfico 8 - Classificação taxonômica em nível de espécie das *reads* analisadas pelo Genome Detective.

Biblioteca _Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
L1_C1	1	Cultura	CHIKV	289		CHIKV
L1_C2	2	Cultura	DENV	9237		DENV
L1_C3	3	Cultura	ZIKV	7726		ZIKV
L2_A1	4	Clínica	HIV	750		HIV
L2_A2	5	Clínica	HCV	1085		HCV
L3_A3	2	Clínica	CHIKV	2304		CHIKV
L3_A4	3	Clínica	DENV	5274		DENV
L3_A5	4	Clínica	HCV	0		Indetectável
L3_A6	5	Clínica	HIV	0		Indetectável
L3_A7	6	Clínica	SARS-CoV-2	33980		SARS-Cov-2
L3_A8	7	Clínica	HCV	1604		HCV
L3_C4	8	Cultura	ZIKV	32468		ZIKV
L4_A9	1	Clínica	CHIKV	500		CHIKV
L4_A10	2	Clínica	CHIKV	2123		CHIKV
L4_A11	3	Clínica	DENV	110		DENV
L4_A12	4	Clínica	DENV	1518		DENV
L4_A13	5	Clínica	HCV	0		Indetectável
L4_A14	6	Clínica	HCV	0		Indetectável
L4_A15	7	Clínica	HIV	39		DENV+HIV
L4_A16	8	Clínica	HIV	22		DENV
L4_A17	9	Clínica	SARS-CoV-2	70		SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	182		SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	221		SARS-Cov-2
L5_A20	1	Clínica	CHIKV	254		CHIKV
L5_A21	2	Clínica	CHIKV	374		CHIKV
L5_A22	3	Clínica	DENV	278		DENV
L5_A23	4	Clínica	DENV	973		DENV
L5_A24	5	Clínica	HCV	21		SARS-Cov-2
L5_A25	6	Clínica	HCV	0		Indetectável
L5_A26	7	Clínica	HIV	4		HIV
L5_A27	8	Clínica	HIV	12		CHIKV+DENV+HIV
L5_A28	9	Clínica	HIV	7		HIV
L5_A29	10	Clínica	HIV	0		Indetectável
L5_A30	11	Clínica	SARS-CoV-2	0		Indetectável
L5_A31	12	Clínica	SARS-CoV-2	20		DENV+SARS-Cov-2

* Classified as virus (virome). Cut-off = 1,0%.

Fonte: Elaborado pela autora

A acuraria, sensibilidade e especificidade do *wf3* foi de 70%, 74% e 66%, respectivamente para o *cut-off* = 1,0%. Na tabela 13 as análises feitas pelo Genome Detective estão sumarizadas.

Tabela 13 - Cálculos de sensibilidade, especificidade e acurácia em diferentes *cut-offs* utilizando o classificador taxonômico Genome Detective

Genome Detective

Sumário	Cut-off = 0.0%		Cut-off ≥ 0,5%		**Cut-off ≥ 1,0%**		Cut-off ≥ 5,0%	
Total	35	100%	35	100%	**35**	**100%**	35	100%
Sensibilidade	34	97%	26	74%	**26**	**74%**	19	54%
Especificidade	7	20%	20	57%	**23**	**66%**	18	51%
Acurácia	41	59%	46	66%	**49**	**70%**	37	53%

Fonte: Elaborado pela autora

Por fim, nossos resultados utilizando a plataforma online da ONT, o Epi2ME, com *cut-off* de 0,5%, obtivemos 23 diagnósticos sensíveis em um universo de 35 amostras (gráfico 9).

Gráfico 9 - Classificação taxonômica em nível de espécie das *reads* analisadas pelo Epi2ME (qc9).

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. reads class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0% 20% 40% 60% 80% 100%)	Diag. Obtido
L1_C1	1	Cultura	CHIKV	249		CHIKV+DENV
L1_C2	2	Cultura	DENV	7974		DENV
L1_C3	3	Cultura	ZIKV	7561		ZIKV
L2_A1	4	Clínica	HIV	359		HIV
L2_A2	5	Clínica	HCV	627		HCV
L3_A3	2	Clínica	CHIKV	2005		CHIKV
L3_A4	3	Clínica	DENV	2627		DENV
L3_A5	4	Clínica	HCV	0		Indeterminado
L3_A6	5	Clínica	HIV	190		Indetectável
L3_A7	6	Clínica	SARS-CoV-2	33758		SARS-Cov-2
L3_A8	7	Clínica	HCV	1406		HCV
L3_C4	8	Cultura	ZIKV	31761		ZIKV
L4_A9	1	Clínica	CHIKV	430		CHIKV
L4_A10	2	Clínica	CHIKV	1879		CHIKV
L4_A11	3	Clínica	DENV	89		CHIKV+DENV+ZIKV+SARS-Cov-2
L4_A12	4	Clínica	DENV	779		DENV
L4_A13	5	Clínica	HCV	10		Indetectável
L4_A14	6	Clínica	HCV	17		Indetectável
L4_A15	7	Clínica	HIV	48		CHIKV+DENV+HIV+SARS-Cov-2
L4_A16	8	Clínica	HIV	29		DENV+ZIKV
L4_A17	9	Clínica	SARS-CoV-2	70		SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	180		SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	221		SARS-Cov-2
L5_A20	1	Clínica	CHIKV	201		CHIKV
L5_A21	2	Clínica	CHIKV	316		CHIKV
L5_A22	3	Clínica	DENV	0		Indetectável
L5_A23	4	Clínica	DENV	0		Indetectável
L5_A24	5	Clínica	HCV	3		Indetectável
L5_A25	6	Clínica	HCV	0		Indetectável
L5_A26	7	Clínica	HIV	0		Indetectável
L5_A27	8	Clínica	HIV	7		CHIKV+DENV+HIV
L5_A28	9	Clínica	HIV	5		DENV+HIV
L5_A29	10	Clínica	HIV	4		SARS-Cov-2
L5_A30	11	Clínica	SARS-CoV-2	3		SARS-Cov-2
L5_A31	12	Clínica	SARS-CoV-2	9		ZIKV+SARS-Cov-2

* Classified as virus (virome). Cut-off = 0,5%.

Fonte: Elaborado pela autora

O melhor *cut-off* analisado para esse classificador taxonômico foi o de 0,5%, apresentando sensibilidade de 66%, especificidade de 49% e acurácia de 57%. Na tabela 14 há a análise dos valores dos diferentes *cut-offs* analisados com suas respectivas sensibilidades, especificidades e acurácias para análises pelo Epi2ME.

Tabela 14 - Cálculos de sensibilidade, especificidade e acurácia em diferentes *cut-offs* utilizando o classificador taxonômico Epi2ME

Epi2ME - qc9

Sumário	Cut-off = 0.0%		**Cut-off ≥ 0,5%**		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	35	100%	**35**	**100%**	35	100%	35	100%
Sensibilidade	26	74%	**23**	**66%**	21	60%	17	49%
Especificidade	9	26%	**17**	**49%**	16	46%	17	49%
Acurácia	35	50%	**40**	**57%**	37	53%	34	49%

Fonte: Elaborado pela autora

A análises dos controles (DENV_FTA1, sequencias randômicas e da OPV), oferecem resultados importantes sobre os desempenhos dos *workflows*. Para os dados de treinamento DENV_FTA_1, nenhum *cut-off* ou um *cut-off* de 0,5% foi o suficiente para fornecer diagnóstico acurado das amostras. Entretanto, para as análises da diluição seriada da vacina, a contaminação presente possivelmente influenciou o resultado obtido, necessitando a utilização de um *cut-off* de 5% para que a melhor acurácia avaliada nessa biblioteca (50%) fosse alcançada.
Comparativamente, entre os 4 *workflows* testados, o *wf3* que usa o Genome detective, possui a melhor acurácia (70%), melhor especificidade (66%) e melhor sensibilidade (74%). Foi possível também comparar os resultados de parâmetros mais liberais para o *workflow* 1 e 2, e notar uma piora na sensibilidade e especificidade da análise, quando comparadas aos parâmetros recomentados.

5.3 DIAGNÓSTICO POR CLASSIFICAÇÃO TAXONÔMICA A NÍVEL DE *CONTIGS*

Após finalização de todas as análises referentes a classificação taxonômica do vírus a nível de *reads*, realizamos análises a nível de *contigs*. Primariamente avaliamos as *contigs* obtidas dos dados de treinamento e os *cut-offs* 0%, 0,5%, 1% e 5%, respectivamente de A à D (gráfico 10). Esses resultados demonstram graficamente que foi possível montar *contigs* de 9 amostras diferente com o *cut-off* de 0% e que esse número decresceu com o aumento do *cut-off*, aumentando assim o número de amostras consideradas indetectáveis.

Gráfico 10 - Classificação taxonômica das *contigs* pelo Kraken2 – dados de treinamento

A

Biblioteca_Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Abundância relativa	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	2		DENV
DS_P2	2	Produto PCR	DENV	2		DENV
DS_P3	3	Produto PCR	DENV	40		DENV+ZIKV
DS_P4	4	Produto PCR	DENV	2		DENV
DS_P5	5	Produto PCR	DENV	0		Indetectável
DS_P6	6	Produto PCR	DENV	5		DENV
DS_P7	7	Produto PCR	DENV	0		Indetectável
DS_P8	8	Produto PCR	DENV	7		DENV
DS_P9	9	Produto PCR	DENV	40		DENV
DS_P10	10	Produto PCR	DENV	1		DENV
DS_P11	11	Produto PCR	DENV	2		DENV
DS_P12	12	Produto PCR	DENV	0		Indetectável

* Classified as virus (virome). Cut-off = 0,0%.

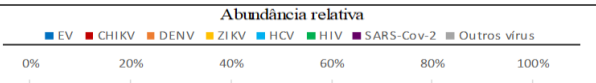
B

Biblioteca_Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Abundância relativa	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	0		Indeterminado
DS_P2	2	Produto PCR	DENV	0		Indetectável
DS_P3	3	Produto PCR	DENV	36		DENV
DS_P4	4	Produto PCR	DENV	2		DENV
DS_P5	5	Produto PCR	DENV	0		Indetectável
DS_P6	6	Produto PCR	DENV	5		DENV
DS_P7	7	Produto PCR	DENV	0		Indetectável
DS_P8	8	Produto PCR	DENV	7		DENV
DS_P9	9	Produto PCR	DENV	40		DENV
DS_P10	10	Produto PCR	DENV	1		DENV
DS_P11	11	Produto PCR	DENV	2		DENV
DS_P12	12	Produto PCR	DENV	0		Indetectável

* Classified as virus (virome). Cut-off = 0,5%.

C

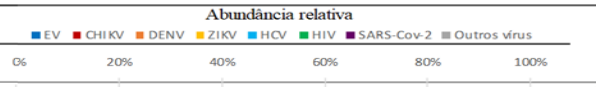
Biblioteca_Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Abundância relativa	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	0		Indetectável
DS_P2	2	Produto PCR	DENV	0		Indetectável
DS_P3	3	Produto PCR	DENV	36		DENV
DS_P4	4	Produto PCR	DENV	2		DENV
DS_P5	5	Produto PCR	DENV	0		Indetectável
DS_P6	6	Produto PCR	DENV	3		DENV
DS_P7	7	Produto PCR	DENV	0		Indetectável
DS_P8	8	Produto PCR	DENV	4		DENV
DS_P9	9	Produto PCR	DENV	35		DENV
DS_P10	10	Produto PCR	DENV	0		Indetectável
DS_P11	11	Produto PCR	DENV	2		DENV
DS_P12	12	Produto PCR	DENV	0		Indetectável



* Classified as virus (virome). Cut-off = 1,0%.

D

Biblioteca_Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Abundância relativa	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	0		Indetectável
DS_P2	2	Produto PCR	DENV	5006		DENV
DS_P3	3	Produto PCR	DENV	0		Indetectável
DS_P4	4	Produto PCR	DENV	0		Indetectável
DS_P5	5	Produto PCR	DENV	0		Indetectável
DS_P6	6	Produto PCR	DENV	0		Indetectável
DS_P7	7	Produto PCR	DENV	0		Indetectável
DS_P8	8	Produto PCR	DENV	0		Indetectável
DS_P9	9	Produto PCR	DENV	0		Indetectável
DS_P10	10	Produto PCR	DENV	0		Indetectável
DS_P11	11	Produto PCR	DENV	0		Indetectável
DS_P12	12	Produto PCR	DENV	0		Indetectável



* Classified as virus (virome). Cut-off = 5,0%.

Fonte: Elaborado pela autora

Dentre os diferentes *cut-offs* analisados nas *contigs*, a acuraria, sensibilidade e especificidade do *wf1* foi de 71%, 75% e 67%, respectivamente para o *cut-off* de 0% (tabela 15).

Tabela15 - Cálculos de sensibilidade, especificidade e acurácia de *contigs* em diferentes *cut-offs* utilizando o classificador taxonômico Kraken2 – dados de treinamento

Kraken2 - controles

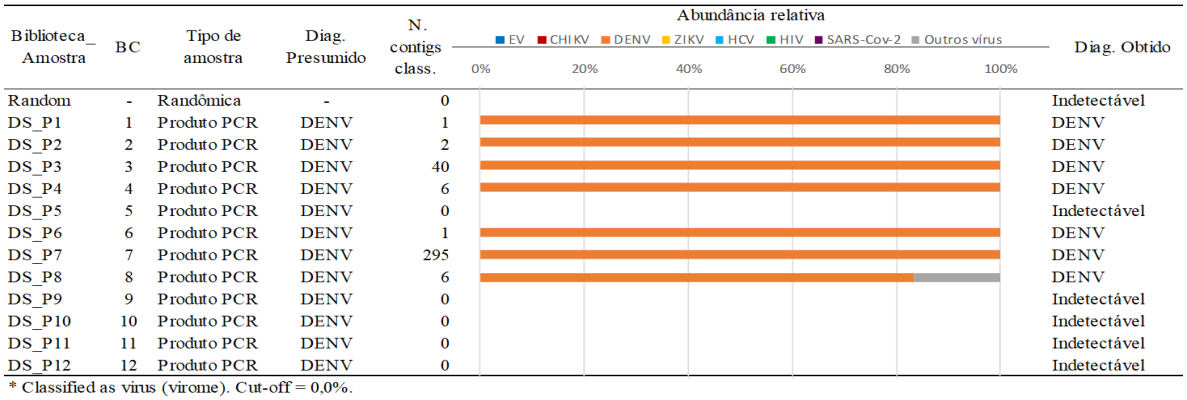
Sumário	**Cut-off = 0.0%**		Cut-off ≥ 0.5%		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	**12**	**100%**	12	100%	12	100%	12	100%
Sensibilidade	**9**	**75%**	7	58%	6	50%	1	8%
Especificidade	**8**	**67%**	7	58%	6	50%	1	8%
Acurácia	**17**	**71%**	14	58%	12	50%	2	8%

Fonte: Elaborado pela autora

Após as análises do *wf1*, realizamos análises a nível de *contigs* usando o *wf2*. Nos dados de treinamento testamos os *cut-offs* 0%, 0,5%, 1% e 5%, respectivamente de A à D (gráfico 11). Diferentemente do método anterior, foi possível classificar apenas 7 amostras diferentes com o *cut-off* de 0% e esse número também vai decrescendo com o aumento dos *cut-offs*.

Gráfico 11 - Classificação taxonômica das *contigs* pelo BLAST – dados de treinamento

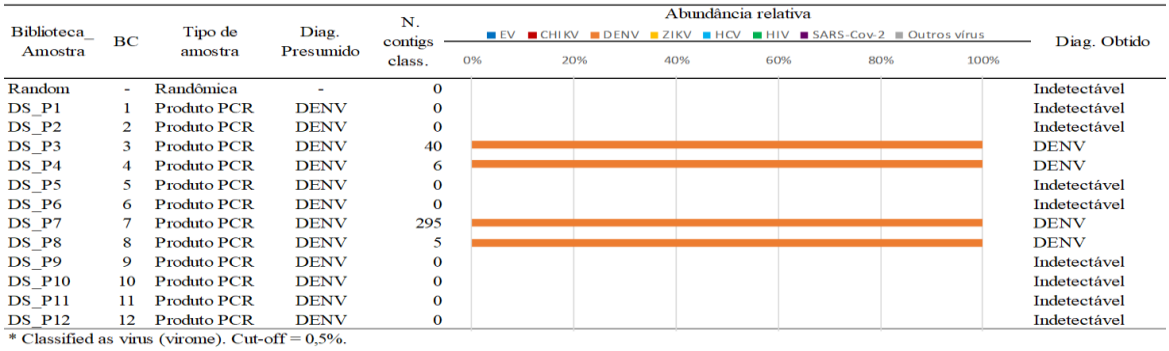
A



Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Diag. Obtido
Random	-	Randômica	-	0	Indetectável
DS_P1	1	Produto PCR	DENV	1	DENV
DS_P2	2	Produto PCR	DENV	2	DENV
DS_P3	3	Produto PCR	DENV	40	DENV
DS_P4	4	Produto PCR	DENV	6	DENV
DS_P5	5	Produto PCR	DENV	0	Indetectável
DS_P6	6	Produto PCR	DENV	1	DENV
DS_P7	7	Produto PCR	DENV	295	DENV
DS_P8	8	Produto PCR	DENV	6	DENV
DS_P9	9	Produto PCR	DENV	0	Indetectável
DS_P10	10	Produto PCR	DENV	0	Indetectável
DS_P11	11	Produto PCR	DENV	0	Indetectável
DS_P12	12	Produto PCR	DENV	0	Indetectável

* Classified as virus (virome). Cut-off = 0,0%.

B



Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Diag. Obtido
Random	-	Randômica	-	0	Indetectável
DS_P1	1	Produto PCR	DENV	0	Indetectável
DS_P2	2	Produto PCR	DENV	0	Indetectável
DS_P3	3	Produto PCR	DENV	40	DENV
DS_P4	4	Produto PCR	DENV	6	DENV
DS_P5	5	Produto PCR	DENV	0	Indetectável
DS_P6	6	Produto PCR	DENV	0	Indetectável
DS_P7	7	Produto PCR	DENV	295	DENV
DS_P8	8	Produto PCR	DENV	5	DENV
DS_P9	9	Produto PCR	DENV	0	Indetectável
DS_P10	10	Produto PCR	DENV	0	Indetectável
DS_P11	11	Produto PCR	DENV	0	Indetectável
DS_P12	12	Produto PCR	DENV	0	Indetectável

* Classified as virus (virome). Cut-off = 0,5%.

C

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	0		Indetectável
DS_P2	2	Produto PCR	DENV	0		Indetectável
DS_P3	3	Produto PCR	DENV	40		DENV
DS_P4	4	Produto PCR	DENV	6		DENV
DS_P5	5	Produto PCR	DENV	0		Indetectável
DS_P6	6	Produto PCR	DENV	0		Indetectável
DS_P7	7	Produto PCR	DENV	295		DENV
DS_P8	8	Produto PCR	DENV	5		DENV
DS_P9	9	Produto PCR	DENV	0		Indetectável
DS_P10	10	Produto PCR	DENV	0		Indetectável
DS_P11	11	Produto PCR	DENV	0		Indetectável
DS_P12	12	Produto PCR	DENV	0		Indetectável

* Classified as virus (virome). Cut-off = 1,0%.

D

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Abundância relativa (EV, CHIKV, DENV, ZIKV, HCV, HIV, SARS-Cov-2, Outros vírus; 0%, 20%, 40%, 60%, 80%, 100%)	Diag. Obtido
Random	-	Randômica	-	0		Indetectável
DS_P1	1	Produto PCR	DENV	0		Indetectável
DS_P2	2	Produto PCR	DENV	0		Indetectável
DS_P3	3	Produto PCR	DENV	40		DENV
DS_P4	4	Produto PCR	DENV	0		Indetectável
DS_P5	5	Produto PCR	DENV	0		Indetectável
DS_P6	6	Produto PCR	DENV	0		Indetectável
DS_P7	7	Produto PCR	DENV	0		Indetectável
DS_P8	8	Produto PCR	DENV	0		Indetectável
DS_P9	9	Produto PCR	DENV	0		Indetectável
DS_P10	10	Produto PCR	DENV	0		Indetectável
DS_P11	11	Produto PCR	DENV	0		Indetectável
DS_P12	12	Produto PCR	DENV	0		Indetectável

* Classified as virus (virome). Cut-off = 5,0%.

Fonte: Elaborado pela autora

Dentre os diferentes *cut-offs* analisados nas *contigs*, o *cut-off* de 1,0% apresentou acuraria, sensibilidade e especificidade de 58% (tabela 16).

Tabela 16 - Cálculos de sensibilidade, especificidade e acurácia de *contigs* em diferentes *cut-offs* utilizando o classificador taxonômico BLAST – dados de treinamento

BLAST - controles

Sumário	**Cut-off = 0.0%**		Cut-off ≥ 0.5%		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	**12**	**100%**	12	100%	12	100%	12	100%
Sensibilidade	**7**	**58%**	4	33%	4	33%	1	8%
Especificidade	**7**	**58%**	4	33%	4	33%	1	8%
Acurácia	**14**	**58%**	8	33%	8	33%	2	8%

Fonte: Elaborado pela autora

Seguindo com a análise de nossos controles, no gráfico 12, são representados os diagnósticos obtidos da biblioteca de diluição seriada da OPV a partir da classificação taxonômica a nível de *contigs* pelo *wf1* e *wf2*, A e B, respetivamente. Concordando com os resultados a nível de *reads*, contaminação foi possível identificar os enterovírus presentes na OPV até a diluição de 10^{-4} nas análises com o Kraken2 e BLAST.

Gráfico 12 - Classificação taxonômica das *contigs* – vacina anti pólio oral (OPV)

A

Classificação taxonômica das contigs do viroma em nível de espécie pelo Kraken2 - vacina anti pólio

Amostras	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Abundância relativa (EV CHIKV DENV ZIKV HCV HIV SARS-Cov-2 Outros vírus; 0% 20% 40% 60% 80% 100%)	Diag. Obtido
1	1	Vacina (1E0)	EV	6		EV
2	2	Vacina (1E-1)	EV	4		EV
3	3	Vacina (1E-2)	EV	2		EV
4	4	Vacina (1E-3)	EV	13		EV
5	5	Vacina (1E-4)	EV	10		EV
6	6	Vacina (1E-5)	EV	0		Indetectável
7	7	Vacina (1E-6)	EV	0		Indetectável

* Classified as virus (virome). Cut-off = 0,0%.

B

Classificação taxonômica das contigs do viroma em nível de espécie pelo BLAST - vacina anti pólio

Amostras	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Abundância relativa (EV CHIKV DENV ZIKV HCV HIV SARS-Cov-2 Outros vírus; 0% 20% 40% 60% 80% 100%)	Diag. Obtido
1	1	Vacina (1E0)	EV	5		EV
2	2	Vacina (1E-1)	EV	4		EV
3	3	Vacina (1E-2)	EV	146		EV
4	4	Vacina (1E-3)	EV	2		EV
5	5	Vacina (1E-4)	EV	1		EV
6	6	Vacina (1E-5)	EV	0		Indetectável
7	7	Vacina (1E-6)	EV	0		Indetectável

* Classified as virus (virome). Cut-off = 0,0%.

Fonte: Elaborado pela autora

Na análise do *cut-off* de 0% a acurácia, sensibilidade e especificidade obtiveram os mesmos valores nos *workflows* 1 e 2 (71%).

Já nas análises das demais amostras para classificação taxonômica a nível de *contigs*, obtivemos 19 diagnósticos sensíveis com o *wf1* em um universo de 35 amostras (gráfico 13).

Gráfico 13 - Classificação taxonômica em nível de espécie das *contigs* analisadas pelo Kraken2

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs	Diag. Obtido
L1_C1	1	Cultura	CHIKV	3	CHIKV
L1_C2	2	Cultura	DENV	1	DENV
L1_C3	3	Cultura	ZIKV	2	ZIKV
L2_A1	4	Clínica	HIV	1	HIV
L2_A2	5	Clínica	HCV	47	HCV
L3_A3	2	Clínica	CHIKV	2	CHIKV
L3_A4	3	Clínica	DENV	1	DENV
L3_A5	4	Clínica	HCV	0	Indetectável
L3_A6	5	Clínica	HIV	0	Indetectável
L3_A7	6	Clínica	SARS-CoV-2	1	SARS-Cov-2
L3_A8	7	Clínica	HCV	130	HCV
L3_C4	8	Cultura	ZIKV	126	ZIKV
L4_A9	1	Clínica	CHIKV	1	CHIKV
L4_A10	2	Clínica	CHIKV	1	CHIKV
L4_A11	3	Clínica	DENV	2	DENV
L4_A12	4	Clínica	DENV	1	DENV
L4_A13	5	Clínica	HCV	0	Indetectável
L4_A14	6	Clínica	HCV	0	Indetectável
L4_A15	7	Clínica	HIV	0	Indetectável
L4_A16	8	Clínica	HIV	0	Indetectável
L4_A17	9	Clínica	SARS-CoV-2	2	SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	3	SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	5	SARS-Cov-2
L5_A20	1	Clínica	CHIKV	2	CHIKV
L5_A21	2	Clínica	CHIKV	1	CHIKV
L5_A22	3	Clínica	DENV	1	DENV
L5_A23	4	Clínica	DENV	74	DENV
L5_A24	5	Clínica	HCV	0	Indetectável
L5_A25	6	Clínica	HCV	0	Indetectável
L5_A26	7	Clínica	HIV	0	Indetectável
L5_A27	8	Clínica	HIV	0	Indetectável
L5_A28	9	Clínica	HIV	0	Indetectável
L5_A29	10	Clínica	HIV	0	Indetectável
L5_A30	11	Clínica	SARS-CoV-2	0	Indetectável
L5_A31	12	Clínica	SARS-CoV-2	0	Indetectável

* Classified as virus (virome). Cut-off = 0,0%.

Fonte: Elaborado pela autora

Os *cut-offs* analisados de 0% e 0,5%, apresentaram resultados de sensibilidade, especificidade e acurácia de 54% (tabela 17).

Tabela 17 - Cálculos de sensibilidade, especificidade e acurácia de *contigs* em diferentes *cut-offs* utilizando o classificador taxonômico Kraken2

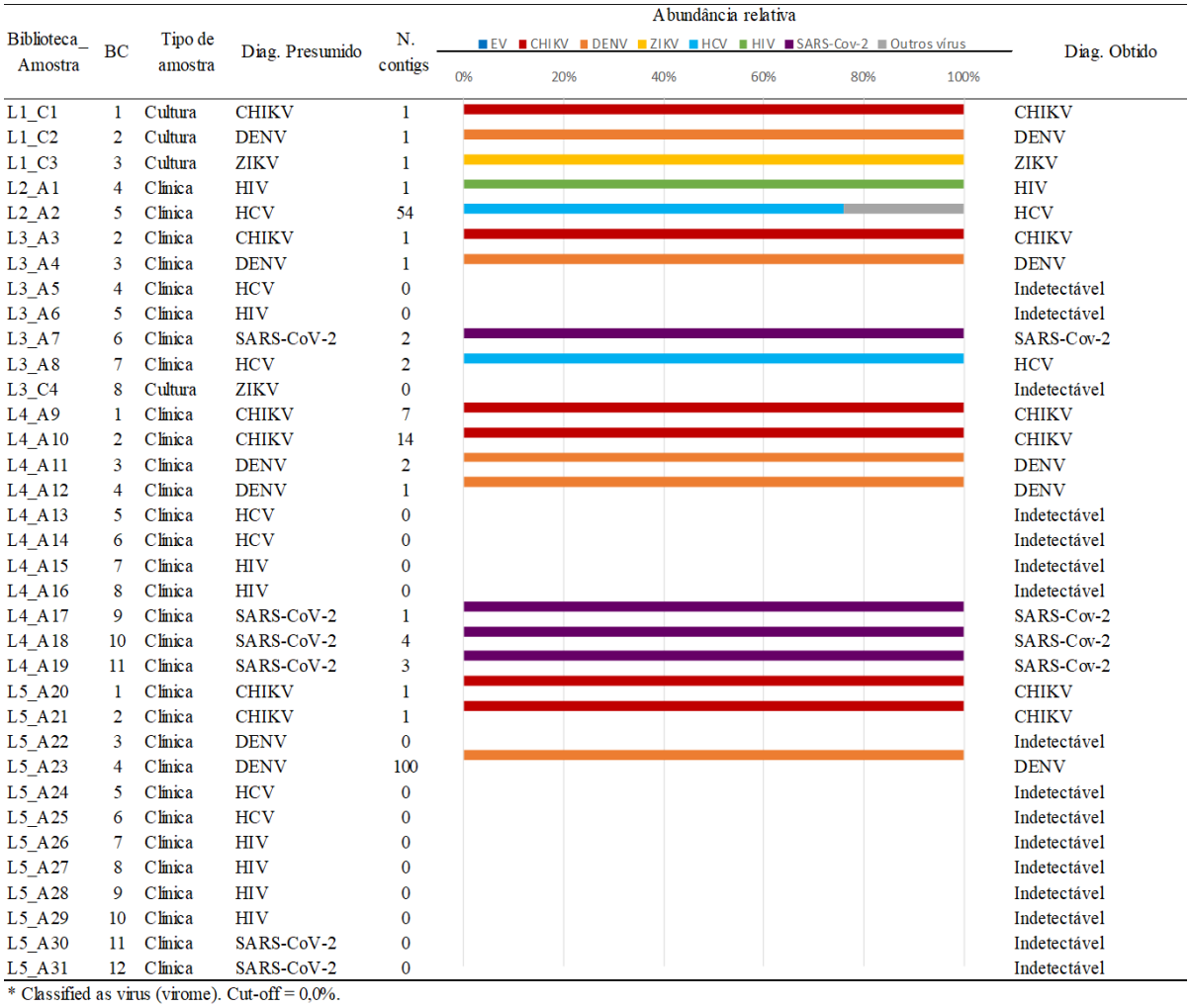
Kraken2 - qc9

Sumário	**Cut-off = 0.0%**		Cut-off ≥ 0,5%		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	**35**	**100%**	35	100%	35	100%	35	100%
Sensibilidade	**19**	**54%**	19	54%	18	51%	14	40%
Especificidade	**19**	**54%**	19	54%	18	51%	14	40%
Acurácia	**38**	**54%**	38	54%	36	51%	28	40%

Fonte: Elaborado pela autora

Em nosso segundo *workflow*, o BLAST, ao analisar as *contigs*, foi sensível para diagnosticar 18 amostras de 35 totais (gráfico 14).

Gráfico 14 - Classificação taxonômica em nível de espécie das *contigs* analisadas pelo BLAST



Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs	Diag. Obtido
L1_C1	1	Cultura	CHIKV	1	CHIKV
L1_C2	2	Cultura	DENV	1	DENV
L1_C3	3	Cultura	ZIKV	1	ZIKV
L2_A1	4	Clínica	HIV	1	HIV
L2_A2	5	Clínica	HCV	54	HCV
L3_A3	2	Clínica	CHIKV	1	CHIKV
L3_A4	3	Clínica	DENV	1	DENV
L3_A5	4	Clínica	HCV	0	Indetectável
L3_A6	5	Clínica	HIV	0	Indetectável
L3_A7	6	Clínica	SARS-CoV-2	2	SARS-Cov-2
L3_A8	7	Clínica	HCV	2	HCV
L3_C4	8	Cultura	ZIKV	0	Indetectável
L4_A9	1	Clínica	CHIKV	7	CHIKV
L4_A10	2	Clínica	CHIKV	14	CHIKV
L4_A11	3	Clínica	DENV	2	DENV
L4_A12	4	Clínica	DENV	1	DENV
L4_A13	5	Clínica	HCV	0	Indetectável
L4_A14	6	Clínica	HCV	0	Indetectável
L4_A15	7	Clínica	HIV	0	Indetectável
L4_A16	8	Clínica	HIV	0	Indetectável
L4_A17	9	Clínica	SARS-CoV-2	1	SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	4	SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	3	SARS-Cov-2
L5_A20	1	Clínica	CHIKV	1	CHIKV
L5_A21	2	Clínica	CHIKV	1	CHIKV
L5_A22	3	Clínica	DENV	0	Indetectável
L5_A23	4	Clínica	DENV	100	DENV
L5_A24	5	Clínica	HCV	0	Indetectável
L5_A25	6	Clínica	HCV	0	Indetectável
L5_A26	7	Clínica	HIV	0	Indetectável
L5_A27	8	Clínica	HIV	0	Indetectável
L5_A28	9	Clínica	HIV	0	Indetectável
L5_A29	10	Clínica	HIV	0	Indetectável
L5_A30	11	Clínica	SARS-CoV-2	0	Indetectável
L5_A31	12	Clínica	SARS-CoV-2	0	Indetectável

* Classified as virus (virome). Cut-off = 0,0%.

Os *cut-offs* de 0% e o de 0,5%, apresentaram sensibilidade, especificidade e acurácia de 51% (tabela 18).
Fonte: Elaborado pela autora

Tabela 18 - Cálculos de sensibilidade, especificidade e acurácia de *contigs* em diferentes *cut-offs* utilizando o classificador taxonômico BLAST

BLAST - qc9

Sumário	Cut-off = 0.0%		**Cut-off ≥ 0,5%**		Cut-off ≥ 1,0%		Cut-off ≥ 5,0%	
Total	35	100%	**35**	**100%**	35	100%	35	100%
Sensibilidade	18	51%	**18**	**51%**	17	49%	14	40%
Especificidade	18	51%	**18**	**51%**	17	49%	14	40%
Acurácia	36	51%	**36**	**51%**	34	49%	28	40%

Fonte: Elaborado pela autora

Nossos resultados utilizando a ferramenta web Genome Detective para análise de *contigs* desempenhou de forma igual nos *cut-offs* de 0%, 0,5% e 1%. Apesar de "ruídos" de sequenciamento, o *wf3* obteve o diagnóstico sensível de 29 das 35 amostras testadas (gráfico 15).

Gráfico 15 - Classificação taxonômica em nível de espécie das *contigs* analisadas pelo Genome Detective

Biblioteca_ Amostra	BC	Tipo de amostra	Diag. Presumido	N. contigs class.	Diag. Obtido
L1_C1	1	Cultura	CHIKV	2	CHIKV
L1_C2	2	Cultura	DENV	1	Indetectável
L1_C3	3	Cultura	ZIKV	18	CHIKV+DENV+ZIKV
L2_A1	4	Clínica	HIV	7	HIV
L2_A2	5	Clínica	HCV	9	HCV
L3_A3	2	Clínica	CHIKV	4	CHIKV+ZIKV
L3_A4	3	Clínica	DENV	1	DENV
L3_A5	4	Clínica	HCV	28	DENV+HCV+SARS-Cov-2
L3_A6	5	Clínica	HIV	34	DENV+ZIKV+HIV+SARS-Cov-2
L3_A7	6	Clínica	SARS-CoV-2	1	SARS-Cov-2
L3_A8	7	Clínica	HCV	33	DENV+ZIKV+HCV+SARS-Cov-2
L3_C4	8	Cultura	ZIKV	33	CHIKV+DENV+ZIKV+SARS-Cov-2
L4_A9	1	Clínica	CHIKV	17	CHIKV
L4_A10	2	Clínica	CHIKV	12	CHIKV+DENV+SARS-Cov-2
L4_A11	3	Clínica	DENV	30	CHIKV+DENV+ZIKV+SARS-Cov-2
L4_A12	4	Clínica	DENV	1	DENV
L4_A13	5	Clínica	HCV	3	DENV
L4_A14	6	Clínica	HCV	5	HCV
L4_A15	7	Clínica	HIV	20	DENV+HIV+SARS-Cov-2
L4_A16	8	Clínica	HIV	22	CHIKV+DENV+ZIKV+HIV
L4_A17	9	Clínica	SARS-CoV-2	9	SARS-Cov-2
L4_A18	10	Clínica	SARS-CoV-2	7	SARS-Cov-2
L4_A19	11	Clínica	SARS-CoV-2	5	SARS-Cov-2
L5_A20	1	Clínica	CHIKV	4	CHIKV
L5_A21	2	Clínica	CHIKV	13	CHIKV
L5_A22	3	Clínica	DENV	1	DENV
L5_A23	4	Clínica	DENV	1	DENV
L5_A24	5	Clínica	HCV	15	DENV+SARS-Cov-2
L5_A25	6	Clínica	HCV	3	HCV
L5_A26	7	Clínica	HIV	3	HIV
L5_A27	8	Clínica	HIV	8	CHIKV+HIV
L5_A28	9	Clínica	HIV	7	DENV+HIV
L5_A29	10	Clínica	HIV	1	HIV
L5_A30	11	Clínica	SARS-CoV-2	0	Indetectável
L5_A31	12	Clínica	SARS-CoV-2	11	DENV+SARS-Cov-2

* Classified as virus (virome). Cut-off = 0,0%.

Fonte: Elaborado pela autora

Nas análises de *contigs* por desse classificador taxonômico apresentou-se sensibilidade de 83%, especificidade de 51% e acurácia de 67% nos *cut-offs* de 0%, 0,5% e 1% (tabela 19).

Tabela 19 - Cálculos de sensibilidade, especificidade e acurácia de *contigs* em diferentes *cut-offs* utilizando o classificador taxonômico Genome Detective

Genome Detective

Sumário	Cut-off = 0.0%		Cut-off >= 0,5%		**Cut-off >= 1,0%**		Cut-off >= 5,0%	
Total	35	100%	35	100%	**35**	**100%**	35	100%
Sensibilidade	29	83%	29	83%	**29**	**83%**	28	80%
Especificidade	18	51%	18	51%	**18**	**51%**	18	51%
Acurácia	47	67%	47	67%	**47**	**67%**	46	66%

Fonte: Elaborado pela autora

Para o Genome Detective, adotamos um critério de aceitar sequencias que possuíam ≥10% de cobertura do genoma referência do vírus. Tanto nas análises a nível de *reads*, quanto nas análises a nível de *contigs*, o *workflow* 3 possui os melhores valores de acurácia, especificidade e sensibilidade quando comparado com os dois outros *workflows*.-*

5.4 TEMPO DE DIAGNÓSTICO

Por fim, com todos nossos resultados, foi possível testar diferentes parâmetros de otimização, colocá-los em prática e avaliar comparativamente os resultados obtidos, fornecendo informações cruciais sobre o desempenho das nossas análises, limitações das técnicas e possíveis melhorias. Dentre eles obtivemos uma estimativa de tempo de performance para execução da metodologia proposta por esse trabalho (figura 7).

Figura 7 – Performances
Fonte: Elaborado pela autora

6 **DISCUSSÃO**

Conforme publicado pela Sociedade Brasileira de Patologia Clínica estima-se que 70% das decisões médicas são baseadas em resultados de exames laboratoriais (SBPC, 2021). A identificação de patógenos em laboratórios clínicos depende majoritariamente de técnicas tradicionais de diagnóstico dependentes de cultura, sorologia e diagnóstico molecular baseado em PCR (SIMNER; MILLER; CARROLL, 2018). A "era genômica" trouxe inúmeros avanços ao diagnóstico, entre eles o sequenciamento do DNA, que evoluiu de uma abordagem de baixo rendimento para plataformas altamente tecnológicas e de alto rendimento (MOREIRA, 2021). O sequenciamento metagenômico permitiu uma nova compreensão da diversidade genética das espécies virais presentes dentro e entre os indivíduos (MINOT et al., 2011) e pode ser a peça-chave para diagnósticos mais precisos de doenças infecciosas, para o avanço da medicina de precisão e o tratamento personalizado de pacientes (CHIU; MILLER, 2019).

Inúmeros fatores podem afetar a qualidade e integridade do material genético recuperado de uma amostra, incluindo o manuseio e armazenamento de amostras, extração de ácidos nucléicos, estratégia de sequenciamento e até mesmo o tipo de desenho do estudo (COSTEA et al., 2017). A quantidade de material genético recuperado por extração de ácido nucleico é bastante variável e a depender da amostra biológica, uma grande parte do DNA genômico extraído pode ser o DNA hospedeiro, enquanto o DNA microbiano pode ser responsável por uma pequena fração (FOUHY et al., 2016; COSTEA et al., 2017; KNUDSEN et al., 2016). Essa foi uma das primeiras dificuldades que encontramos ao iniciar o processo de padronização do método, nas primeiras bibliotecas realizadas a quantificação do material genético na biblioteca era muito menor do que o esperado, devido a escolha das amostras e estado de preservação delas (dados não exibidos). Optamos por testar novas amostras com concentração viral superior já conhecidas e confirmadas por algum outro método molecular.

Outro ponto importante é o uso do SISPA, ao invés dos outros primers randômicos que são comumente usados em estudos de metagenômica (FROUSSARD, 1992; DJIKENG et al., 2008; KIRZHNER et al., 2016). No SISPA temos uma amplificação não seletiva de ácido nucléico que envolve a ligação de um iniciador assimétrico em cada extremidade do DNA e após vários ciclos de desnaturação, anelamento e amplificação, quantidades mínimas do DNA inicial são enriquecidas e então são amplificadas (REYES; KIM, 1991). Já os primers randômicos são oligodesoxirribonucleotídeos, não precisam necessariamente de uma preparação da sequência de DNA, como no caso do uso do SISPA, e numericamente gera uma quantidade maior de sequências amplificadas.

Uma dificuldade inerente da amplificação por SISPA, assim como outros métodos que se utilizam de transcrição reversa aleatória e PCR para gerar *amplicons* é a probabilidade de detectar sequencias contaminantes dentre as sequencias de interesse (DJIKENG et al., 2008).

Ao aplicar essa metodologia em uma padronização da avaliação do viroma de amostras clínicas, nós tivemos a presença de diversos contaminantes, sejam eles das próprias amostras, ambientais ou de sequenciamentos anteriores. Tentamos minimizar este problema que enfrentamos com os artefatos de sequenciamento com soluções de bioinformática que incluíram a avaliação do uso de filtros de qualidade e complexidade diversos, além da estratégia arbitrária do uso de *cut-offs.*

Os vieses de metodologia precisam ser integrados à análise de erros pois podem influenciar fortemente a representação de espécies no sequenciamento, como por exemplo, é necessário considerar que em amostras respiratórias, os vírus de RNA são o constituinte mais comum (VANSPAUWEN et al., 2014), entretanto não são os únicos. De acordo com o nosso *pipeline* a etapa de tratamento com DNAse e a estruturação das etapas dessa metodologia impossibilitou o diagnóstico em amostras contendo vírus de DNA, como o citomegalovírus e o vírus causador da hepatite B (dados não exibidos). A baixa qualidade de algumas das sequências, a inexistência de uma abordagem padronizada única e a proporção de falsos positivos pode representar um grande obstáculo para a metagenômica quantitativa (DELMONT; SIMONET; VOGEL, 2013) e todos esses fatores foram observados durante a execução desse estudo.

Além da parte laboratorial, realizamos múltiplas análises de bioinformática que nos permitiu otimizar uma das etapas de maior custo computacional, o *basecalling*. Através dos nossos *benchmarks* foi possível reduzir cerca de 10% do tempo de análise para o modelo *fast* e cerca de 81% para o modelo *hac*. Também podemos observar que a utilização de parâmetros mais liberais, que recuperam um maior número de *reads*, não significa que obtivemos melhores resultados. Comparativamente, os parâmetros qc7_bc2 obtiveram valores de sensibilidade, especificidade e acurácia inferiores aos parâmetros de qualidade recomendados.

O maior desafio da introdução de uma tecnologia, como a proposta por esse estudo, em um laboratório de microbiologia clínica é a análise dos dados. As bibliotecas metagenômica de sequenciamentos de nova geração podem conter vários erros que diminuem a qualidade dos dados e podem atrapalhar a interpretação dos dados (EDWARDS; HOLT, 2013). Dentre os quatro *workflows* testados, consideramos o *workflow 3* o mais de fácil de ser utilizado e que possui melhor acurácia (70%), facilitando a análise de dados e montagem de genomas virais.
O *workflow* 1, criado por nós, possui uma maior complexidade analítica, possuindo dois filtros

de qualidade e complexidade, *softwares* para alinhamento e remoção do genoma humano, além de *softwares* para processamento e correção de erros das *reads* antes da classificação taxonômica com o Kraken2. Devido a um banco de dados maior, o Kraken2 apresentou a vantagem de detectar novos vírus, como o GB-C vírus em coinfecção com o HCV. O *workflow 2* possui menos filtros que o primeiro e se utiliza de um banco de dados de sequências referência local para análise e classificação taxonômica das *reads* utilizando o BLAST. O *workflow* 3 passa pelo Guppy *Basecaller* e Guppy *Barcoder*, depois disso, a plataforma virtual do Genome Detective se encarrega das demais etapas, bem como o *workflow* 4, que se utiliza do Epi2ME para classificação taxonômica.

O classificador taxonômico Kraken2 virtualmente identifica mais vírus que o BLAST (CAMACHO et al., 2009; WOOD; SALZBERG, 2014), essa diferença é ocasionada pelo fato que o BLAST se utiliza de uma base de dados limitada a poucos genomas referência no NCBI. No nosso caso, ainda mais limitado por se tratar de um banco de dados constituído localmente com aproximadamente 1 a 10 sequências referência de cada agente viral de interesse. Idealmente, seria necessário um maior número de sequências referência para que a sensibilidade desse *workflow* seja aumentada, ou a utilização integral de todas as sequências do NCBI.

O uso dos *workflows* 3 e 4, que possuem uma interface mais "amigável" ao usuário inexperiente é um artifício utilizado visando simplificar a análise de dados de bioinformática. Dentre esses, o *workflow 3* desempenhou melhor com uma sensibilidade (74%), especificidade (66%) e acurácia (70%) maiores do que as encontradas pelo *workflow* 4 em seu melhor *cut-off* (1%), quando comparado aos valores de sensibilidade (66%), especificidade (49%) e acurácia (57%) do *workflow* 4 em seu melhor *cut-off* (0,5%).

É intuitivo associar a quantidade de *reads* ao diagnóstico correto da amostra, porém em sequenciamentos metagenômico a quantidade de *reads* específicas necessárias para afirmar com confiança o diagnóstico é bastante variável (FREY et al., 2014). Tendo em vista essa problemática, decidimos avaliar o desempenho do diagnóstico taxonômico a nível de *contigs*. A melhor acurácia alcançada nos dados de treinamento foi de 92% para o Kraken2 e 100% para o BLAST nas análises a nível de *reads*, em contrapartida as melhores acurácias para as análises a nível de *contigs* foram de 65% e 54%, respectivamente. Já nas análises da biblioteca de diluição seriada da vacina anti pólio, houve uma melhoria de acurácia nos *workflows* 1 e 2, subindo de 50% nas análises de *reads*, para 71% nas análises de *contigs*. Entretanto o *workflow*3 saiu de uma acurácia de 50% para 43%.

Ao serem analisadas, as *contigs* obtidas das bibliotecas sequenciadas apresentaram um rendimento geral abaixo do esperado, ou seja, um desempenho de sensibilidade, especificidade

e acurácia inferiores aos obtidos nas análises a nível de *reads.* Este fato pode ser justificado pela diversidade, cobertura recuperada, número e natureza das *reads* encontradas (MACDONALD; PARKS; BEIKO, 2012). A análise de *contigs* neste trabalho se trata de uma forma complementar de checar e validar o resultado encontrado neste estudo. Importante ressaltar que, caso implementado numa rotina clínica, tais análises se tornam caras e demoradas, requerendo um profissional bioinformata que saiba correlacionar tais dados e liberar o diagnóstico encontrado, portanto não seriam necessárias a nível de diagnóstico clínico.

Alguns estudos sugerem a existência de uma "zona cinza" de diagnóstico (AGAPOW et al., 2004; QUEIROZ, 2007; FUNK; OMLAND, 2003; BICKHART et al., 2021), sendo uma área onde não é possível ter total certeza do resultado encontrado. Se esta possibilidade for aplicada a interpretação deste estudo, talvez não comprometessem nossos resultados a nível de *reads*, podendo até mesmo aumentar nossos valores de sensibilidade. Lembrando que não investigamos profundamente a existência e a interferência da zona cinza nesse estudo de validação. Outra estratégia interessante vem sendo aplicada pela Chan Zuckerberg Biohub (CZ Biohub) com o CZ ID, uma plataforma de metagenômica gratuita baseada em nuvem para pesquisadores, as análises na plataforma online CZ ID visam identificar e eliminar os artefatos de sequenciamento, bem como oferecer métricas de qualidade para os sequenciamentos metagenômicos Illumina. Atualmente essa tecnologia não está disponível para sequenciamentos MinION, porém em breve pode ser adaptada e desenvolvida para tal plataforma.

Softwares de bioinformática mais robustos e que possuam uma interface mais simples de usar, como armazenamento em nuvem e análise em tempo real se tornam mais atrativos para uma possível implantação da tecnologia em outros laboratórios (WANG et al., 2021). A reprodutibilidade da metodologia, os custos de implementação clínica, o valor dos custos por teste, o treinamento de pessoal, a padronização das metodologias e de análise de dados, o processo de acreditação diagnóstica e busca de métodos para determinar a relevância clínica com diretrizes de interpretação para os médicos representam os maiores desafios atuais da implementação da metagenômica para diagnóstico de infecções clínicas (FORBES et al., 2018).

O sequenciamento por nanoporos, como o MinION, é uma tecnologia que possui duas vantagens principais: leituras mais longas e a capacidade de executar análises de sequência em tempo real (GRENINGER et al., 2015). O MinION permite a vigilância genômica em campo e em tempo real de doenças infecciosas emergentes (WANG et al., 2021) podendo facilitar significativamente a detecção de vírus (WALTER et al., 2017) e outros patógenos de interesse da saúde pública. O MinION tem sido usado para detectar uma variedade de vírus, incluindo

Ebola (HOENEN et al., 2016; QUICK et al., 2016), dengue (MONGAN et al., 2015), Zika (FARIA et al., 2017; GRUBAUGH et al., 2017; QUICK et al., 2017), influenza (ECKERT et al., 2016; WANG et al., 2015), varíola bovina (KILIANSKI et al., 2015), bem como em estudos de metagenômica ambiental e até mesmo para uso a bordo da Estação Espacial Internacional (CASTRO-WALLACE et al., 2017).

Nosso trabalho propõe algumas alternativas viáveis para análises de bioinformática em um sequenciamento metagenômico usando o MinION e fornece resultados críticos acerca de um *workflow* para sequenciamento metagenômico de vírus. É possível dizer que este estudo consiste na fase 1 de um projeto maior que contemplaria 3 fases, visando a validação dos fluxos de trabalho e metodologias propostas. A fase um consiste em testar apenas amostras de diagnóstico conhecido e confirmado por biologia molecular. A fase dois inclui a testagem de amostras sem diagnóstico e/ou comparação de desempenho com outros métodos de sequenciamento para confirmação do resultado. Por fim, a fase três consiste na coleta e testagem de amostras clínicas em hospitais referências para doenças infectocontagiosas.

Não buscamos substituir a metodologia de cultivo, sorologia e diagnóstico por PCR, mas sim complementá-las a fim de preencher a lacuna de alguns métodos tradicionais de diagnóstico. Inicialmente nosso estudo se concentra na investigação do viroma, entretanto, objetiva-se atingir a análise metagenômica ampla de diversos patógenos, principalmente para investigação de infecções de difícil diagnóstico e vigilância genômica não direcionada.

O MinION e esses fluxos de trabalho propostos neste trabalho ainda não permitem um diagnóstico rápido o suficiente para um serviço laboratorial de diagnósticos. Com a otimização desses fluxos de trabalho e a melhoria da precisão dos testes de diagnóstico na prática clínica, acreditamos que essa tecnologia desempenhará um papel maior no diagnóstico de infecções num futuro próximo. Ainda são necessários mais estudos como o nosso, visando de padronização e validação de técnicas e *pipelines* em quipamentos de sequenciamento diversos para melhorar vários aspectos do NGS metagenômico, a fim de reduzir o tempo de resposta diagnóstica, de preparação da biblioteca e das execuções nas plataformas NGS e, concomitantemente, reduzir ainda mais os custos associados a essas novas tecnologias.

7 **CONCLUSÕES**

Novas intervenções são urgentemente necessárias para auxiliar o diagnóstico de patógenos presentes em casos complexos ou raros e na identificação de agentes emergentes, uma vez que as medidas atuais não têm sido suficientemente eficazes. Esperamos contribuir com o desenvolvimento de testes de diagnóstico avançados que permitirão que os médicos diagnostiquem pacientes com doenças infecciosas na fase inicial da doença e deem suporte à decisão em relação ao início precoce de tratamentos específicos, a fim de prevenir a progressão para formas graves da doença e mortalidade.

Os resultados deste estudo demonstram que a utilização de uma metodologia de NGS metagenômica possibilita o diagnóstico acurado de patógenos virais de importância clínica. O MinION foi capaz de gerar dados das sequencias das amostras para a realização da análise do viroma e diagnóstico de infecções virais com acurácia. Avanços nessa metodologia podem reduzir custos e poderão possibilitar viabilizar sua utilização na rotina de diagnóstico, com a vantagem de permitir a vigilância e descoberta de agentes emergentes.

REFERÊNCIAS

AGAPOW, P. et al. The Impact of Species Concept on Biodiversity Studies. **The Quarterly Review of Biology**, v. 79, n. 2, p. 161–179, 2004.

ALLWOOD, A. C. et al. Stromatolite reef from the Early Archaean era of Australia. **Nature**, v. 441, n. June, p. 714–718, 2006.

ALTERMANN, W.; KAZMIERCZAK, J. Archean microfossils: a reappraisal of early life on Earth. **Research in Microbiology**, v. 154, p. 611–617, 2003.

ALTSCHUL, S. F. et al. Basic Local Alignment Search Tool. **Journal of Molecular Biology**, v. 215, p. 403–410, 1990.

ARAÚJO, A. M. DE. Sp*read*ing the evolutionary synthesis: Theodosius Dobzhansky and genetics in Brazil. **Genetics and Molecular Biology**, v. 27, n. 3, p. 467–475, 2004.

BATOVSKA, J. et al. Metagenomic arbovírus detection using MinION nanopore sequencing. **Journal of Virological Methods**, v. 249, p. 79–84, 2017.

BERNADELI, João. **Infecções virais**: como é feito o diagnóstico delas?. Varsomics, 2022. Disponível em: < https://blog.varsomics.com/diagnostico-de-infeccoes-virais>. Acesso em: 26 de junho de 2022.

BERTELLI, C.; GREUB, G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. **Clinical Microbiology and Infection**, v. 19, n. 9, p. 803–813, 2013.

BICKHART, D. M. et al. **Generation of lineage-resolved complete metagenome-assembled genomes by precision phasing.** bioRxiv, 2021.

BIO-MANGUINHOS. **Poliomielite oral**. Fundação Oswaldo Cruz, 2022. Disponível em: < https://www.bio.fiocruz.br/index.php/br/produtos/vacinas/poliomielite>. Acesso em: 30 de agosto de 2022.

BONECKER, S. **PCR em tempo real**: a metodologia padrão ouro para o diagnóstico. Rio de Janeiro: IPEDMOL, 2020.

BRIESE, T. et al. Genetic detection and characterization of Lujo vírus, a new hemorrhagic fever-associated arenavírus from southern Africa. **PLoS Pathogens**, v. 5, n. 5, p. 1–8, 2009.

BROWN, J. R. et al. Astrovírus VA1/HMO-C: An increasingly recognized neurotropic pathogen in immunocompromised patients. **Clinical Infectious Diseases**, v. 60, n. 6, p. 881–888, 2015.

CAMACHO, C. et al. BLAST +: architecture and applications. **BMC Bioinformatics,** v. 10, n. 412, p. 1–9, 2009.

CASTRO-WALLACE, S. L. et al. Nanopore DNA Sequencing and Genome Assembly on the International Space Station. **Scientific Reports**, v. 7, n. 18022, p. 1–12, 2017.

CHARALAMPOUS, T. et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. **Nature Biotechnology**, v. 37, n. 7, p. 783–792, 2019.

CHARGAFF, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. **Experientia,** v. 6, n. 6, p. 368–376, 1950.

CHENG, J. et al. Identification of pathogens in culture-negative infective endocarditis cases by metagenomic analysis. **Annals of Clinical Microbiology and Antimicrobials,** v. 17, n. 43, p. 1–11, 2018.

CHIU, C. Y.; MILLER, S. A. Clinical metagenomics. **Nature Reviews Genetics**, v. 20, n. 6, p. 341–355, 2019a.

CHIU, C. Y.; MILLER, S. A. **Clinical metagenomics nature reviews genetics**. Nature Publishing Group, , 1 jun. 2019b.

COSTA, E. et al. Severe forms of leptospirosis: clinical, demographic and environmental aspects. **Revista da Sociedade Brasileira de Medicina Tropical,** v. 34, n. 3, p. 261–267, 2001.

COSTEA, P. I. et al. Towards standards for human fecal sample processing in metagenomic studies. **Nature Biotechnology,** v. 35, p. 1069–1076, 2017.

DE JESUS, J. G. et al. Acute Vector-Borne Viral Infection: Zika and MinION Surveillance. **Microbiology Spectrum,** v. 7, n. 4, p. 1–12, 2019.

DELMONT, T. O.; SIMONET, P.; VOGEL, T. M. Mastering methodological pitfalls for surviving the metagenomic jungle. **BioEssays**, v. 35, n. 8, p. 744–754, 2013.

DEURENBERG, R. H. et al. Application of next generation sequencing in clinical microbiology and infection prevention. **Journal of Biotechnology**, v. 243, p. 16–24, 2017.

DJIKENG, A. et al. Viral genome sequencing by random priming methods. **BMC Genomics,** v. 9, n. 5, p. 1–9, 2008.

DUERKOP, B. A.; HOOPER, L. V. Resident víruses and their interactions with the immune system. **Nature Immunology**, v. 14, n. 7, p. 654–659, 2013.

ECKERT, S. E. et al. Enrichment by hybridisation of long DNA fragments for Nanopore sequencing. **Microbial Genomics**, v. 2, n. 9, p. 1–10, 2016.

EDWARDS, D. J.; HOLT, K. E. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. **Microbial Informatics and Experimentation,** v. 3, n. 2, p. 1–9, 2013.

FARIA, N. R. et al. Establishment and cryptic transmission of Zika vírus in Brazil and the Americas. **Nature,** v. 546, n. 7658, p. 406–410, 2017.

FAUCI, A. S.; MORENS, D. M. The Perpetual Challenge of Infectious Diseases. **New England Journal of Medicine,** v. 366, n. 5, p. 454–461, 2012.

FEDURCO, M. et al. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. **Nucleic Acids Research**, v. 34, n. 3, 2006.

FENOLLAR, F.; RAOULT, D. Molecular diagnosis of bloodstream infections caused by non-cultivable bacteria. **International Journal of Antimicrobial Agents**, v. 30S, p. S7–S15, 2007.

FORBES, J. D. et al. Metagenomics: The Next Culture-Independent Game Changer. **Frontiers in Microbiology**, v. 8, n. July, p. 1–21, 2017.

FORBES, J. D. et al. **Highlighting Clinical metagenomics for enhanced diagnostic decision-making**: a step towards wider implementationcomputational and structural biotechnology. Journal Elsevier B.V., , 1 jan. 2018.

FOUCAULT, Michel. **Microfísica do poder.** 10. ed. Rio de Janeiro: Graal, 1992.

FOUHY, F. et al. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. **BMC Microbiology**, v. 16, n. 123, p. 1–13, 2016.

FREY, K. G. et al. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. **BMC Genomics,** v. 15, n. 96, p. 1–14, 2014.

FROUSSARD, P. A random-POR method (rPCR) to construct whole cDNA library from low amounts of RNA. **Nucleic Acids Research,** v. 20, n. 11, p. 2900, 1992.

FUNK, D. J.; OMLAND, K. E. Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. **Annual Review of Ecology, Evolution, and Systematics**, v. 34, p. 397–423, 2003.

GARGIS, A. S.; KALMAN, L.; LUBIN, M. Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. **Journal of Clinical Microbiology,** v. 54, n. 12, p. 2857–2865, 2016.

GIRE, S. K. et al. Genomic surveillance elucidates Ebola vírus origin and transmission during the 2014 outbreak. **Science,** v. 345, n. 6002, p. 1369–1372, 2014.

GLASER, C. A. et al. Beyond Víruses: Clinical Profiles and Etiologies Associated with Encephalitis. **Clinical Infectious Diseases**, v. 43, n. 12, p. 1565–1577, 2006.

GLASER, C.; BLOCH, K. C. Encephalitis: Why We Need to Keep Pushing the Envelope. **Clinical Infectious Diseases**, v. 49, n. 12, p. 1848–1850, 2009.

GÓES, A. C. DE S.; OLIVEIRA, B. V. X. DE. Projeto Genoma Humano: um retrato da construção do conhecimento científico sob a ótica da revista Ciência Hoje. **Ciência & Educação,** Bauru, v. 20, n. 3, p. 561–577, 2014.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. **Nature,** v. 17, n. 6, p. 333–351, 2016.

GOODWIN, S.; WAPPEL, R.; MCCOMBIE, W. R. 1D Genome Sequencing on the Oxford Nanopore MinION. **Current Protocols in Human Genetics**, v. 94, n. 7, p. 18.11.1-18.11.14, 2017.

GORRIE, C. L. et al. Antimicrobial-Resistant Klebsiella pneumoniae Carriage and Infection in Specialized Geriatric Care Wards Linked to Acquisition in the Referring Hospital. **Clinical Infectious Diseases,** v. 67, n. 2, p. 161–170, 2018.

GRENINGER, A. L. et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. **Genome Medicine,** v. 7, n. 1, p. 1–13, 2015.

GRUBAUGH, N. D. et al. Genomic epidemiology reveals multiple introductions of Zika vírus into the United States. **Nature**, v. 546, p. 401–405, 2017.

GUO, J. et al. Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. **Proceedings of the National Academy of Sciences**, v. 105, n. 27, p. 9145–9150, 2008.

GWINN, M.; MACCANNELL, D.; ARMSTRONG, G. L. Next-Generation Sequencing of Infectious Pathogens. **JAMA Insights**, v. 321, n. 9, p. 893–894, 2019.

HAN, D. et al. mNGS in clinical microbiology laboratories: on the road to maturity. **Critical Reviews in Microbiology**, v. 45, n. 5–6, p. 668–685, 2019.

HEATHER, J. M.; CHAIN, B. The sequence of sequencers: The history of sequencing DNA. **Genomics**, v. 107, n. 1, p. 1–8, 2015.

HOENEN, T. et al. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. **Emerging Infectious Diseases**, v. 22, n. 2, p. 331–334, 2016.

HOULDCROFT, C. J.; BEALE, M. A.; BREUER, J. Clinical and biological insights from viral genome sequencing. **Nature Reviews Microbiology**, v. 15, n. 3, p. 183–192, 2017.

HSIUNG, G. D. Diagnostic Virology: From Animals to Automation. **The Yale Journal of Biology and Medicine**, v. 57, p. 727–733, 1984.

IP, C. L. C. et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. **F1000Research**, v. 4, n. 1075, p. 1–35, 2015.

JU, J. et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. **Proceedings of the National Academy of Sciences**, v. 103, n. 52, p. 19635–19640, 2006.

KANZI, A. M. et al. Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. **Frontiers in Genetics**, v. 11, n. 10, p. 1–18, 2020.

KAZ, R. **Tempos da peste**: uma biografia improvável. Piauí. 1641 ed., [s.n], 2020. Disponível em: <https://piaui.folha.uol.com.br/materia/uma-biografia-improvavel/>. Acesso em: 30 de maio de 2022

KILIANSKI, A. et al. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. **GigaScience**, v. 4, n. 12, p. 1–8, 2015.

KIRZHNER, V. et al. Analysis of Metagenome Composition by the Method of Random Primers. **arXiv**, p. 1–18, 2016.

KNUDSEN, B. E. et al. Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. **ASM Journals**, v. 1, n. 5, p. 1–15, 2016.

KO, F. W. et al. Molecular detection of respiratory pathogens and typing of human rhinovírus of adults hospitalized for exacerbation of asthma and chronic obstructive pulmonary disease. **Respiratory Research,** v. 20, n. 210, p. 1–9, 2019.

KONO, N.; ARAKAWA, K. Nanopore sequencing: Review of potential applications in functional genomics. **Development, Growth & Differentiation**, v. 61, n. 3, p. 316–326, 2019.

KORSMAN, Stephen N J. **Virologia**. Rio de Janeiro: Grupo GEN, 2014. 9788595151871. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788595151871/>. Acesso em: 30 de maio de 2022

KUBISTA, M. I See the Light! And I See It Again and Again! **Clinical Chemistry**, v. 58, n. 11, p. 1505–1506, 2012.

LE GOFF, Jacques. Uma história dramática. In: LE GOFF, Jacques. (org.). A**s doenças têm história**. Lisboa: Terramar, 1991.

LEFTEROVA, M. I. et al. Next-Generation Sequencing for Infectious Disease Diagnosis and Management: A Report of the Association for Molecular Pathology. **Journal of Molecular Diagnostics**, v. 17, n. 6, p. 623–634, 2015.

LEGER, A.; LEONARDI, T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. **Journal of Open Source Software**, v. 4, n. 34, p. 1–4, 2019.

LELAND, D. S.; GINOCCHIO, C. C. Role of Cell Culture for Vírus Detection in the Age of Technology. **Clinical Microbiology Reviews**, v. 20, n. 1, p. 49–78, 2007.

LI, H. Minimap2: Pairwise alignment for nucleotide sequences. **Bioinformatics,** v. 34, n. 18, p. 3094–3100, 2018.

LIN, B.; HUI, J.; MAO, H. Nanopore Technology and Its Applications in Gene Sequencing. **Biosensors,** v. 214, n. 11, p. 1–17, 2021.

LIU, L. et al. Comparison of next-generation sequencing systems. **Journal of Biomedicine and Biotechnology**, p. 1–25, 2012.

LOCEY, K. J.; LENNON, J. T. Scaling laws predict global microbial diversity. **Proceedings of the National Academy of Sciences**, v. 113, n. 21, p. 5970–5975, 2016.

LU, H.; GIORDANO, F.; NING, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. **Genomics, Proteomics and Bioinformatics**, v. 14, n. 5, p. 265–279, 2016.

MACDONALD, N. J.; PARKS, D. H.; BEIKO, R. G. Rapid identification of high-confidence taxonomic assignments for metagenomic data. **Nucleic Acids Research**, v. 40, n. 14, p. 1–13, 2012.

MACFARLAN, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovírus activity. **Nature,** v. 486, p. 57–63, 2012.

MARDIS, E. R. The impact of next-generation sequencing technology on genetics. Cell Press Trends in Genetics, v. 24, n. 3, p. 133–141, 2008.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing *reads*. **Advances in Environmental Biology,** v. 7, n. 10, p. 2803–2809, 2011.

MEDINI, D. et al. Microbiology in the post-genomic era. **Nature Reviews Microbiology**, v. 6, n. 6, p. 419–430, 2011.

MIAO, Q. et al. Microbiological Diagnostic Performance of Metagenomic Next-generation Sequencing When Applied to Clinical Practice. **Clinical Infectious Diseases**, v. 67, n. Suppl 2, p. S231–S240, 2018.

MIKHEYEV, A.; TIN, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. **Molecular Ecology**, v. 14, n. 6, p. 1097–1102, 2014.

MILLER, M. B.; TANG, Y. W. Basic concepts of microarrays and potential applications in clinical microbiology. **Clinical Microbiology Reviews**, v. 22, n. 4, p. 611–633, 2009.

MILLER, R. R. et al. Metagenomics for pathogen detection in public health. **Genome Medicine**, v. 5, n. 81, p. 1–14, 2013.

MINOT, S. et al. The human gut virome: Inter-individual variation and dynamic response to diet. **Genome Research,** v. 21, p. 1616–1625, 2011.

MITCHELL, A. B.; GLANVILLE, A. R. Introduction to techniques and methodologies for characterizing the human respiratory virome. **Methods in Molecular Biology,** v. 1838, p. 111–123, 2018.

MITCHELL, A. B.; OLIVER, B. G. G.; GLANVILLE, A. R. Translational aspects of the human respiratory virome. **American Journal of Respiratory and Critical Care Medicine**, v. 194, n. 12, p. 1458–1464, 2016.

MONGAN, A. E. et al. The Evaluation on Molecular Techniques of Reverse Transcription Loop-Mediated Isothermal Amplification (RT-LAMP), Reverse Transcription Polymerase Chain Reaction (RT-PCR), and Their Diagnostic Results on MinION Nanopore Sequencer for the Detection of Den. **American Journal of Microbiological Research**, v. 3, n. 3, p. 118–124, 2015.

MOON, J. et al. International Journal of Medical Microbiology Rapid diagnosis of bacterial meningitis by nanopore 16S amplicon sequencing : A pilot study. **International Journal of Medical Microbiology,** v. 309, n. 6, p. 1–7, 2019.

MOORE, J. A. Science as a way of knowin genetics. **American Zoologist**, v. 26, p. 583–747, 1986.

MOREIRA, L. M. **Ciências genômicas**: fundamentos e aplicações. Ribeirão Preto, SP: Sociedade Brasileira de Genética, 2015.

MOREY, M. et al. A glimpse into past, present, and future DNA sequencing. **Molecular Genetics and Metabolism**, v. 110, n. 1–2, p. 3–24, 2013.

MULLIS, K. et al. Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 51, p. 263–273, 1986.

MULLIS, K. B. The Unusual Origin of the Polymerase Chain Reaction. **Scientific American**, v. April, p. 56–65, 1990.

NACCACHE, S. N. et al. Diagnosis of neuroinvasive astrovírus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. **Clinical Infectious Diseases**, v. 60, n. 6, p. 919–923, 2015.

NOWROUSIAN, M. Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems. **Eukaryotic Cell**, v. 9, n. 9, p. 1300–1310, 2010.

OSTERHOLM, M. T.; OLSHAKER, M. I**nimigo mortal**: nossa guerra contra os germes assassinos. Rio de Janeiro: Intrinseca, 2020.

OXFORD NANOPORE TECHNOLOGIES. **Introduction to real time analysis**. 2020. Disponível em: <https://www.youtube.com/watch? v=8oNEjt5Ov_Q>. Acesso em: 30 de maio de 2022

OXFORD NANOPORE TECHNOLOGIES. **MinION IT requirements**: version: 1.0.0. p. 1–6, [s.d.].

PALACIOS, G. et al. A new arenavírus in a cluster of fatal transplant-associated diseases. **The New England Journal of Medicine,** v. 358, n. 10, p. 991–998, 2008.

PALLEN, M. J. Diagnostic metagenomics: potential applications to bacterial , viral and parasitic infections. Cambridge University Press. **Parasitology,** v. 141, n. 14, p. 1856–1862, 2014.

PARKER, J.; CHEN, J. Application of next generation sequencing for the detection of human viral pathogens in clinical specimens. **Journal of Clinical Virology**, v. 86, p. 20–26, 2018.

PRAY, B. L. A. **Discovery of DNA structure and function**: watson and crick the landmark ideas of watson and crick relied heavily on the work of other scientists. Levene Inves. p. 1–6, [s.d.].

QUARESMA, P. S. A. As doenças e a história do homem: um itinerário em comum. ANAIS DO XXVI SIMPÓSIO NACIONAL DE HISTÓRIA ANPUH. **Anais**... São Paulo: 2011.

QUEIROZ, K. DE. Species Concepts and Species Delimitation. **Systematic Biology,** v. 56, n. 6, p. 879–886, 2007.

QUICK, J. et al. Real-time, portable genome sequencing for Ebola surveillance. **Nature,** v. 530, n. 7589, p. 228–232, 2016.

QUICK, J. et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other vírus genomes directly from clinical samples. Nature Protocols, v. 12, n. 6, p. 1261–1266, 2017.

RANG, F. J.; KLOOSTERMAN, W. P.; RIDDER, J. DE. From squiggle to basepair: computational approaches for improving nanopore sequencing *read* accuracy. **Genome Biology**, v. 19, n. 90, p. 1–11, 2018.

REYES, G. R.; KIM, J. P. Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. **Molecular and Cellular Probes**, v. 5, p. 473–481, 1991.

RIBEIRO, G. et al. Frequent House Invasion of Trypanosoma cruzi-Infected Triatomines in a Suburban Area of Brazil. **PLoS Neglected Tropical Diseases**, v. 9, n. 4, p. 1–11, 2015.

RONDON, M. R. et al. Cloning the soil metagenome: A strategy for accessing the genetic and functional diversity of uncultured microorganisms. **Applied and Environmental Microbiology, v. 66, n. 6, p. 2541–2547, 2000.**

RUNTUWENE, L. R.; TUDA, J. S. B.; MONGAN, A. E. **On-site minion sequencing.** ed. 1129. Singapore: Springer, 2019. p. 143–150.

RUPPÉ, E.; SCHRENZEL, J. Messages from the second International Conference on Clinical Metagenomics (ICCMg2). **Microbes and Infection**, v. 20, n. 4, p. 222–227, 2018.

SAIKI, R. K. et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. **Science,** v. 230, n. 12, p. 1350–1354, 1985.

SALTER, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. **BMC Biology**, v. 12, n. 87, p. 1–12, 2014.

SANDERSON, N. D. et al. Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. **BMC Genomics**, v. 19, n. 714, p. 1–11, 2018.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America,** v. 74, n. 12, p. 5463–5467, 1977.

SANTOS, E. A. et al. I**nfluência da temperatura ambiente na análise do termociclador**. 24 Congresso Brasileiro de Engenharia Biomédica, p. 1522–1525, 2014

SANTOS, N. S. de O. et al. **Virologia humana.** Rio de Janeiro: Grupo GEN, 2021. 9788527738354. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788527738354/>. Acesso em: 30 de maio de 2022

SIMNER, P. J.; MILLER, S.; CARROLL, K. C. Understanding the promises and hurdles of metagenomic next-generation sequencing as a diagnostic tool for infectious diseases. **Clinical Infectious Diseases**, v. 66, n. 3, p. 778–788, 2018.

SIMPSON, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. **Nature Methods**, v. 14, n. 1, p. 407–410, 2017.

SOCIEDADE BRASILEIRA DE PATOLOGIA CLÍNICA-MEDICINA LABORATORIAL. **Lista de laboratórios acreditados no PALC**. 2021. Disponível em: <http://www.sbpc.org.br/programa-da-qualidade/laboratorios-acreditados/>. Acesso em: 29 de maio de 2022.

SÖDING, J. Protein homology detection by HMM – HMM comparison. bioinformatics, v. 21, n. 7, p. 951–960, 2005.

STREIT, W. R.; SCHMITZ, R. A. Metagenomics: the key to the uncultured microbes. **Current Opinion in Microbiology,** v. 7, p. 492–498, 2004.

STRONG, M. J. et al. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. **Plos Pathogens,** v. 10, n. 11, p. 1–6, 2014.

TEIXEIRA, R. DOS S. Reflexões sobre a origem e a evolução das doenças infecciosas e parasitárias no estado da bahia. **Gazeta Médica da Bahia**, v. 2, n. 77, p. 158–181, 2007.

THOMAS, M. K. et al. Estimates of foodborne illness–related hospitalizations and deaths in canada for 30 specified pathogens and unspecified agents. F**oodborne Pathogens and Disease**, v. 12, n. 10, p. 820–827, 2015.

TURCATTI, G. et al. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for dna sequencing by synthesis. N**ucleic Acids Research**, v. 36, n. 4, p. 1–13, 2008.

UJVARI, S. C. A **história da humanidade contada pelos vírus**: bactérias, parasitas e outros microrganismos. São Paulo: Contexto, 2012.

VAN GAGELDONK-LAFEBER, A. B. VAN et al. A case-control study of acute respiratory tract infection in general practice patients in the netherlands. **Clinical Infectious Diseases**, v. 41, n. 4, p. 490–497, 2005.

VANSPAUWEN, M. J. et al. Comparison of three different techniques for the isolation of viral RNA in sputum. **Journal of Clinical Virology,** v. 61, p. 265–269, 2014.

VASER, R. et al. *Fast* and accurate de novo genome assembly from long uncorrected *reads*. **Genome Research**, v. 27, p. 737–746, 2017.

VILSKER, M. et al. Genome Detective: An automated system for vírus identification from high-throughput sequencing data. **Bioinformatics**, v. 35, n. 5, p. 871–873, 2019.

VIRGIN, H. W.; WHERRY, E. J.; AHMED, R. Redefining chronic viral infection. **Cell**, v. 138, n. 1, p. 30–50, 2009.

VOELKERDING, K. V.; DAMES, S.; DURTSCHI, J. D. Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: a paper from the 2009 william beaumont hospital symposium on molecular pathology. **Journal of Molecular Diagnostics,** v. 12, n. 5, p. 539–551, 2010.

WALTER, M. C. et al. MinION as part of a biomedical rapidly deployable laboratory. **Journal of Biotechnology**, v. 250, n. 5, p. 16–22, 2017.

WANG, J. et al. MinION nanopore sequencing of an influenza genome. **Frontiers in Microbiology**, v. 6, p. 1–7, 2015.

WANG, Y. et al. Nanopore sequencing technology, bioinformatics and applications. **Nature Biotechnology,** v. 39, n. 11, p. 1348–1365, 2021.

WARD, N.; FRASER, C. M. How genomics has affected the concept of microbiology. **Current Opinion in Microbiology**, v. 8, n. 5, p. 564–571, 2005.

WATSON, J.; CRICK, F. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. **Nature**, v. 171, n. 4356, p. 737–738, 1953.

WELLE, Deutsche. Calendário Histórico 1931: físico Alemão cria o microscópio eletrônico. **Folha de São Paulo**, 2005. Disponível em: < https://www1.folha.uol.com.br/folha/dw/ult1908u2254.shtmlL>. Acesso em: 27 de junho de 2022.

WHITMAN, W. B.; COLEMAN, D. C.; WIEBE, W. J. Perspective Prokaryote : The unseen majority. **Proceedings of the National Academy of Sciences of the United States of America**, v. 95, n. 6, p. 6578–6583, 1998.

WILLNER, D. et al. Case Studies of the Spatial Heterogeneity of DNA Víruses in the Cystic Fibrosis Lung. **American Journal of Respiratory Cell and Molecular Biology**, v. 46, p. 127–131, 2012.

WILSON, M. R. et al. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. **New England Journal of Medicine**, v. 370, n. 25, p. 2408–2417, 2014.

WOOD, D. E.; LU, J.; LANGMEAD, B. Improved metagenomic analysis with Kraken 2. **Genome Biology,** v. 20, n. 257, p. 1–13, 2019.

WOOD, D. E.; SALZBERG, S. L. Kraken: ultra*fast* metagenomic sequence classification using exact alignments. **Genome Biology**, v. 15, n. R46, p. 1–12, 2014.

WORLD HEALTH ORGANIZATION. **Constituição da organização mundial de saúde.** Geneva. : [s.n.], [s.d.].
Disponível em: <http://www.who.int/governance/eb/who_constitution_sp.pdf >.

WORLD HEALTH ORGANIZATION. **Working to overcome the global impact of neglected tropical diseases**. first who report on neglected tropical diseases. Geneva: [s.n.], [s.d.].

WU, T. D.; WATANABE, C. K. Sequence analysis GMAP: a genomic mapping and alignment program for mRNA and EST sequences. **Bioinformatics**, v. 21, n. 9, p. 1859–1875, 2005.

YANG, J. et al. Unbiased Parallel Detection of Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach. **Journal of Clinical Microbiology**, v. 49, n. 10, p. 3463–3469, 2011.

ZAHA, A.; FERREIRA, H. B.; PASSAGLIA, L. M. P. (org.). **Biologia molecular básica.** 4. ed. [s.l.]: ArtMed, 2012.

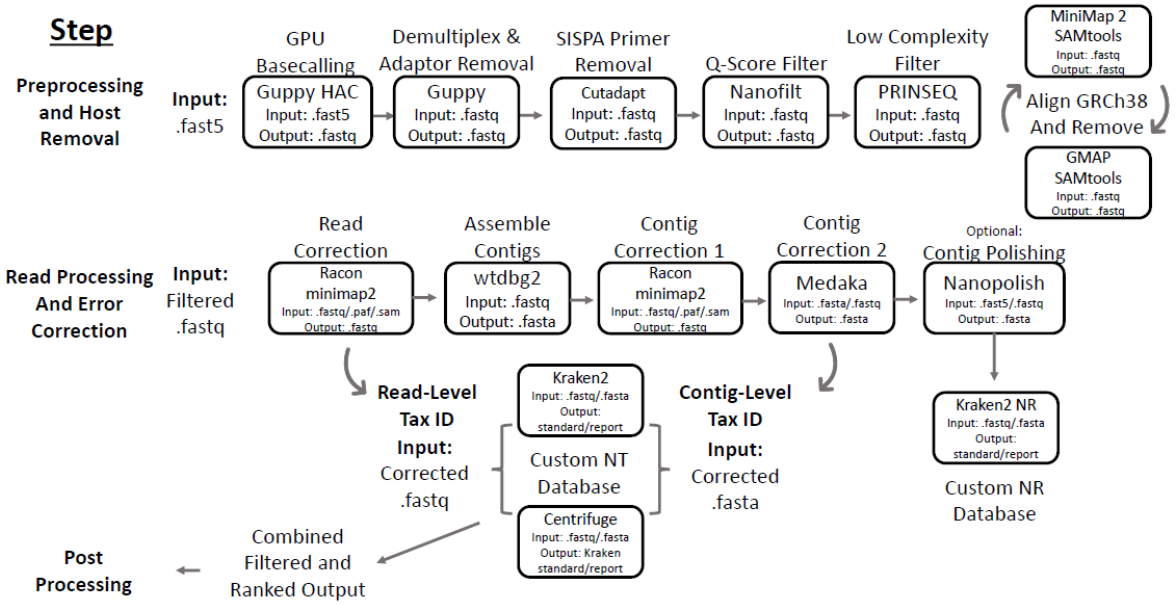
ZEPEDA MENDOZA, M. L.; SICHERITZ-PONTÉN, T.; THOMAS GILBERT, M. P. Environmental genes and genomes: Understanding the differences and challenges in the approaches and *software* for their analyses. **Briefings in Bioinformatics**, v. 16, n. 5, p. 745–758, 2015.

ZLOJUTRO, A.; REY, D.; GARDNER, L. A decision-support framework to optimize border control for global outbreak mitigation. **Scientific Reports**, v. 9, n. 2216, p. 1–14, 2019.

Apêndice A - Checagem das *flowcells* e controle de corrida.

Num. das *flowcells*	Código	Num. de poros ativos	Tempo de uso da *flowcell* (h)	Total de *reads*
1	FAN01632	1.315	10,00	700
2	FAM97448	1.743	48,00	645282
3	FAM97324	1.413	10,00	23000
4	FAM96095	1.588	1,00	135
5	FAM97841	1.212	1,30	365
6	FAM95086	1.418	24,00	2000
7	FAM97824	1.498	22,00	2100
8	FAM97777	1.100	1,30	300840
9	FAM94927	1.393	0,13	35
10	FAM94813	1.314	0,20	0
11	FAM96139	1.415	0,13	3
12	FAM95140	1.200	0,22	1200
13	FAM94926	1.496	24,00	5520000
14	FAN03236	1.329	24,00	50000
15	FAM97860	1531	38,00	1800000
16	FAM97720	1.654	37,00	55420
17	FAM97389	1.672	40,00	4900000
18	FAM97443	1.547	43,00	2500000
19	FAM94928	1.519	16,00	392000
20	FAM94815	716	24,00	257000
21	FAM97387	1.685	23,00	1200000
22	FAN01632	684	25,00	679000

Apêndice B – *Workflow* base



Anexo A: Arquivos de configuração do Guppy *Basecaller* dna_r9.4.1_450bps_*fast*.cfg

Basic configuration file for ONT Guppy *Basecaller software*.

# Data trimming.
trim_strategy = dna
trim_threshold = 2.5
trim_min_events = 3

# *Basecalling*.
model_file = template_r9.4.1_450bps_*fast*.jsn
chunk_size = 2000
gpu_runners_per_device = 8
chunks_per_runner = 160
chunks_per_caller = 10000
overlap = 50
qscore_offset = -0.8012
qscore_scale = 0.8476
builtin_scripts = 1
beam_width = 32
noisiest_section_scaling_max_size = 8000

# Calibration strand detection
calib_reference = lambda_3. 6kb.fasta
calib_min_sequence_length = 3000
calib_max_sequence_length = 3800
calib_min_coverage = 0.6

Output.
records_per_*fast*q = 4000
min_qscore = 8.0

# Telemetry
ping_url = https://ping.oxfordnanoportal.com/basecall
ping_segment_duration = 60

dna_r9.4.1_450bps_*hac*.cfg

Basic configuration file for ONT Guppy *Basecaller software*.

# Compatibility
compatible_flowcells = FLO-FLG001, FLO-FLGOP1, FLO-MIN106
compatible_kits = SQK-CAS109, SQK-CS9109, SQK-DCS108, SQK-DCS109, SQK-LRK001,SQK-LSK108,SQK-LSK109,SQK-LSK109-XL,SQK-LSK110,SQK-LSK110-XL,SQK-LWP001,SQK-PCS108,SQK-PCS109,SQK-PCS110,SQK-PRC109,SQK-PSK004,SQK-RAD002,SQK-RAD003,SQK-RAD004,SQK-RAS201,SQK-RLI001,SQK-ULK001,VSK-VBK001,VSK-VSK001,VSK-VSK002,VSK-VSK003compatible_kits_with_barcoding =OND-SQK-LP0096M,OND-SQK-LP0096MA,OND-SQK-LP0096S,OND-SQK-LP0768L,OND-SQK-LP1152S,OND-SQK-LP9216,OND-SQK-RP0096M,OND-SQK-RP0096MA,OND-SQK-RP0384L,SQK-16S024,SQK-NBD110-24,SQK-NBD110-96,SQK-PCB109,SQK-PCB110,SQK-RBK001,SQK-RBK004,SQK-RBK110-96,SQK-RLB001,SQK-LWB001,SQK-PBK004,SQK-RAB201,SQK-RAB204,SQK-RPB004,VSK-PTC001,VSK-VMK001,VSK-VMK002,VSK-VMK003

Data trimming.

trim_strategy = dna
trim_threshold = 2.5
trim_min_events = 3

# *Basecalling*.
model_file = template_r9.4.1_450bps_*hac*.jsn
chunk_size = 2000
gpu_runners_per_device = 4
chunks_per_runner = 256
chunks_per_caller = 10000
overlap = 50
qscore_offset = -0.1721
qscore_scale = 0.9356

dna_r9.4.1_450bps_sup.cfg

Basic configuration file for ONT Guppy *Basecaller software*.

# Data trimming.
trim_strategy = dna
trim_threshold = 2.5
trim_min_events = 3

# *Basecalling*.
model_file = template_r9.4.1_450bps_sup.jsn
chunk_size = 1000
gpu_runners_per_device = 12
chunks_per_runner = 256
chunks_per_caller = 10000
overlap = 100
qscore_offset = 0.3498
qscore_scale = 0.9722
builtin_scripts = 1
beam_cut = 100.0
beam_width = 32
noisiest_section_scaling_max_size = 8000

# Calibration strand detection
calib_reference = lambda_3. 6kb.fasta
calib_min_sequence_length = 3000
calib_max_sequence_length = 3800
calib_min_coverage = 0.6

Output.
records_per_*fast*q = 4000
min_qscore = 10.0

# Telemetry
ping_url = https://ping.oxfordnanoportal.com/basecall
ping_segment_duration = 60

builtin_scripts = 1
beam_width = 32
noisiest_section_scaling_max_size = 8000

# Calibration strand detection
calib_reference = lambda_3. 6kb.fasta
calib_min_sequence_length = 3000
calib_max_sequence_length = 3800
calib_min_coverage = 0.6

Output.
records_per_*fast*q = 4000
min_qscore = 9.0

# Telemetry
ping_url = https://ping.oxfordnanoportal.com/basecall
ping_segment_duration = 60

Anexo B: Lista de sequências referência

N°	Família	ACC	TAX_ID	TAX_NAME
1	Caliciviridae	NC_029645	340017	Norovirus GIII
2	Caliciviridae	NC_039475	552592	Norovirus GII.17
3	Caliciviridae	NC_039476	490039	Norovirus GII.2
4	Caliciviridae	NC_039477	122929	Norovirus GII
5	Caliciviridae	NC_040876	122929	Norovirus GII
6	Caliciviridae	NC_044045	1529918	Norovirus GII
7	Caliciviridae	NC_044046	1529924	Norovirus GII
8	Caliciviridae	NC_044932	122929	Norovirus GII
9	Caliciviridae	NC_001959	122928	Norovirus GI
10	Caliciviridae	NC_039897	1529909	Norovirus GI
11	Caliciviridae	NC_044853	122928	Norovirus GI
12	Caliciviridae	NC_044854	122928	Norovirus GI
13	Caliciviridae	NC_044856	122928	Norovirus GI
14	Caliciviridae	NC_029647	262897	Norovirus GIV
15	Caliciviridae	NC_044855	262897	Norovirus GIV
16	Caliciviridae	NC_045762	262897	Norovirus GIV
17	Caliciviridae	NC_008311	1246677	Norovirus GV
18	Coronaviridae	NC_045512	2697049	Severe Acute Respiratory Syndrome coronavirus 2
19	Flaviviridae	FJ850094	11069	Dengue virus 3
20	Flaviviridae	GU131872	11069	Dengue virus 3
21	Flaviviridae	GU131873	11069	Dengue virus 3
22	Flaviviridae	GU131874	11069	Dengue virus 3
23	Flaviviridae	GU131875	11069	Dengue virus 3
24	Flaviviridae	GU131876	11069	Dengue virus 3
25	Flaviviridae	GU131877	11069	Dengue virus 3
26	Flaviviridae	GU131878	11069	Dengue virus 3
27	Flaviviridae	JF808120	11069	Dengue virus 3
28	Flaviviridae	JF808121	11069	Dengue virus 3
29	Flaviviridae	JX286519	11060	Dengue virus 2
30	Flaviviridae	JX286521	11060	Dengue virus 2
31	Flaviviridae	KP188541	11053	Dengue virus 1
32	Flaviviridae	KP188542	11053	Dengue virus 1
33	Flaviviridae	KP188543	11053	Dengue virus 1
34	Flaviviridae	KP188544	11053	Dengue virus 1
35	Flaviviridae	KP188545	11053	Dengue virus 1
36	Flaviviridae	KP188546	11053	Dengue virus 1
37	Flaviviridae	KP188547	11053	Dengue virus 1
38	Flaviviridae	KP188548	11053	Dengue virus 1
39	Flaviviridae	KP188551	11060	Dengue virus 2
40	Flaviviridae	KP188552	11060	Dengue virus 2

41	Flaviviridae	KP188553	11060	Dengue virus 2
42	Flaviviridae	KP188554	11060	Dengue virus 2
43	Flaviviridae	KP188555	11060	Dengue virus 2
44	Flaviviridae	KP188556	11060	Dengue virus 2
45	Flaviviridae	KP188558	11070	Dengue virus 4
46	Flaviviridae	KP188559	11070	Dengue virus 4
47	Flaviviridae	KP188560	11070	Dengue virus 4
48	Flaviviridae	KP188561	11070	Dengue virus 4
49	Flaviviridae	KP188562	11070	Dengue virus 4
50	Flaviviridae	KP188563	11070	Dengue virus 4
51	Flaviviridae	KP188564	11070	Dengue virus 4
52	Flaviviridae	KP188565	11070	Dengue virus 4
53	Flaviviridae	KP188566	11070	Dengue virus 4
54	Flaviviridae	KP188567	11053	Dengue virus 1
55	Flaviviridae	KP188568	11053	Dengue virus 1
56	Flaviviridae	KP188569	11060	Dengue virus 2
57	Flaviviridae	KU513441	11070	Dengue virus 4
58	Flaviviridae	MN239505	11060	Dengue virus 2
59	Flaviviridae	NC_001437	11072	Japanese encephalitis virus
60	Flaviviridae	NC_001474	11060	Dengue virus 2
61	Flaviviridae	NC_001475	11069	Dengue virus 3
62	Flaviviridae	NC_001477	11053	Dengue virus 1
63	Flaviviridae	NC_001563	11082	West Nile virus
64	Flaviviridae	NC_002031	11089	Yellow Fever virus
65	Flaviviridae	NC_002640	11070	Dengue virus 4
66	Flaviviridae	NC_007580	11080	Saint Louis Encephalitis virus
67	Flaviviridae	NC_009028	59563	Ilheus virus
68	Flaviviridae	NC_009942	11082	West Nile virus
69	Flaviviridae	NC_012532	64320	Zika virus
70	Flaviviridae	NC_035889	64320	Zika virus
71	Flaviviridae	NC_040776	64315	Rocio virus
72	Hepeviridae	AF082843	2848127	Swine hepatitis E virus
73	Hepeviridae	EF077630	1678143	Orthohepevirus A
74	Hepeviridae	M73218	1678143	Orthohepevirus A
75	Hepeviridae	M74506	2773468	Human hepatitis E virus genotype 2a
76	Herpesviridae	NC_001348	10335	Human alphaherpesvirus 3
77	Herpesviridae	NC_001798	10310	Human alphaherpesvirus 2
78	Herpesviridae	NC_001806	10298	Human alphaherpesvirus 1
79	Paramyxoviridae	NC_002200	2560602	Mumps orthorubulavirus
80	Picornaviridae	KF537633	12080	Human poliovirus 1
81	Picornaviridae	KU763188	12086	Human poliovirus 3
82	Picornaviridae	AY184219	12080	Human poliovirus 1
83	Picornaviridae	AY184220	12083	Human poliovirus 2
84	Picornaviridae	AY184221	12086	Human poliovirus 3
85	Picornaviridae	MH484164	1295563	Enterovirus C99

86	Picornaviridae	MH484165	325445	Enterovirus C96
87	Picornaviridae	MH484166	1295563	Enterovirus C99
88	Picornaviridae	MK069966	200154	Enterovirus B73
89	Picornaviridae	MK689070	47512	Echovirus E29
90	Picornaviridae	MK689071	1295563	Enterovirus C99
91	Picornaviridae	NC_001430	138951	Enterovirus D
92	Picornaviridae	NC_001612	138948	Enterovirus A
93	Picornaviridae	NC_001859	12064	Enterovirus E
94	Picornaviridae	NC_002058	138950	Enterovirus C
95	Picornaviridae	NC_003988	310907	Enterovirus H
96	Picornaviridae	NC_004441	64141	Porcine enterovirus 9
97	Picornaviridae	NC_021220	1330520	Enterovirus F
98	Picornaviridae	NC_030454	2760809	Enterovirus A114
99	Picornaviridae	NC_038306	33757	Coxsackievirus A2
100	Picornaviridae	NC_038307	12072	Coxsackievirus B3
101	Picornaviridae	NC_038309	442851	Simian enterovirus SV4
102	Picornaviridae	NC_038311	573824	Rhinovirus A1
103	Picornaviridae	NC_038312	44130	Rhinovirus B3
104	Picornaviridae	NC_038878	992230	Human rhinovirus
105	Reoviridae	NC_007547	36427	Rotavirus C
106	Reoviridae	NC_007546	36427	Rotavirus C
107	Reoviridae	NC_007572	36427	Rotavirus C
108	Reoviridae	NC_007574	36427	Rotavirus C
109	Reoviridae	NC_007570	36427	Rotavirus C
110	Reoviridae	NC_007543	36427	Rotavirus C
111	Reoviridae	NC_007544	36427	Rotavirus C
112	Reoviridae	NC_007571	36427	Rotavirus C
113	Reoviridae	NC_007545	36427	Rotavirus C
114	Reoviridae	NC_007569	36427	Rotavirus C
115	Reoviridae	NC_007573	36427	Rotavirus C
116	Reoviridae	NC_011507	28875	Rotavirus A
117	Reoviridae	NC_011506	28875	Rotavirus A
118	Reoviridae	NC_011508	28875	Rotavirus A
119	Reoviridae	NC_011510	28875	Rotavirus A
120	Reoviridae	NC_011500	28875	Rotavirus A
121	Reoviridae	NC_011509	28875	Rotavirus A
122	Reoviridae	NC_011501	28875	Rotavirus A
123	Reoviridae	NC_011502	28875	Rotavirus A
124	Reoviridae	NC_011503	28875	Rotavirus A
125	Reoviridae	NC_011504	28875	Rotavirus A
126	Reoviridae	NC_011505	28875	Rotavirus A
127	Reoviridae	NC_021541	10942	Human rotavirus B
128	Reoviridae	NC_021545	10942	Human rotavirus B
129	Reoviridae	NC_021551	10942	Human rotavirus B
130	Reoviridae	NC_021543	10942	Human rotavirus B

131	Reoviridae	NC_021546	10942	Human rotavirus B
132	Reoviridae	NC_021544	10942	Human rotavirus B
133	Reoviridae	NC_021547	10942	Human rotavirus B
134	Reoviridae	NC_021548	10942	Human rotavirus B
135	Reoviridae	NC_021542	10942	Human rotavirus B
136	Reoviridae	NC_021550	10942	Human rotavirus B
137	Reoviridae	NC_021549	10942	Human rotavirus B
138	Retroviridae	NC_001436	11908	Human T-cell leukemia virus type I
139	Retroviridae	NC_001722	11709	Human immunodeficiency virus 2
140	Retroviridae	NC_001802	11676	Human immunodeficiency virus 1
141	Retroviridae	NC_011800	318279	Human T-lymphotropic virus 4
142	Togaviridae	KP003813	37124	Chikungunya virus
143	Togaviridae	NC_004162	37124	Chikungunya virus
144	Togaviridae	NC_001547	11034	Sindbis virus
145	Togaviridae	NC_003417	59301	Mayaro virus
146	Hepadnaviridae	X02763	10407	Hepatitis B virus
147	Hepadnaviridae	D00330	106821	Hepatitis B virus
148	Hepadnaviridae	AY123041	2764122	Hepatitis B virus
149	Hepadnaviridae	V01460	10407	Hepatitis B virus
150	Hepadnaviridae	X75657	10407	Hepatitis B virus
151	Hepadnaviridae	X69798	10407	Hepatitis B virus
152	Hepadnaviridae	AF160501	10407	Hepatitis B virus
153	Hepadnaviridae	AY090454	2847141	Hepatitis B virus
154	Flaviviridae	M62321	2847144	Hepacivirus C
155	Flaviviridae	D90208	11103	Hepacivirus C
156	Flaviviridae	D00944	11103	Hepacivirus C
157	Flaviviridae	D10988	11103	Hepacivirus C
158	Flaviviridae	D17763	356415	Hepacivirus C
159	Flaviviridae	D49374	357355	Hepacivirus C
160	Flaviviridae	Y11604	356418	Hepacivirus C
161	Flaviviridae	Y13184	356419	Hepacivirus C
162	Flaviviridae	Y12083	356420	Hepacivirus C
163	Flaviviridae	EF108306	1544902	Hepacivirus C
164	Flaviviridae	MH590698	11103	Hepacivirus C
165	Flaviviridae	NC_001710	54290	GB virus C