MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Mestrado em Programa de Pós-Graduação Biologia Computacional e Sistemas

METAGENÔMICA NA INVESTIGAÇÃO DE AGENTES INFECCIOSOS EM AMOSTRAS CLÍNICAS HUMANAS

LILIANE COSTA CONTEVILLE

Rio de Janeiro
Março de 2016

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

LILIANE COSTA CONTEVILLE

METAGENÔMICA NA INVESTIGAÇÃO DE AGENTES INFECCIOSOS EM AMOSTRAS CLÍNICAS HUMANAS

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Biologia computacional e Sistemas

Orientador (es): Dra. Ana Carolina Paulo Vicente
Dr. Marcos César Lima de Mendonça

RIO DE JANEIRO
Março de 2016

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

C843 Conteville, Liliane Costa

Metagenômica na investigação de agentes infecciosos em amostras clínicas humanas / Liliane Costa Conteville. – Rio de Janeiro, 2016.

xii, 103 f. : il. ; 30 cm.

Dissertação (Mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2016.

Bibliografia: f. 42-51

1. Metagenômica. 2. Sequenciamento de alto desempenho. 3. Análise taxonômica. 4. Dengue. 5. Casos febris negativos. I. Título.

CDD 616.91852075

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

LILIANE COSTA CONTEVILLE

METAGENÔMICA NA INVESTIGAÇÃO DE AGENTES INFECCIOSOS EM AMOSTRAS CLÍNICAS HUMANAS

ORIENTADORES: Dra. Ana Carolina Paulo Vicente
Dr. Marcos César Lima de Mendonça

Aprovada em: 17/03/2016

EXAMINADORES:

Prof. Dr. Gonzalo José Bello Bentancor **– Presidente** (IOC/FIOCRUZ)
Prof. Dr. Wim Maurits Sylvain Degrave (IOC/FIOCRUZ)
Prof. Dr. Cristiane Carneiro Thompson (UFRJ)
Prof. Dr. Edson Oliveira Delatorre (IOC/FIOCRUZ)
Prof. Dr. Luis Caetano Martha Antunes (ENSP/FIOCRUZ)

Rio de Janeiro, 17 de março de 2016

AGRADECIMENTOS

Aos meus pais, por sempre me apoiarem em minhas decisões e nunca medirem esforços para que eu alcançasse os meus objetivos.

Aos meus orientadores, Dra Ana Carolina e Dr Marcos, por confiarem a mim esse projeto, me orientarem e incentivarem.

Ao Dr Michel, pela paciência e auxílio em diversas etapas da realização desse projeto.

À Louise pela amizade e constante colaboração.

Aos amigos do LGMM, por terem me recebido tão bem e pelo apoio que me deram durante todos os momentos.

Às amigas do LABFLA, por sempre torcerem pelo meu sucesso.

Às minhas companheiras de sala, Daiana e Flávia, pela amizade e apoio psicológico.

À Pós-Graduação de Biologia Computacional e Sistemas.

Aos membros da banca examinadora, por aceitarem nosso convite.

À CAPES, pelo suporte financeiro.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

METAGENÔMICA NA INVESTIGAÇÃO DE AGENTES INFECCIOSOS EM AMOSTRAS CLÍNICAS HUMANAS

RESUMO

DISSERTAÇÃO DE MESTRADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

LILIANE COSTA CONTEVILLE

O vírus dengue infecta um número estimado de 50-100 milhões de pessoas anualmente em todo o mundo e no Brasil, dengue foi relatada pela primeira vez em 1981, desde então, a infecção tornou-se hiper-endêmica. As práticas atuais de diagnóstico não detectam o vírus em cerca de 50% dos casos suspeitos. Metagenômica é uma estratégia que pode ser aplicada na identificação de qualquer organismo em uma amostra, uma vez que recupera sequências de ácidos nucléicos que são analisadas contra bases de dados. Neste estudo, o nosso objetivo foi aplicar abordagens de metagenômica para analisar casos fatais de pacientes que apresentaram sintomas similares a dengue, mas com teste negativo para este vírus e também amostras de pacientes com suspeita de febre amarela. DNA e RNA genômico foram extraídos, amplificados com iniciadores randômicos e processados no sequenciador Illumina HiSeq2500. Em seguida, aplicamos um *pipeline* de bioinformática para filtragem de sequências de baixa qualidade e remoção de sequências humanas. Vários programas foram utilizados para classificar as *reads* metagenômicas: Kraken, GOTTCHA, SURPI, Metaphlan2, Taxoner e Blastn. As análises *in silico* identificaram patógenos que foram posteriormente confirmados por ensaios *in vitro* aplicando reagentes específicos. Deste modo, identificamos vírus e bactérias em 13% das amostras. Quatro desses vírus, com potencial de patogenicidade estavam em amostras distintas: Parvovírus B19, vírus da hepatite A, vírus da hepatite G e vírus Torque-teno. Todos eles, com exceção do vírus da Hepatite A estavam nos casos fatais. O Parvovírus B19 é um agente patogênico que tem sido associado a casos fatais. As bactérias identificadas foram *N. meningitidis*

do serogrupo C em duas amostras e *S. pneumoniae* em duas outras amostras. Esses são patógenos que causam surtos e epidemias no Brasil e têm alta taxa de mortalidade. Particularmente, *N. meninigitidis* sorogrupo C é o determinante de surtos atuais de *N. meningitidis* no Brasil. Além disso, dois genomas completos foram recuperados, o do Parvovírus B19 de um dos casos fatais (5,6 kb) e o do vírus Chikungunya (12 kb) que estava presente na amostra utilizada como controle positivo para as análises de metagenômica. Aqui, podemos demonstrar a aplicabilidade da metagenômica tanto na identificação quanto recuperação de genomas completos de patógenos a partir de amostras clínicas humanas.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

METAGENOMICS IN THE INVESTIGATION OF INFECTIOUS AGENTS IN HUMAN CLINICAL SAMPLES

ABSTRACT

MASTER DISSERTATION IN *COMPUTATIONAL AND SYSTEMS BIOLOGY*

LILIANE COSTA CONTEVILLE

Dengue virus infects an estimated 50–100 million people annually worldwide and in Brazil, dengue was first reported in 1981, since then the infection became hyper-endemic. The current diagnostic practices cannot detect the virus in around 50% of suspected cases. Metagenomic is a strategy that can be applied to recover any organism in a sample. In this study, our aim was to apply metagenomic approaches to analyse fatal cases of patients presenting dengue-like symptoms, but testing negative for this virus and samples of patients with suspected yellow fever. Genomic DNA and RNA were extracted, amplified with random primers and sequenced in the Illumina HiSeq2500 sequencer. We then followed a bioinformatic filtering pipeline to remove both low-quality sequences and human sequences. Several tools were used to classify the metagenome reads: Kraken, GOTTCHA, SURPI, Metaphlan2, Taxoner and Blastn. The *in silico* analysis identified pathogens that were further confirmed by *in vitro* assays applying specific reagents. In this way, we were able to detect viruses and bacteria in 13% of the samples. Four viruses, with pathogenicity potential were identified in distinct samples: Parvovirus B19, Hepatitis A virus, Hepatitis G virus and Torque-teno virus. All of them, except Hepatitis A virus were present in fatal cases. The Parvovirus B19 is in fact a pathogen that has been eventually associated to fatal cases. The bacteria identified were *N. meningitidis* serogroup C in two samples and *S. pneumoniae* in two other samples. Those are pathogens that cause outbreaks and epidemics in Brazil and have high mortality rate. Particularly, *N. meninigitidis* serogroup C is the determinant of the current *N. meningitidis* outbreaks in Brazil. Moreover, two complete genomes were recovered, the B19V from one of the fatal

cases (5.6 kb) and the Chikungunya virus (12 kb) that was in the sample used as a positive control to the metagenomic approach. Here, we demonstrated the aplicability of metagenomics in both the identification and recovery of complete genomes of pathogens from human clinical samples.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

B19V	Parvovírus B19
cDNA	DNA complementar
DENV	Vírus dengue
DNA	Ácido desoxirribonucleico
dsDNA	DNA dupla-fita
HAV	Vírus da Hepatite A
HGV	Vírus da Hepatite G
HTS	High-Throughput Sequencing – Sequenciamento de Alto-desempenho
LABFLA	Laboratório de Flavivírus do IOC/Fiocruz
NCBI	National Center for Biotechnology Information - Centro Nacional de Informação Biotecnológica
PCR	Reação em cadeia da polimerase
Refseq	Base de dados de referência do NCBI
RNA	Ácido ribonucleico
RT-PCR	Reação da transcriptase reversa seguida de reação em cadeia da polimerase
ssDNA	DNA fita simples
TTV	Torque-teno vírus

1 INTRODUÇÃO

O termo metagenômica foi empregado pela primeira vez em 1998, referindo-se à recuperação de genomas de bactérias cultiváveis e não cultiváveis, a partir de amostras de solo (1). A estratégia envolvia a extração do DNA total de uma amostra seguida por clonagem em cromossomos artificiais de bactérias e sequenciamento. Aplicando essa lógica, projetos em metagenômica foram desenvolvidos com a finalidade de montar genomas completos de bactérias não cultiváveis (2). Essa abordagem fornece uma visão da diversidade microbiana sem o viés da cultura, e expande o conhecimento relacionado à diversidade genética, estrutura de populações, e papéis ecológicos dos microrganismos (3).

Atualmente, com a metagenômica é possível detectar e identificar desde um microrganismo até um microbioma, sem requerer cultura, clonagem ou um conhecimento *a priori* de sua identidade/composição. A metagenômica pode ser aplicada em amostras complexas, compostas por diferentes ácidos nucleicos e torna possível, inclusive, a recuperação de sequências genômicas completas de organismos ainda não caracterizados (4,5).

O rápido crescimento de estudos de metagenômica de forma quantitativa e qualitativa só foi possível devido à capacidade de produção maciça de dados de forma rápida e sensível dos sequenciadores de alto desempenho (High Throughput Sequencing – HTS) (6). Com o aumento da profundidade e amplitude dessas tecnologias de sequenciamento, a metagenômica passou a contribuir com respostas para questões biológicas até então inatingíveis (7). Os experimentos aplicando HTS geram *terabytes* de dados brutos complexos, que precisam ser armazenados, processados e analisados, o que é um desafio para os bioinformatas. Felizmente, o desenvolvimento de softwares de bioinformática está avançando rapidamente, o que tem possibilitado a exploração dos dados metagenômicos por uma ampla gama de pesquisadores (8).

Para a aplicação da abordagem de metagenômica a uma questão biológica, é fundamental a realização de um estudo prévio do cenário a ser explorado, já que a metodologia e análises de bioinformática a serem aplicadas serão específicas para

diferentes situações. De um modo geral, uma análise de metagenômica compreende três etapas principais: Processamento da amostra biológica; sequenciamento numa plataforma de alto desempenho; e análise dos dados. Integrar as três etapas é de importância crítica para a interpretação e o sucesso da análise.

1.1. Processamento das Amostras e Sequenciamento

O processamento das amostras é o primeiro passo para um projeto de metagenômica partindo do *in vitro* para o *in silico* e, o mesmo, é definido com base no alvo do projeto. Por exemplo, amostras clínicas, tais como fezes, biópsia de tecido, expectoração e urina, são susceptíveis de conter uma grande quantidade de DNA e RNA do hospedeiro (9). Para evitar que o HTS gere uma quantidade significativa de sequências não relacionadas diretamente ao alvo do projeto, são aplicados protocolos que podem remover contaminantes e/ou enriquecer a amostra. Essa etapa é particularmente importante quando a proporção entre alvo e *background* desfavorece o alvo e, portanto, pode levar a uma sobrecarga de sequências do *background* (10).

Quanto a plataformas de HTS, a plataforma de pirosequenciamento desenvolvida pela Roche, o 454, foi muito utilizada inicialmente, pois gerava *reads* substancialmente maiores que as plataformas concorrentes (400-1000 bases). Porém, no caso de projetos metagenômicos, as plataformas Ilumina e SOLiD passaram a ser as mais utilizadas, pois embora gerem *reads* mais curtas do que o 454 (85-250 bases), o *output* é muito mais elevado (11). Da Illumina, o sequenciador HiSeq 2500, utilizado nesse estudo, vem sendo a escolha dos principais centros de sequenciamento e instituições ao redor do mundo, já que possui capacidade de gerar entre 10 gigabases e 1 terabase por corrida de forma rápida, acurada e com tamanho de *reads* consideráveis (100 a 250 bases) (12,13).

1.2. Análise dos Dados Metagenômicos

As centenas de milhares a milhões de *reads* geradas pelo HTS devem ser processadas computacionalmente de modo a se obter uma fração enriquecida com informações relacionadas ao alvo do estudo.

Os *pipeline*s computacionais utilizados para análise de metagenomas normalmente começam com a remoção dos adaptadores e filtragem de sequências de baixa qualidade. Uma pontuação de qualidade, *q*, é atribuída a cada base da *read* relacionada com a probabilidade estimada, *p*, da base estar errada, utilizando a seguinte fórmula: $q = -10 \times \log_{10}(p)$. Assim, uma pontuação de qualidade 20 corresponde a uma taxa de erro de 1 em 100, com uma precisão correspondente a 99%; uma pontuação 30 representa uma taxa de erro de 1 em 1000, com uma precisão correspondente a 99,9%. Na maioria das análises, bases com qualidade ≥20 são consideradas de alta qualidade, enquanto bases com qualidade menor devem ser removidas ou processadas de modo a eliminar a região com baixa qualidade (trimagem), pois as mesmas podem levar a resultados enviesados (14–17). Para essa etapa, podem ser aplicados, por exemplo os programas Cutadapt (18), Fastx-Toolkit (19), Prinseq (20), TagCleaner (21) e Trimmomatic (22).

Em seguida, é possível realizar um mapeamento para remover *reads* não correspondentes ao alvo do estudo, por exemplo, *reads* provenientes do genoma do hospedeiro (23–25). Os alinhadores disponíveis para análise de dados de HTS normalmente apresentam divergências quanto a sua funcionalidade, desempenho e precisão. Alguns dos fatores que podem influenciar seus desempenhos é a plataforma de sequenciamento utilizada, o organismo a ser mapeado e a quantidade de *mismatches* e *indels* presentes nas *reads* em relação a sequência referência (26). Entre os alinhadores mais utilizados estão o BWA (27), o Bowtie2 (28) e o SNAP (29). As *reads* que não alinharem aos genomas referência são as utilizadas para as análises subsequentes.

1.2.1. Análise Taxonômica

Nesta fase, será inferida a similaridade das *reads* com o conjunto de todos os organismos presentes em bancos de dados, que vão desde microrganismos a organismos superiores. Essa análise é normalmente baseada na caracterização de sequências por similaridade e/ou no agrupamento de sequências por grupos taxonômicos particulares (*binning*). Um metagenoma processado dessa forma gera um perfil de táxons ou relações filogenéticas. Essas análises podem ser realizadas tanto com os dados crus quanto com os mesmos montados em *contigs*, *ou seja,* sequências formadas a partir da sobreposição das *reads,* utilizando a montagem à base de referência ou montagem *de novo* (30).

A caracterização de sequências por similaridade implica na comparação do metagenoma com sequências de bases de dados públicas, por exemplo, *GenBank*. E uma vez que sequências relacionadas ao organismo alvo estejam presentes nas bases de dados, pode-se chegar a uma identificação taxonômica (31). Caso contrário, o organismo alvo pode vir a ser identificado a partir de similaridade a uma espécie próxima (32).

Um programa muito utilizado para esse tipo de análise é o programa BLAST (33), porém, ele não foi projetado para análises de metagenômica. Esse programa classifica cada *read* ao encontrar o melhor alinhamento entre ela e sequências presentes na base de dados utilizada. Atualmente, existem diversas ferramentas especializadas em alinhar sequências provenientes de HTS, como os citados anteriormente: BWA (27), o Bowtie2 (28) e o SNAP (29). Alguns *pipelines* de análise metagenômica utilizam esses alinhadores para realizar a análise taxonômica, como por exemplo, o *pipeline* SURPI (23) e o Taxoner (34), que utilizam, respectivamente, os alinhadores SNAP e Bowtie2 em suas análises.

Outros programas baseiam-se em bases de dados de genes "marcadores", ou genes específicos de certos clados microbianos, alguns deles são: Metaphlan2 (35), GOTTCHA (36) e MetaPhyler (37). Uma vez que as bases de dados contêm apenas uma pequena amostra de cada genoma, esses programas classificam uma pequena porcentagem de sequências de um metagenoma. Outros utilizam algoritmos que analisam as sequências com base na sua composição de *k*-mers

(sequências de oligonucleotídeos de comprimento k), comparando-os com os k-mers presentes nas sequências de sua base de dados. Entre esses programas estão o Kraken (32) e o LMAT (38).

A acurácia com a qual qualquer um desses métodos vai estimar a diversidade do metagenoma está relacionada a diversos parâmetros, como a cobertura recuperada, ao número de *reads* e a natureza das *reads* (pareadas ou *singletons*). Por exemplo, uma grande quantidade de reads atribuídas a um único organismo, aumenta a confiança ao resultado obtido pelos programas de análise taxonômica. Além disso, utilizar *reads* pareadas aumenta substancialmente a precisão da análise (39). Em alguns casos, controles negativos podem ser utilizados para identificar possíveis contaminantes tanto originados de reagentes aplicados no processamento das amostras como organismos presentes na microbiota.

1.3. Aplicação da Metagenômica no diagnóstico de doenças infecciosas

A metagenômica tem contribuído no diagnóstico de doenças infecciosas, tanto relacionadas a casos específicos quanto a surtos ou epidemias. Para a identificação dos agentes causadores desses casos, já foram utilizados diversos tipos de amostras clínicas e a escolha das mesmas tem papel fundamental na recuperação do agente infeccioso pesquisado.

Na Nicarágua, foi realizado um estudo com amostras de soro de pacientes febris negativas para dengue, nas quais foram encontradas sequências relacionadas a vírus das famílias *Herpesviridae*, *Flaviviridae*, *Circoviridae*, *Anelloviridae*, *Asfarviridae*, e *Parvoviridae*, assim como novos genótipos e espécies virais até então desconhecidas (40). Um segundo estudo elucidou um surto causado por um novo bunyavirus na China a partir da aplicação da metagenômica em amostras de soro de pacientes com febre, trombocitopenia e leucopenia (41). O vírus Bas-Congo foi outro novo vírus identificado por metagenômica, ao analisarem de amostras de sangue de pacientes durante um surto de febre hemorrágica no Congo Africano (42). Outros novos vírus identificados por metagenômica em amostras clínicas humanas

foi o cyclovirus, encontrado no líquido cefalorraquidiano de pacientes com paraplegia (43) e um novo arenavírus transmitido por transplante de órgãos (44).

Alguns estudos aplicaram a metagenômica em amostras de pacientes com encefalite e identificaram *Leptospira* (45), astrovirus (46), e bornavirus (47). Em 2011, a bactéria *Escherichia coli* Shiga toxigênica, responsável por um surto de gastroenterite na Alemanha, foi identificada a partir da análise de amostras de fezes de pacientes com diarréia (48). Uma pandemia causada pelo vírus da influenza A em 2009 foi elucidada pela análise metagenômica de swabs de nasofaringe coletados durante as primeiras fases da pandemia no México, Canadá e Estados Unidos (49). Assim como os vírus causadores de diversos surtos não resolvidos de gastroenterites na Nova Zelândia foram identificados nas amostras de fezes dos pacientes (50). E essa abordagem não se limita a descoberta de bactérias e vírus, recentemente, um caso de filariose foi solucionado por metagenômica (51).

1.4. Dengue e Febre Amarela

O vírus de dengue (DENV) apresenta os sorotipos: DENV-1, DENV-2, DENV-3 e DENV-4. Atualmente, os quatro sorotipos circulam no Brasil e o país é hiper-endêmico para esta virose (52). A infecção pelo DENV pode causar desde quadros assintomáticos a manifestações clínicas que variam de uma doença febril aguda e autolimitada até síndromes que podem ser fatais. O dengue clássico é uma doença febril aguda caracterizada por dor de cabeça frontal, dor retro ocular, dores musculares e articulares, náuseas, vômitos e exantema. Normalmente, a dengue grave começa com os sintomas do dengue clássico, mas o quadro se agrava com manifestações hemorrágicas, aumento da permeabilidade vascular, trombocitopenia, hepatomegalia, insuficiência circulatória e choque hipovolêmico (53). Esses quadros clínicos não são exclusivos dessa infecção, muitas outras, tanto virais quanto bacterianas, apresentam sintomatologias semelhantes.

O vírus da febre amarela (YFV) é outro arbovírus que causou diversos surtos no Brasil, principalmente durante o século XIX e tornou-se um problema de saúde pública grave até que estratégias bem sucedidas de controle de vetores e de vacinação eliminaram a transmissão urbana em 1942 (54). Desde então, são

reportados apenas casos e surtos humanos esporádicos, resultados da expansão e atividade humana em áreas de selva, onde o ciclo selvagem ainda é endêmico (55). Nos casos sintomáticos, a infecção pelo YFV pode causar manifestações repentinas de febre alta, calafrios, cansaço, dor de cabeça e muscular, náuseas e vômitos. A forma mais grave da doença é rara, porém, pode causar cansaço intenso, icterícia e insuficiência hepática e renal, seguida de febre hemorrágica em cerca de 10% dos indivíduos infectados (56).

No diagnóstico laboratorial de rotina são utilizadas metodologias específicas (RT-PCR, pesquisa de antígeno, IgM e isolamento) para identificação do agente etiológico associado ao quadro clínico. Porém, os casos com diagnóstico laboratorial negativo, muitas vezes, ficam sem a definição do seu agente causal, apesar de serem submetidas a uma bateria de testes laboratoriais convencionais (57).

A pesquisa por outros agentes durante surtos de dengue é absolutamente fundamental para a definição do real cenário epidemiológico presente no Brasil, assim como para vigilância da introdução de potenciais novos agentes infecciosos. Em 2007, no Brasil, de um total de 9287 amostras recebidas de pacientes com suspeitas de dengue e testadas para esse vírus por isolamento viral, apenas 1791 (19,2%) foram confirmadas (58). No banco de dados do Laboratório de Flavivírus do IOC/Fiocruz (LABFLA), uma porcentagem significativa de amostras de pacientes com suspeita de dengue teve resultados negativos para dengue, mesmo após serem testadas pelas abordagens utilizadas na rotina dos laboratórios de referência: RT-PCR, pesquisa de antígeno NS1 e IgM. (Figura 1)

Considerando que outras infecções virais e/ou bacterianas com potencial de causar surtos e/ou epidemias podem estar ocorrendo durante epidemias de dengue, a exploração de outros agentes infecciosos nos casos não confirmados é fundamental. Isto é realizar vigilância epidemiológica de ponta, principalmente em um país onde vetores de diferentes agentes infecciosos são endêmicos. Portanto, a metagenômica seria a abordagem de escolha para se aplicar a situações emergentes e não esclarecidas.

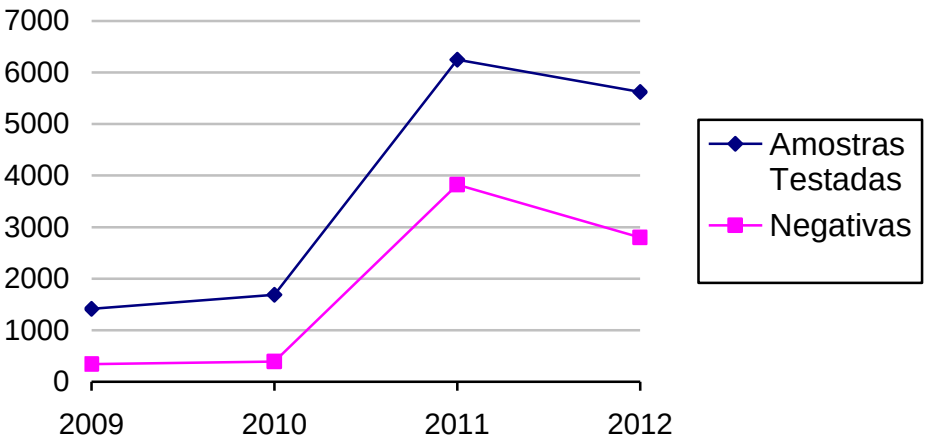


Figura 1: Amostras com suspeita de dengue testadas pelo LABFLA (em azul) e amostras não confirmadas para dengue (em rosa) entre 2009 e 2012.

2 OBJETIVO

2.1 Objetivo Geral

Aplicar a abordagem de metagenômica e análises computacionais para a pesquisa de agentes infecciosos, em amostras de pacientes com suspeita clínica de dengue ou febre amarela, mas com diagnóstico laboratorial negativo.

2.2 Objetivos Específicos

- Recuperar e processar sequências metagenômicas das amostras em questão
- Estabelecer uma identificação taxonômica das sequências processadas considerando bancos de dados de bactérias e vírus.
- Realizar análises *in vitro* com iniciadores específicos para comprovar os resultados *in silico.*
- Realizar análises genéticas relativas aos organismos identificados.

3 METODOLOGIA

3.1 Amostras

Neste estudo foi analisado um total de 99 amostras clínicas, 94 dessas eram de óbitos que apresentaram quadro clínico para dengue, mas com diagnóstico laboratorial negativo. E as outras cinco amostras eram de pacientes com suspeita clínica de febre amarela, mas com diagnóstico laboratorial negativo tanto para febre amarela quanto para dengue também foram incluídas. Como controle positivo da metodologia utilizamos um vírus sem relação genética com dengue e febre amarela da seguinte forma: uma amostra de uma cultura de células inoculada com chikungunya e outra de soro de paciente com presença de chikungunya confirmada laboratorialmente, foram processadas e submetidas ao sequenciamento de alto-desempenho. As 99 amostras eram compostas de amostras clínicas de diferentes naturezas: soro ou macerado de vísceras, como fígado. Essas amostras foram coletadas entre os anos 2008 e 2015, estavam preservadas em freezer -70º C dentro do banco de amostras do LABFLA. Essas amostras fazem parte do programa Apoio ao Estudo de Doenças Negligenciadas e Reemergentes (E-25/010.001558/2014).

3.2 Extração de Ácidos Nucléicos

Sem o conhecimento prévio da natureza do ácido nucléico que caracterizaria um eventual patógeno, as amostras foram processadas considerando tanto RNA quanto DNA. 110 µL das amostras clínicas foram homogeneizados e centrifugados de modo a retirar material particulado e restos celulares. Em seguida, o material foi submetido a distintos processamentos:

1- Amostras tratadas com DNAse

2 - Amostras tratadas com DNAse e RNAse

3 – Amostras sem tratamento com DNAse e RNAse

Nesta etapa, as enzimas DNAse (TURBO DNAse I Life Technologies) e/ou RNAse (RNAse One Ribonuclease Promega) foram utilizados para eliminar contaminantes de ácido nucléico do hospedeiro, o que levaria a um enriquecimento dos ácidos nucléicos relativos aos patógenos.

Estas amostras foram então, submetidas a extração de ácidos nucléicos. O processo de extração foi realizado utilizando o *kit High Pure Viral Nucleic Acid* (Roche), seguindo o protocolo de recomendações do fabricante. O protocolo foi modificado para o processamento de cinco amostras, pois não foi utilizado o *carrier poly(A)* sugerido no *kit*.

3.3 Reação de RT-PCR Randômico

Aplicamos essa abordagem tanto para transcrever o RNA fita simples em DNA dupla fita (dsDNA) quanto para enriquecer o DNA presente na amostra, pois o preparo das bibliotecas genômicas parte de fragmentos de dsDNA (30,59).

Antes do procedimento, os produtos extraídos foram aquecidos a 65º C por 5 minutos, e mantidos em gelo até o momento do uso, de modo a disponibilizar fitas simples lineares, tanto de RNA quanto de DNA. A síntese do cDNA foi realizada utilizando o kit Cloned AMV Reverse Transcriptase da Invitrogen, o iniciador GACCATCTAGCGACCTCCACNNNNNNNN a 10 µM e o mix recomendado pelo fabricante em um volume final de 25µL. A transcrição reversa foi realizada por uma etapa inicial de 25º C por 5 minutos para ligação randômica dos iniciadores, em seguida prosseguimos com 45º C por 60 minutos e inativação da enzima a 85º C por 5 minutos. Após a transcrição reversa, foi realizada a reação de PCR, utilizando os iniciadores GACCATCTAGCGACCTCCACNNNNNNNN e GACCATCTAGCGACCTCCAC, que se diferenciam pela adição de 8 N na extremidade 3’ de um dos iniciadores, sendo esse, o mesmo utilizado na transcrição reversa.

A reação foi conduzida com duas etapas consecutivas. A PCR foi então realizada com cinco ciclos de desnaturação a 94º C por 2 minutos, hibridização a 25º C por 30 segundos e extensão a 72º C por 1 minuto seguido de 40 ciclos de 94º C

por 30 segundos, 55º C por 30 segundos e 72º C por 1 minuto, seguido por 72° C durante 5 minutos e 4° C (60). A primeira etapa foi realizada com uma temperatura de hibridização baixa (25º C) para possibilitar a ligação randômica dos iniciadores (que possuem 8N). Enquanto a segunda etapa teve uma temperatura de hibridização de 55º C, a ideal para que as bases complementares se anelassem. O produto da PCR foi submetido a eletroforese em gel de agarose a 1,5%, corado por brometo de etídeo.

3.4 Sequenciamento de alto-desempenho em Illumina Hi-Seq

As amostras foram organizadas em *pools* de cinco amostras cada, que foram submetidos ao sequenciamento de alto-desempenho no sequenciador Illumina HiSeq 2500 no modo rápido, que apresenta um desempenho mínimo necessário para nossos objetivos. Foram realizadas duas corridas.

Na primeira, foram incluídos 20 pools, sendo que um deles continha uma amostra controle positivo de cultura de células inoculada com chikungunya. Os *pools* foram quantificados em Qubit, e 1 nanograma de dsDNA foi utilizado para a produção de biblioteca genômica por tagmentação e indexação - marcação por indexes - utilizando o kit Nextera XT.

Para a segunda corrida, foram escolhidas aleatoriamente, 20 amostras das 100 previamente sequenciadas, com as quais formamos quatro *pools*. Além desses quatro *pools*, também foi sequenciada uma amostra de soro positivo para chikungunya. Cada *pool* foi marcado na etapa de construção de biblioteca. Os *pools* foram processados da mesma forma da primeira corrida.

Para a realização do sequenciamento, utilizamos o serviço da Plataforma de Sequenciamento de Alto Desempenho do IOC/FIOCRUZ.

3.5 Análise Computacional

3.5.1 Análise de Qualidade e Pré-processamento

Para análise de qualidade dos dados gerados foi utilizada a ferramenta FastQC, que gera gráficos e tabelas baseados em diferentes aspectos de qualidade, como qualidade por base, conteúdo GC e presença de adaptadores nas *reads*. Os arquivos *fastq* de cada *pool* passaram por uma etapa de pré-processamento, que consiste de: (i) trimagem das *reads* com qualidade abaixo de 20 e retirada dos adaptadores utilizando o programa cutadapt (18) e; (ii) remoção de sequências com baixa complexidade utilizando o algoritmo dust com score 49 e trimagem das *reads* menores que 30 bp pelo programa PRINSEQ (20). Sequências de baixa complexidade são definidas como tendo trechos de nucleotídeos repetidos, com conteúdo de informação limitada e biologicamente insignificante. O algoritmo dust calcula a complexidade das sequências e as mesmas são escaladas de 0 a 100. Valores mais altos implicam menor complexidade. Por exemplo, uma sequência de homopolímeros tem pontuação 100, enquanto sequências de dinucleotídieos, como por exemplo uma sequência TATATATATA, tem uma pontuação em torno de 49 (20).

3.5.2 Programas de Análise Taxonômica

Após o pré-processamento, as *reads* pareadas e não pareadas foram mapeadas com o programa Bowtie2 contra a sequência referência de genoma humano mais recente, o Hg38. As *reads* que não alinharam ao Hg38 foram submetidas a programas que caracterizam metagenomas taxonomicamente por similaridade com sequências referências presentes em bases de dados. Nesta etapa, foram utilizados os programas: Kraken, Taxoner, Metaphlan2, GOTTCHA e BLASTn. O *pipeline* SURPI também foi utilizado, porém, a ele foram submetidas as *reads* cruas, pois este *pipeline* utiliza seus próprios parâmetros de pré-processamento e mapeamento contra o genoma humano. Os programas foram executados seguindo as configurações presentes no *default*.

Kraken é uma ferramenta baseada em *k-mers*, especializada em recuperar o sinal taxonômico presente em metagenomas. O programa extrai k-mers das *reads* e pesquisa pelos *k-mers* no banco de dados de referência. Utilizamos o programa com o banco standard, formado por genomas completos de bactérias, arquéias e vírus presentes na base de dados do National Center for Biotechnology Information (NCBI) (32).

Taxoner utiliza bancos de dados com genomas concatenados e indexados pelo programa Bowtie2. Cada genoma é um segmento que está anotado por identificadores, incluindo nome taxonômico e identificador. Utilizamos as bases de dados pré-indexadas de bactérias, vírus e fungos disponibilizadas no programa. No Taxoner, as *reads* são mapeadas diretamente contra esses bancos de dados e os táxons são identificados com um algoritmo que busca seu ancestral comum mais próximo (LCA) (34).

Metaphlan2 utiliza o Bowtie2 para o mapeamento das *reads* metagenômicas contra uma base de dados de marcadores de genes clado-específicos de bactérias, arquéias, eucariotos, vírus e fungos. Sua base de dados foi construída a partir de ~ 17.000 genomas referência (35).

GOTTCHA é um *pipeline* que consiste em três etapas: (i) trimagem e fragmentação das *reads* em *30-mers* não sobrepostos; seguido de (ii) mapeamento das *reads* contra a base de dados referência e (iii) filtragem dos resultados. Utilizamos as bases de dados pré-computadas disponibilizadas pelos desenvolvedores do GOTTCHA. Essas bases de dados são formadas por segmentos únicos de genomas de vírus e bactérias (36).

No BLASTn, são formadas sequências curtas (*seeds*) de 11 bases a partir das *reads* e essas são comparadas com as *seeds* formadas a partir das sequências presente no banco de dados (33). Utilizamos como referência, sequências de vírus presentes no RefSeq do NCBI.

O *pipeline* SURPI foi criado para identificação de patógenos em dados complexos de metagenômica gerados a partir de amostras clínicas. Esse *pipeline* consiste de ferramentas para trimagem, mapeamento contra genoma humano, alinhamento contra a base de dados do NCBI e montagem *de novo* (23).

Alguns dos resultados obtidos com esses programas foram testados utilizando o alinhador BWA, com a finalidade de mapear as *reads* com uma sequência referência específica, correspondente ao táxon de um potencial candidato identificado. A referência específica para cada organismo foi aquela que recebeu mais *hits* pelo programa SNAP do pipeline SURPI ou a sequência referência do NCBI. Em seguida, foram utilizados os programas SAMtools, para gerar um arquivo que contém todas as informações das *reads* mapeadas, e UGENE, para visualizar o mapeamento. Para montagem *de novo* de alguns genomas, também foi utilizado o programa SPAdes 3.5.0. Ambas etapas tornaram possível a recuperação desde genomas parciais a genomas completos de organismos candidatos.

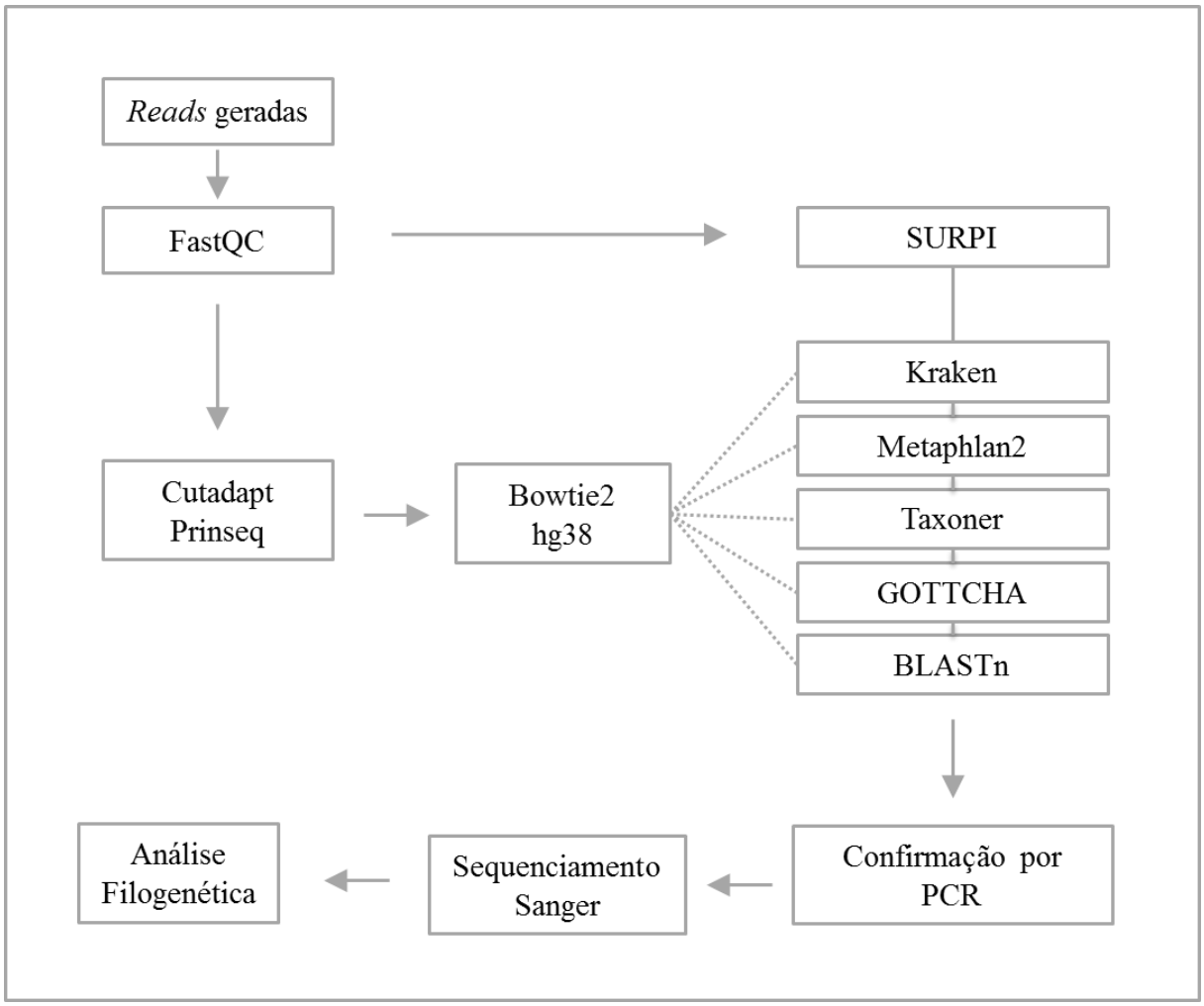


Figura 2: Fluxograma utilizado para identificação de patógenos.

3.6 Ensaios *in vitro*

Iniciadores específicos para alguns grupos taxonômicos identificados *in silico* foram aplicados em reações de RT-PCR ou PCR. Essas reações foram realizadas com o objetivo de confirmar a identificação metagenômica. As amostras foram testadas em função do resultado obtido para o pool ao qual pertenciam. Os *amplicons* foram sequenciados na Plataforma de Sequenciamento de DNA pelo método de Sanger do IOC/FIOCRUZ.

Tabela 1: Iniciadores específicos utilizados para identificar organismos identificados in silico.

Organismo	Primers (5'-3')	Referência
Parvovirus B19	TTCTTTTCAGCTTTTAGG TTTATACAGTGTCCTTAT TATAAGTTTCCTCCAGTGCC GTACTTCTGGTACGTTAAGT	(61)
Vírus da Hepatite A	GATACCTCACCGCCGTTTGC TCAATGCATCCACTGGATGAG CGTTTGCCTAGGCTATAGGCT CAGTCCTYCGGCGTTGAATGG	(62)
Vírus da Hepatite G	TGCCACCCGCCCTCACCCGAA AGGTGGTGGATGGGTGAT GGRGCTGGGTGGCCYCATGCWT TGGTAGGTCGTAAATCCCGGT	(63)
Torque-Teno Vírus	GTGGGACTTTCACTTGTCGGTGTC GACAAATGGCAAGAAGATAAAGGCC CAGACTCCGAGTTGCCATTGGAC CACGTGTCGGGGCCTACTTCCG	(64)
Neisseria meningitidis	TGTTCCGCTTCGACTGCCAAC TCCCCGTCGTAAAAAACAATC	(65)
Streptococcus pneumoniae	ATGGACAAACCAGCNAGYTT GCTTGAGGTCCCATRCTNCC	(65)

3.7 Análise Filogenética/Filogenômica

As sequências correspondentes aos organismos identificados nas amostras com suspeita clínica de dengue e febre amarela, mas não confirmados, foram submetidas a análises filogenéticas ou filogenômicas. O modelo de evolução foi escolhido para cada conjunto de dados pelo teste Bayesian Information Criterion (BIC), usando o programa ModelGenerator 0.85. Todas as árvores filogenéticas foram geradas através do método de Máxima Verossimilhança pelo programa MEGA 6 considerando 1000 réplicas de *bootstrap*.

4 RESULTADOS

O sequenciamento de alto desempenho na plataforma Illumina HiSeq 2500 operando no modo rápido pode, dependendo do protocolo aplicado, produzir *reads* de 100 a 250 bases tanto de sequências pareadas quanto fitas simples, num total de ~ 300 milhões *reads* por corrida.

Em nosso estudo, realizamos duas corridas no modo rápido com protocolos que geram *reads* pareadas de 100 bases. Na primeira corrida, quando 100 amostras foram combinadas em 20 *pools* de 5 amostras cada (chamados de *pool* 1 ao *pool* 20), foram obtidas 46.820.670 *reads.* Na segunda corrida, quando foram sequenciados um controle positivo e 4 *pools* cada um com 5 amostras (chamados de *pool 2*1 ao *pool* 25)*, foram obtidas* 434.315.516 *reads.*

Houve uma grande discrepância entre a quantidade de reads geradas pelas duas corridas realizadas. Isso ocorreu devido ao número de amostras sequenciadas na mesma corrida. Na primeira corrida, além dos nossos 20 *pools*, 76 amostras de outros grupos foram sequenciadas*.* Na segunda corrida, além dos nossos 5 *pools,* outras 19 amostras foram sequenciadas*.*

4.1 Processamento dos dados

O primeiro passo nesse tipo de análise é avaliar a qualidade das reads geradas. Para isso, utilizamos o programa FastQC e o resultado desta análise está na Figura 3. Pudemos verificar que de um modo geral a qualidade das reads foi boa, uma vez que Q estava acima de 20.

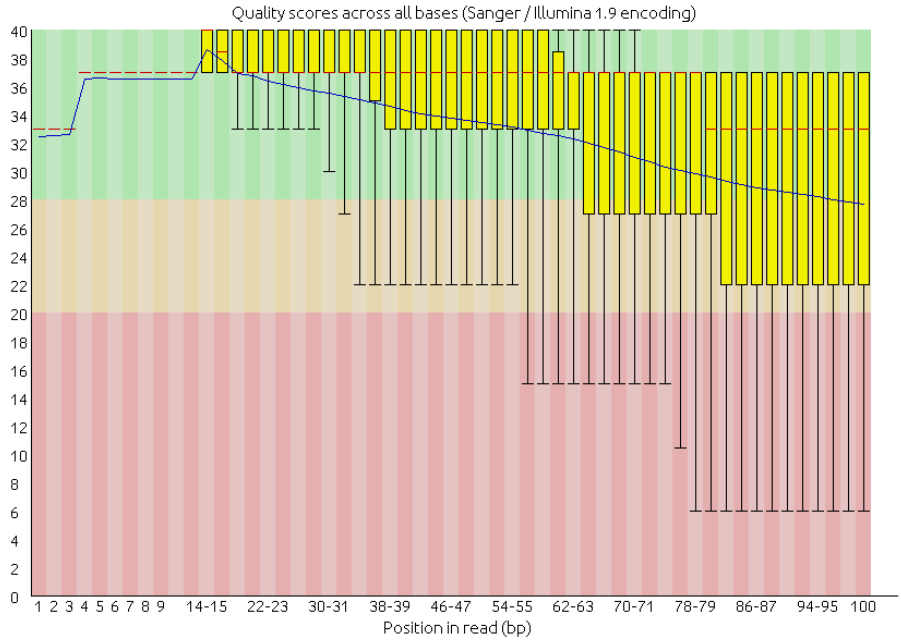


Figura 3: Imagem representativa da qualidade das *reads* geradas pelo sequenciamento.

Ao realizarmos a análise da primeira corrida, observamos que havia uma grande porcentagem de *reads* homopoliméricas, que foram trimadas uma vez que *reads* com regiões repetitivas tendem a ser alinhadas de forma incorreta, causam um viés nos resultados da análise e sobrecarregam o processamento das mesmas (66) (Figura 4).

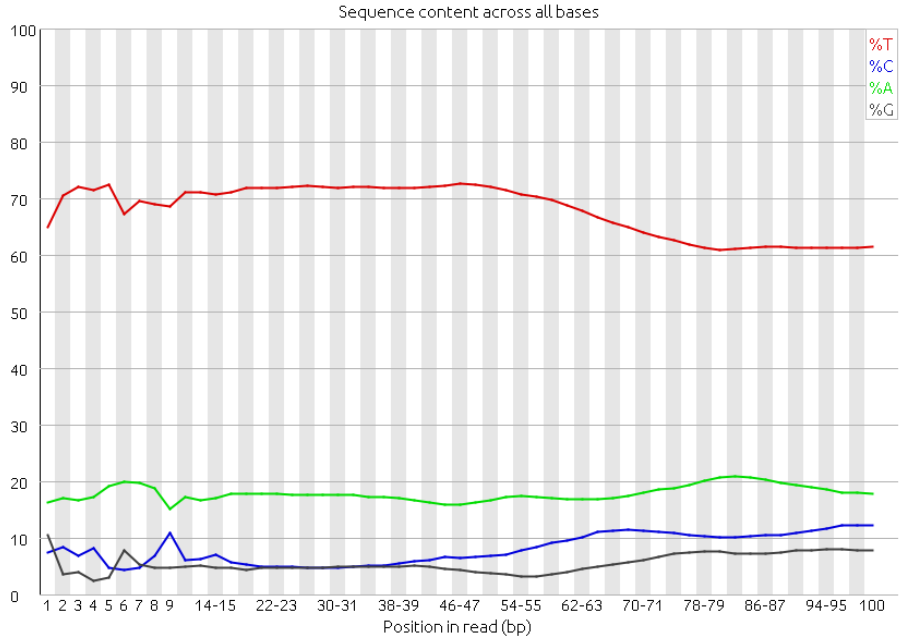


Figura 4: Conteúdo de bases das *reads* geradas a partir dos pools da primeira corrida.

Concluímos que isso provavelmente ocorreu devido a utilização do carrier RNA de poly(A) presente no kit de extração utilizado. Para contornar esse problema, alguns grupos: i) utilizam carriers que não são formados por ácidos nucléicos, como poliacrilamida linear (67); ii) não utilizam carrier (68,69); iii) utilizam o carrier poly(A), mas na amplificação, utilizam primers com hexameros modificados, desenhados especificamente para bloquear a transcrição reversa do RNA do carrier (70,71). Atualmente, documentos da Illumina sobre preparação de bibliotecas aconselham utilizar apenas carriers que não servem como template ou bloqueiam o template, como acrilamida linear, n- ou p- carriers (72).

Após a etapa de pré-processamento, restaram 18.499.960 *reads* da primeira corrida, sendo essas pa*read*as ou *singletons* (Figura 5, Tabela S1).

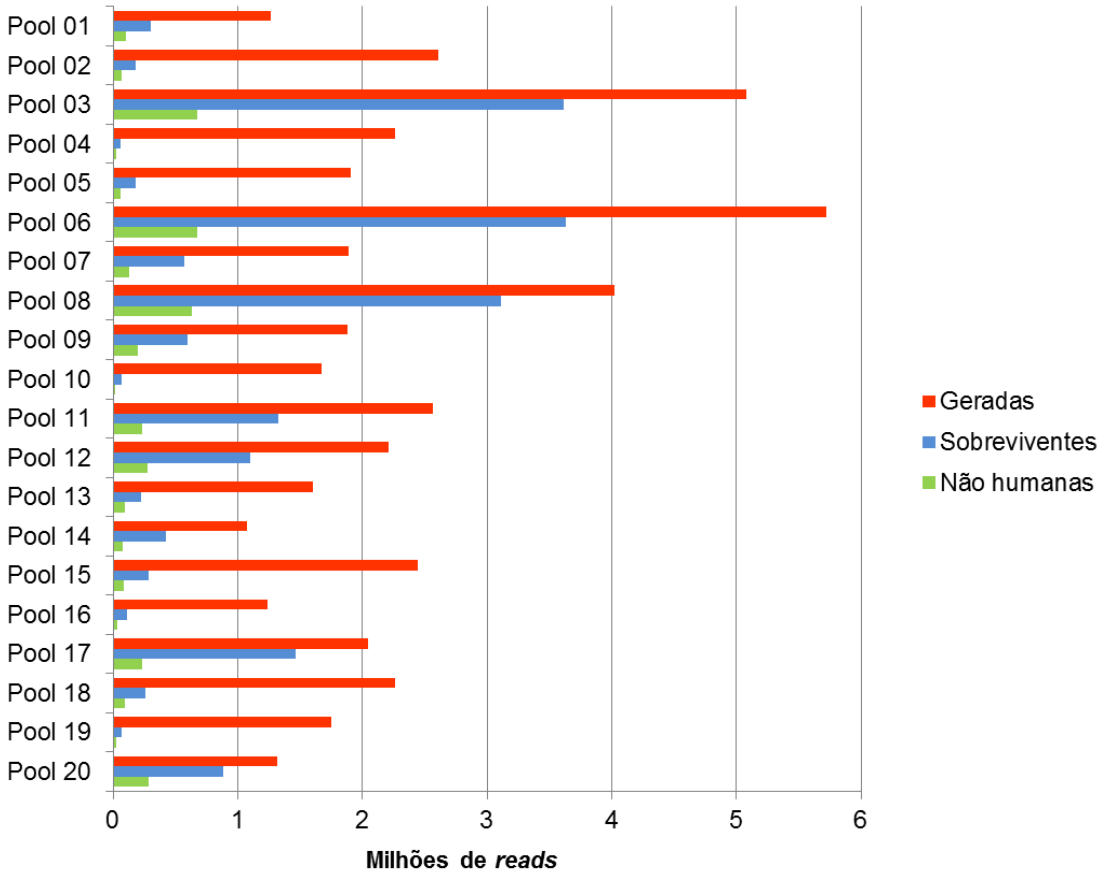


Figura 5: Quantidade de *reads* geradas e processadas de cada pool da primeira corrida. Em vermelho, *reads* geradas pelo sequenciamento; em azul, *reads* após o pré-processamento; e em verde, *reads* não mapeadas com o genoma humano.

Realizamos a segunda corrida com o objetivo de mudar o protocolo de processamento das amostras, ou seja, utilizamos RNAse em todas amostras e além disso, em um dos *pools* não foi utilizado o *carrier poly(A)* durante a extração dos ácidos nucleicos. Nesta corrida, não observamos a prevalência de *reads* com sequências homopoliméricas, mesmo utilizando ou não o *carrier poly(A)*.

Após a trimagem das *reads* geradas *a* partir dos *pools* 21 ao 25, restaram 125.488.671 *reads*, sendo essas pa*read*as ou *singletons* (Figura 6, Tabela S1).

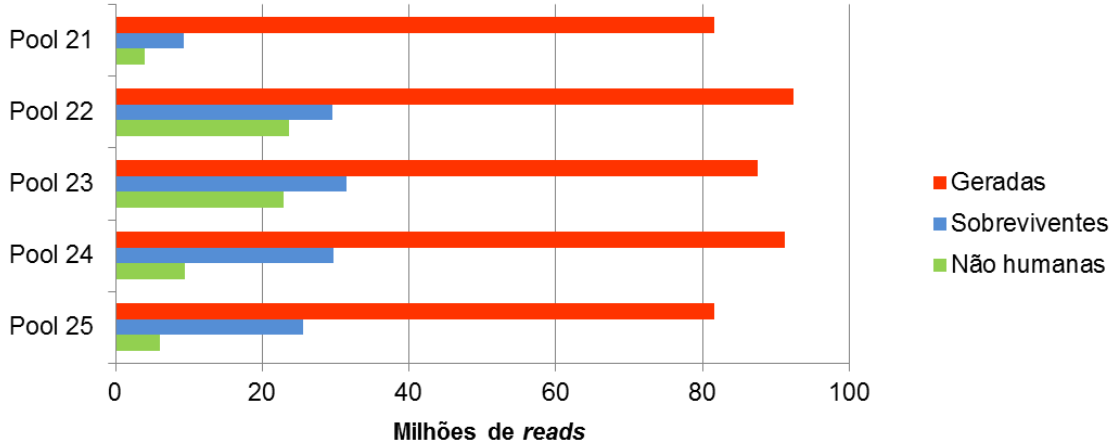


Figura 6: Quantidade de *reads* geradas e processadas de cada pool da segunda corrida. Em vermelho, *reads* geradas pelo sequenciamento; em azul, *reads* após o pré-processamento; e em verde, *reads* não mapeadas com o genoma humano.

O próximo passo foi mapear as *reads* que passaram pelo pré-processamento contra um genoma humano, uma vez que as amostras têm origem em material clínico humano. O resultado foi que ~ 50% a 67% de *reads* da primeira corrida alinharam ao genoma humano (Figura 4, Tabela S1). Enquanto que na segunda corrida ~ 2% a 30% das *reads*, dependendo do *pool,* alinharam ao genoma humano (Figura 5, Tabela S1). Comparando as duas corridas, a segunda teve menos *reads* mapedas contra o genoma humano, demonstrando que a utilização de DNAse e RNAse no processamento das amostras da segunda corrida foi importante para o enriquecimento do nosso alvo em relação ao hospedeiro.

4.2 Mapeamento

As *reads* foram comparadas com sequências de vírus e bactérias presentes em bancos de dados, com a finalidade de analisá-las taxonomicamente. Cada programa utilizado gerou arquivos de saída com os organismos identificados em cada *pool.* Por apresentarem algoritmos e bancos de dados distintos, esses programas geraram, muitas das vezes, resultados distintos. A análise dos dados da primeira corrida identificou 365 gêneros de bactérias (Figura 7) e 157 vírus (incluindo fagos) (Figura 8). Cada organismo identificado em cada *pool* por cada programa utilizado encontra-se nas Tabelas Suplementares (S2 e S3).

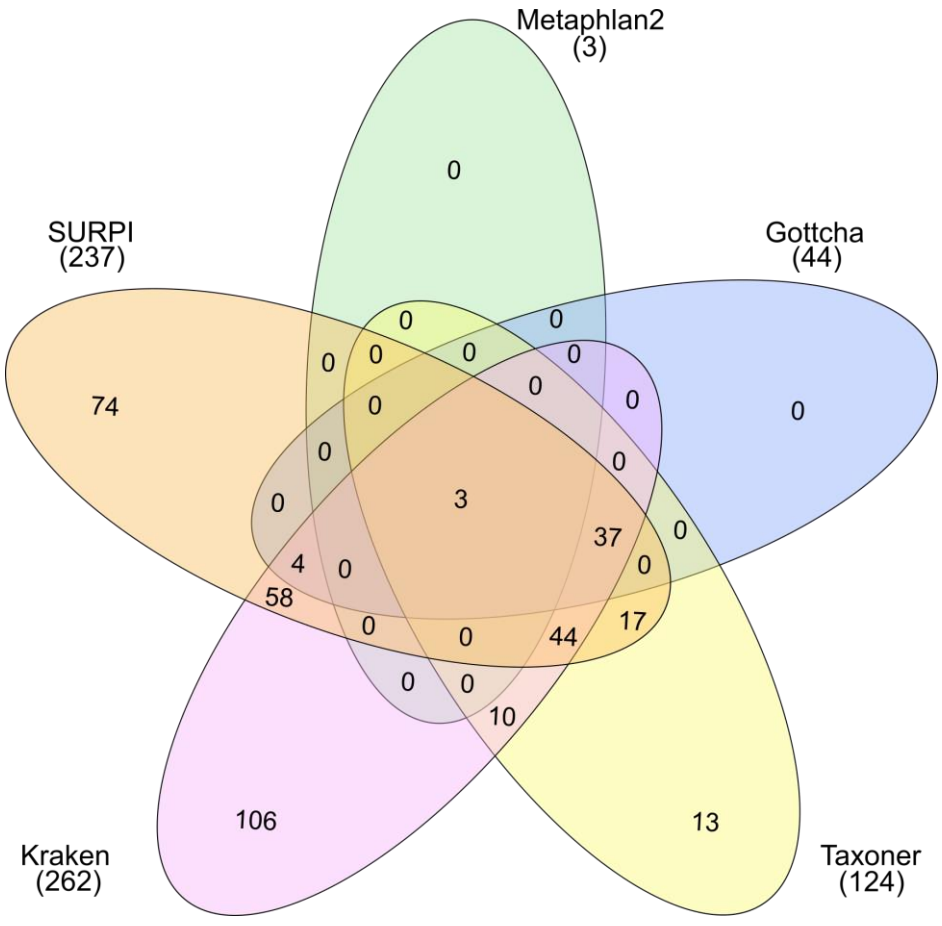


Figura 7: Diagrama representando o número de gêneros de bactérias identificados pelos programas de análise taxonômica a partir dos pools da primeira corrida.

Na figura 8, está representado os resultados gerados por cinco dos seis programas aplicados, já que o programa Metaphlan2 identificou apenas três vírus, que também foram identificados pelos demais e sua exclusão possibilitou uma melhor resolução do diagrama. Em relação ao controle positivo dessa corrida, um inóculo do vírus chikungunya em cultura de células presente no *pool* 16, nenhum dos programas identificou *reads* deste vírus. Entretanto, nesse *pool* foram identificadas *reads* com similaridade a membros da família *Mesoniviridae*: os vírus Nam Dinh, Cavally, Hana e Casuarina, que são vírus de mosquito (73). Provavelmente os mesmos se encontravam nas células nas quais o vírus chikungunya foi cultivado.

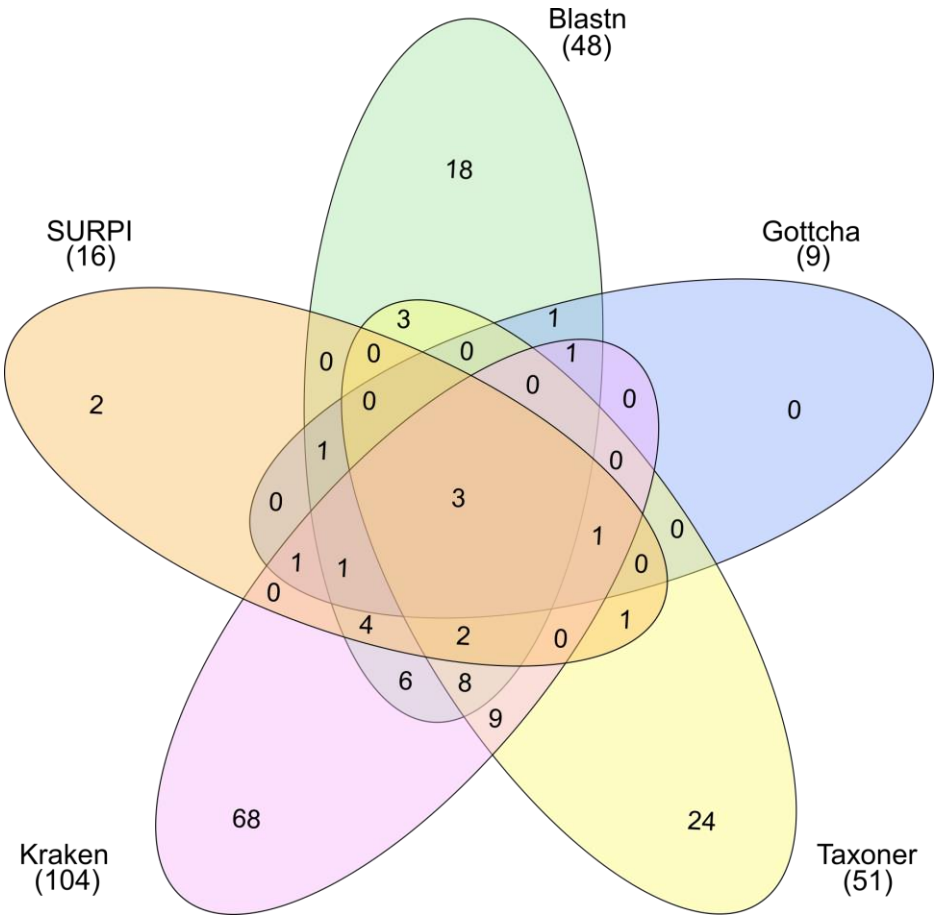


Figura 8: Diagrama representando o número de vírus identificados pelos programas de análise taxonômica a partir dos pools da primeira corrida.

Na segunda corrida, foram identificados 540 gêneros de bactérias (Figura 9) e 217 vírus (incluindo fagos) (Figura 10). Cada organismo identificado em cada *pool* por cada programa utilizado encontra-se nas Tabelas Suplementares (S4 e S5).

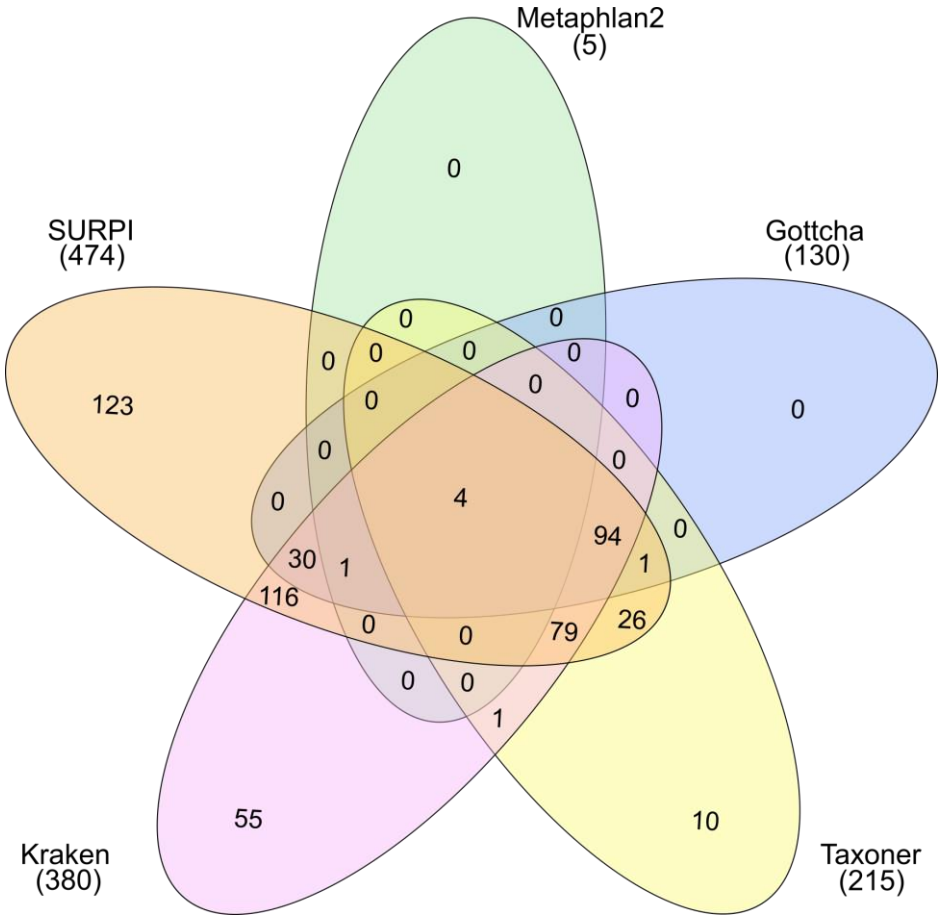


Figura 9: Diagrama representando o número de gêneros de bactérias identificados pelos programas de análise taxonômica a partir dos pools da segunda corrida.

Na figura 10, estão representados os resultados gerados por cinco dos seis programas aplicados, já que ocorreu o mesmo da análise da análise anterior com o programa Metaphlan2. No controle positivo dessa corrida, um soro de paciente com a febre chikungunya, todos os programas identificaram *reads* deste vírus, inclusive, foi possível recuperar o genoma completo do mesmo. Possivelmente o vírus chikungunya estava super-representado na amostra. O genoma completo do vírus chikungunya, assim como análises subsequentes, veio a consubstanciar a publicação no anexo 1. Além disso, encontramos *reads* com similaridade ao vírus dengue 4 no *pool* 22. Provavelmente essa amostra apresentava uma carga viral tão baixa que o vírus não foi previamente identificado por RT-PCR.

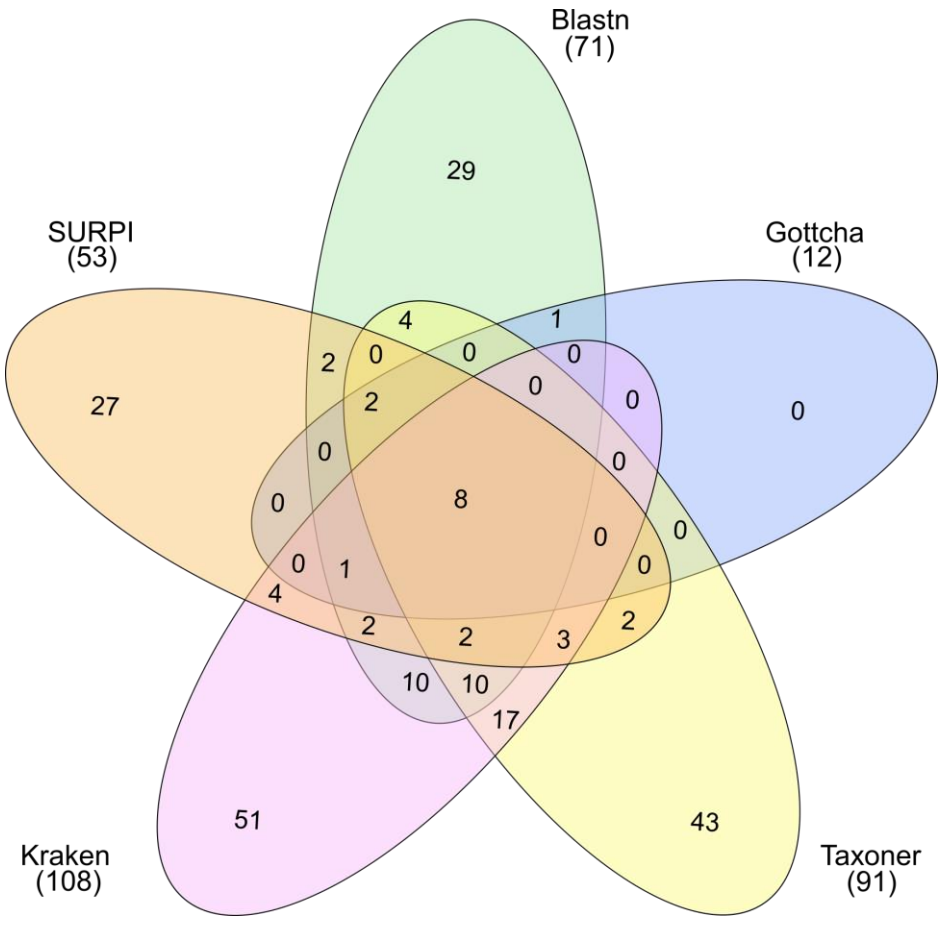


Figura 10: Diagrama representando o número de vírus identificados pelos programas de análise taxonômica a partir dos pools da segunda corrida.

Também comparamos os resultados obtidos a partir dos dados gerados em cada corrida, pois já que foram sequenciadas as mesmas amostras, pudemos verificar a semelhança dos resultados obtidos em cada uma. Na figura 11 e 12 estão apresentados os números de gêneros de bactérias e vírus identificados nas duas corridas e quantos deles foram identificados em ambas corridas.

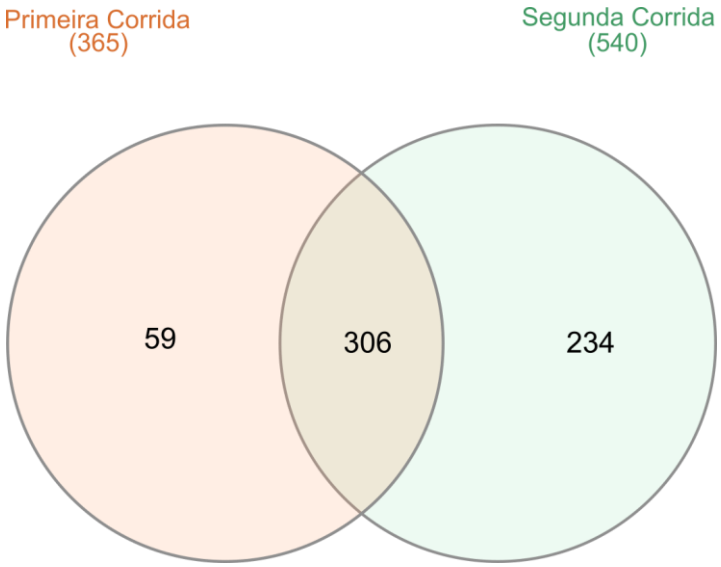


Figura 11: Gêneros de bactérias identificados nas duas corridas.

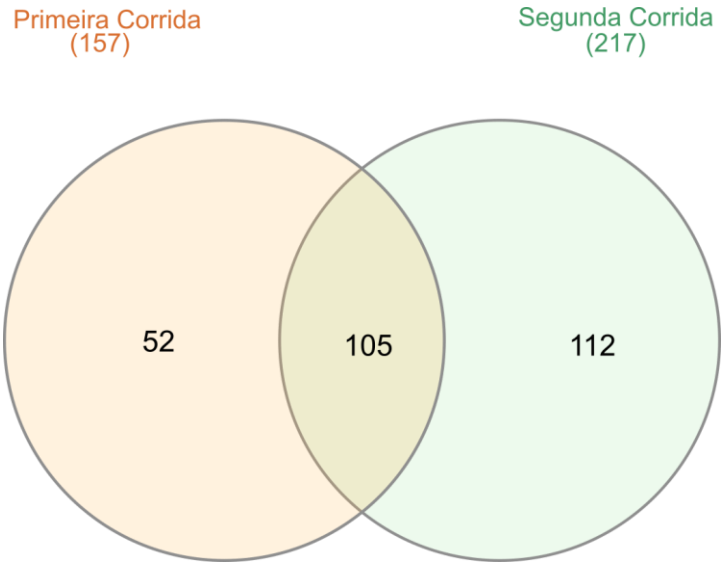


Figura 12: Vírus identificados nas duas corridas.

Apesar de diversos gêneros de bactérias e vírus terem sido identificados, a maior parte das *reads* não foi classificada por nenhum dos programas de análise taxonômica utilizando bases de dados de bactérias e vírus. O programa Kraken, por exemplo, utilizando seu banco de dados *standard*, não conseguiu classificar entre 71,87% a 99,65% das *reads* de cada *pool*.

4.3 Análises específicas in vitro para comprovação dos resultados in silico.

Nas análises *in silico*, foram identificados vírus e bactérias como possíveis agentes infecciosos presentes nas amostras clínicas. Dentre estes organismos identificados *in silico,* alguns foram selecionados para comprovar *in vitro* a sua presença. Foram selecionados organismos identificados em apenas um dos *pools* por mais de um dos programas de análise taxonômica. Entre esses, estão alguns que tiveram identificação taxonômica a partir de um grande número de *reads* (>1000), assim como organismos com um número reduzido *reads* identificadas (<10).

Aplicando esta lógica, selecionamos quatro vírus para comprovação *in vitro*: Parvovírus B19 (B19V), vírus da Hepatite A (HAV), vírus da Hepatite G (HGV) e vírus Torque-Teno (TTV). Esses vírus foram identificados *in silico* nos *pools* 04, 20, 22 e 24, respectivamente, e os quatro foram comprovados *in vitro* (Tabela 1). A partir das *reads* da amostra 2929/11 do pool 04, positiva para B19V, foi possível recuperar o genoma completo desse vírus. Esse genoma completo e análises subsequentes consubstanciaram a publicação no anexo 2.

O HAV foi positivo em duas amostras (1018/12 e 1019/12) do pool 20, que era formado por amostras de pacientes que tiveram suspeita clínica de febre amarela. Neste caso, trabalhamos com poucas *reads*, mas as mesmas possibilitaram que o vírus fosse identificado pelos programas de análise taxonômica e uma região de 580 bp do genoma viral fosse recuperada.

Em relação ao HGV e TTV, embora tenham sido recuperadas um número limitado de *reads* para ambos os vírus na segunda corrida e nenhuma *read* com

similiaridade a esses vírus na primeira corrida, três amostras do pool 22 foram positivas para HGV (417/10, 3671/13 e 2201/12), e duas amostras do pool 24 foram positivas para o TTV (5034/13 e 83/14).

Em relação a bactérias, um número elevado de *reads* com similaridade a *Streptococcus pneumoniae,* foi identificado no *pool* 03. Assim, aplicamos identificação específica *in vitro* para este organismo e uma amostra (01/09) foi positiva. A partir dos resultados da segunda corrida, notamos que uma grande quantidade de *reads* similares a *Neisseria menininginitidis* e *S. pneumoniae* foram recuperadas no *pool* 22 e 23, respectivamente. Ambas as bactérias foram confirmadas *in vitro*: a amostra 417/10 do *pool* 22 foi positiva para *N. menininginitidis* e a amostra 5073/13 foi positiva para *S. pneumoniae.*

A lógica de selecionar organismos identificados em apenas um dos *pools* por mais de um dos programas, não foi aplicada a todos os casos. Análises *in silico* demonstraram que todos os *pools* da primeira corrida apresentavam *reads* com identidade a *N. menininginitidis*, o que poderia ser indicativo de contaminação, por estarem em todos os *pools*. Entretanto, o pool 09 apresentava uma característica distinta, já que um número expressivo de *reads* similares a esse organismo (~ 14.000 reads) foi identificado, enquanto os outros pools apresentaram de 27 a 1600 *reads*. Assim, *N. menininginitidis* também foi selecionada para comprovação *in vitro* no *pool* 09. Esta confirmação foi positiva para amostra 201/14 desse *pool.*

Como na primeira corrida o número de *reads* foi reduzido em comparação com os da segunda corrida, alguns dos organismos identificados em amostras que foram analisadas nas duas corridas, apresentaram distintos números de *reads* nas duas situações, sendo que sempre, os dados da segunda corrida foram mais robustos.

Na tabela 2 estão discriminados os organismos confirmados, o número dos *pools* em que as amostras positivas faziam parte e a quantidade de *reads* identificadas por cada programa de análise taxonômica utilizada. Nessa tabela também se encontram os resultados recuperados a partir do vírus chikungunya do *pool* 21, utilizado como controle positivo. Os vírus HGV e TTV não foram identificados por nenhum programa em nenhum dos *pools* da primeira corrida as quais as amostras positivas faziam parte, por isso, não foram adicionados à tabela.

Tabela 2: R*eads* identificadas pelos programas de análise taxonômica para cada um dos organismos testados *in vitro*. (M): Metaphlan2; (G): Gottcha; (K): Kraken; (S): Surpi; (T): Taxoner; (B): Blastn.

Organismos Comprovados	Pool	M	G	K	S	T	B
Parvovirus B19	04	Sim	928	1.672	655	1.269	1.109
Hepatite A virus	20	Sim	0	13	12	6	8
Hepatite G virus	22	Sim	0	0	6	3	0
Torque-Teno virus	24	Sim	0	0	5	0	0
S. pneumoniae	03	Sim	3.047	21.328	15.453	15.823	X
N. meningitidis	09	Sim	3.624	24.786	14.425	17.434	X
S. pneumoniae	23	Sim	1.264	6.810	11.821	3.902	X
	09	Não	185	1.247	551	1.060	X
N. meningitidis	22	Sim	166	1.126	1.966	531	X
	16	Sim	105	495	1.206	402	X
Chikungunya virus	21	Sim	387	693	17.986	611	708

Na tabela 3, estão apresentadas a quantidade de bases recuperadas pelo BWA, o número de acesso (GenBank) e o tamanho do genoma do organismo referência. Nessa tabela também se encontram os resultados recuperados a partir do vírus chikungunya utilizado como controle positivo.

Tabela 3: Dados *in silico* dos patógenos confirmados.

			Referência	
	Pools	Bases recuperadas	GenBank	Tamanho (bp)
Parvovirus B19	04	5.594	FN598217	5.594
Hepatite A virus	20	580	EU526088	7.451
Hepatite G virus	22	171	NC_001710	9.392
Torque-Teno virus	24	62	EU305674	2.809
S. pneumoniae	03	489.712	CP000936	2.245.615
N. meningitidis	09	522.271	NC_003112	2.272.360
S. pneumoniae	23	225.792	CP000936	2.245.615
N. meningitidis	22	91.820	NC_003112	2.272.360
Chikungunya virus	21	10.654	KJ451624	12.011

4.4 Análises genéticas

A análise genética das amostras positivas para *N. meningitidis* foi baseada na região determinante dos sorogrupos. Foram utilizados iniciadores específicos para os cinco sorogrupos (A, B, C, W135 e Y) relacionados a mais de 90% dos casos da infecção por essa bactéria no mundo. Ambas as amostras foram positivas para o sorogrupo C.

Em relação aos vírus identificados, utilizamos tanto sequências recuperadas pelo HTS quanto sequências obtidas pela análise *in vitro*, para identificação de genótipos. As sequências das duas amostras positivas para TTV apresentaram uma qualidade ruim e não foram suficientes para construção da árvore filogenética, porém, os resultados das análises por BLAST identificaram apenas TTV.

Utilizando o genoma completo do B19V, geramos uma árvore filogenômica sob o modelo TrN+G. Existem três genótipos de B19V circulando no mundo e o genoma da amostra 2929/11 foi agrupada ao genótipo 1. (Figura 13)

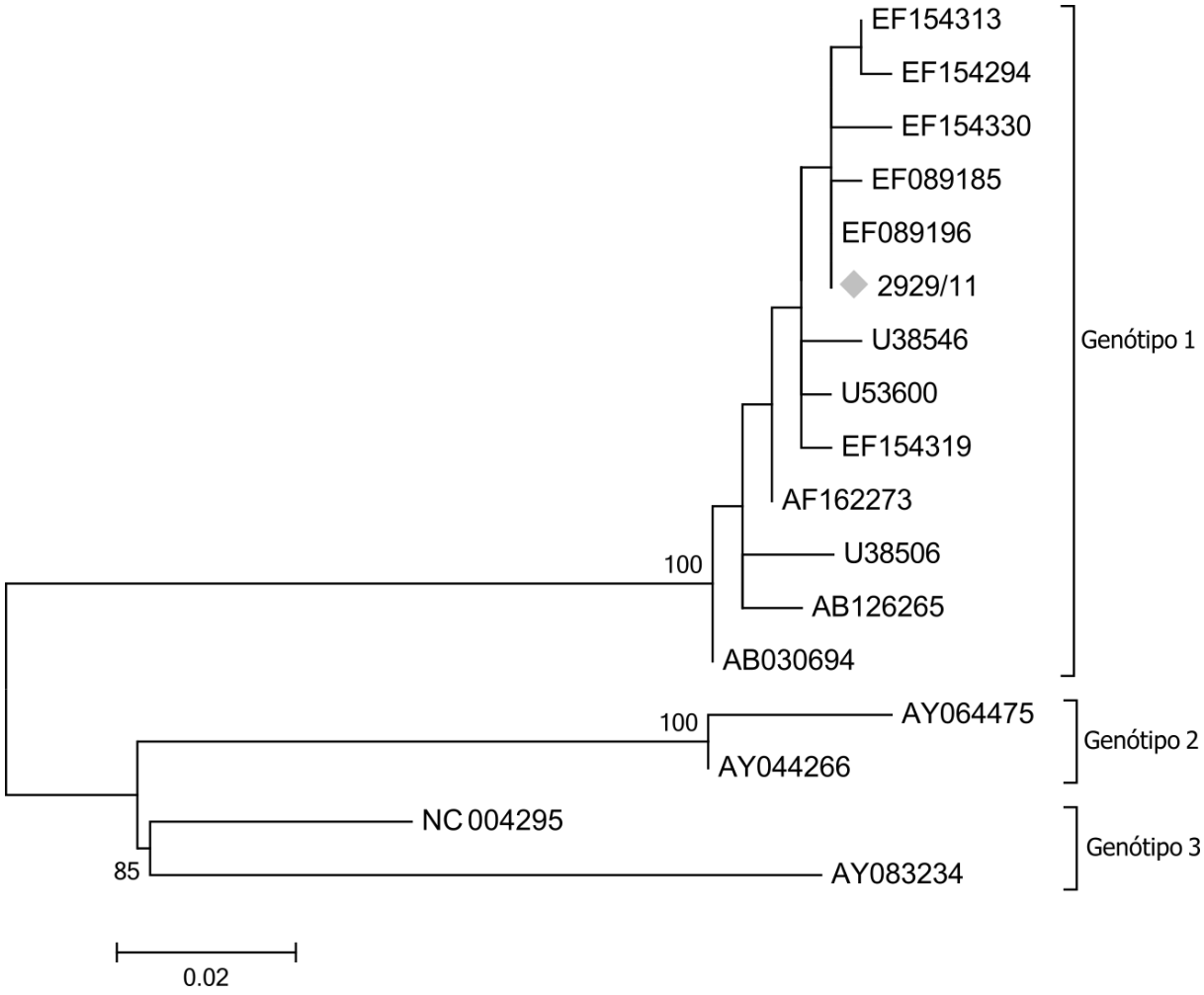


Figura 13: Árvore filogenômica do Parvovirus B19 com os três genótipos indicados. A sequência recuperada nesse estudo está representada por um losango cinza e a escala representa o número de substituições por sítio. Valores de bootstrap acima de 75 estão apresentados.

Em relação ao vírus Chikungunya, também existem três genótipos circulando no mundo. Utilizando o genoma completo recuperado nesse estudo (RJ/CHIKV/2015), geramos uma árvore filogenômica sob o modelo GTR+I. O genoma RJ/CHIKV/2015 pertence ao genótipo Asiático. (Figura 14)

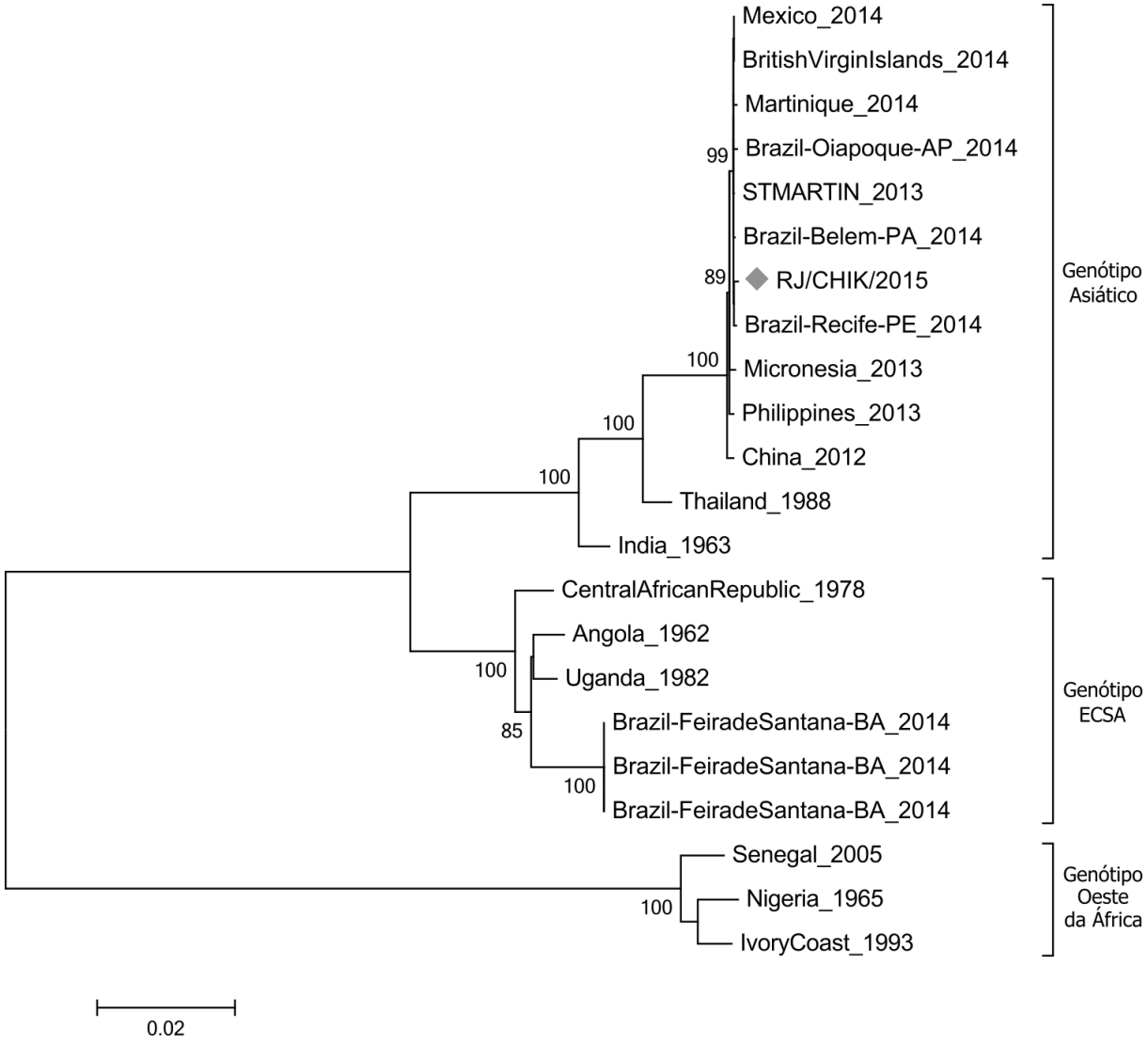


Figura 14: Árvore filogenômica do vírus Chikungunya com os três genótipos indicados. Genótipo ECSA refere-se ao genótipo East/Central/South Africa, encontrado no leste, no sul e na região central da África. A sequência recuperada nesse estudo está representada por um losango cinza e a escala representa o número de substituições por sítio. Valores de bootstrap acima de 75 estão apresentados.

Baseada no gene 5’ UTR do vírus da Hepatite A, geramos uma análise filogenética com os três genótipos previamente relatados sob o modelo HKY+I. As amostras sequenciadas nesse estudo (1018/12 e 1019/12) agruparam-se ao subgenótipo 1A. (Figura 15)

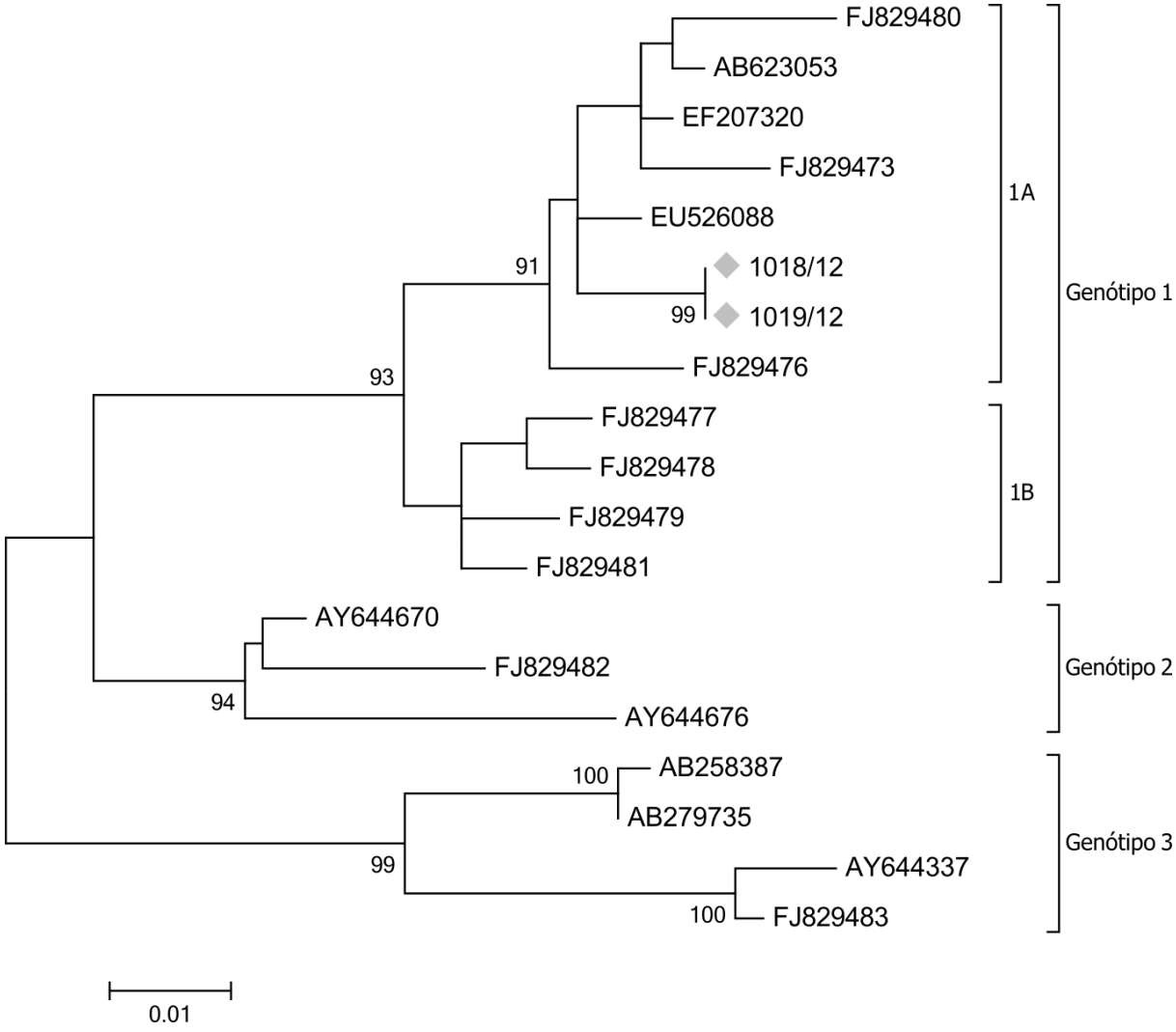


Figura 15: Árvore filogenética do vírus da Hepatite A com genótipos e os subgenótipos indicados. As sequências recuperadas nesse estudo estão representadas por um losango cinza e a escala representa o número de substituições por sítio. Valores de bootstrap acima de 75 estão apresentados.

Também baseada no gene 5’ UTR, geramos uma análise filogenética com os seis genótipos do vírus da Hepatite G sob o modelo GTR+I+G. As amostras sequenciadas nesse estudo (417/10, 2201/12 e 3671/13), agruparam-se ao genótipo 2. (Figura 16)

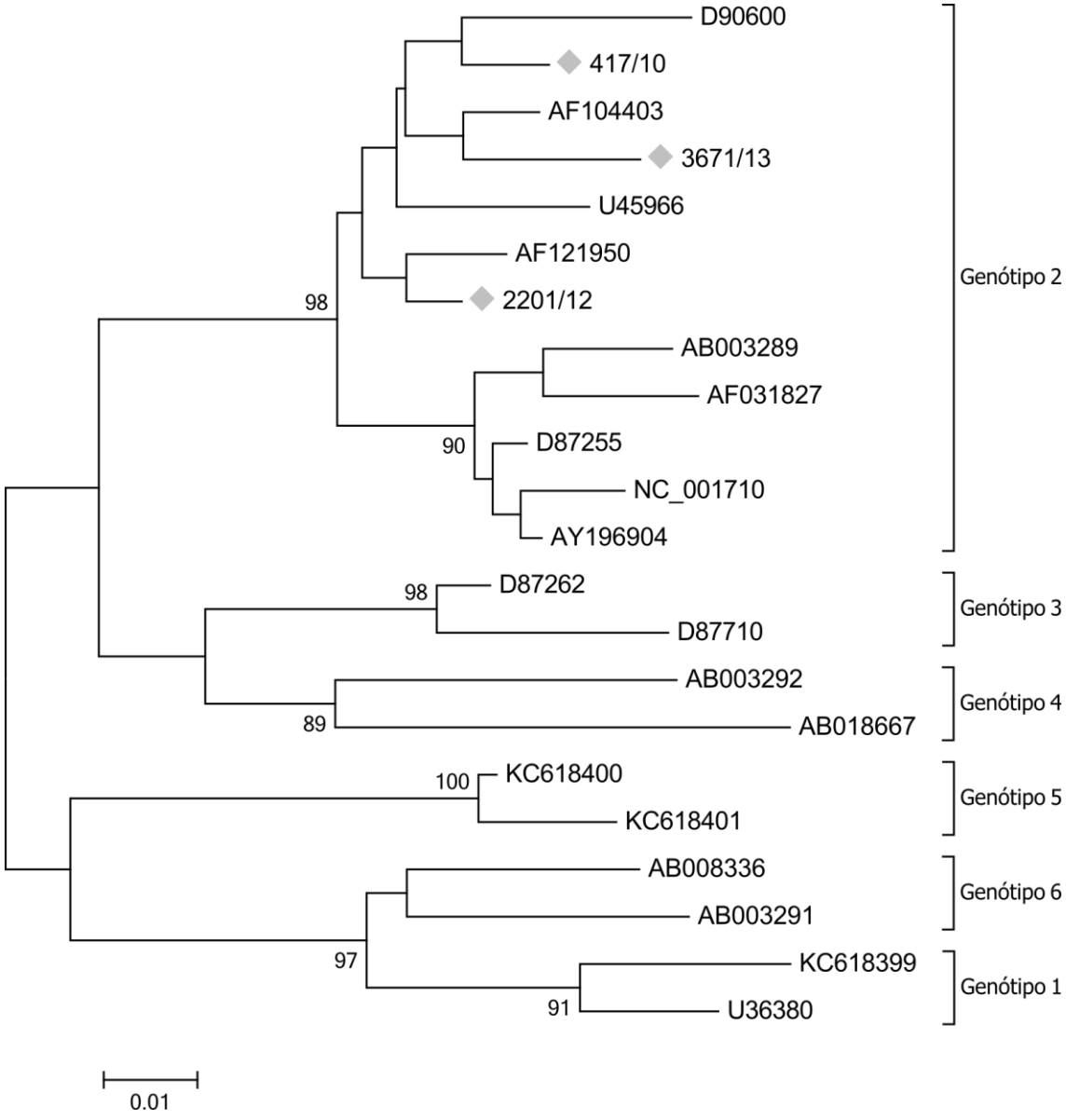


Figura 16: Árvore filogenética do vírus da Hepatite G com os seis genótipos indicados. As sequências recuperadas nesse estudo estão representadas por um losango cinza e a escala representa o número de substituições por sítio. Valores de bootstrap acima de 75 estão apresentados.

5 DISCUSSÃO

Nesse estudo, a abordagem de metagenômica foi aplicada para elucidar organismos presentes em amostras de pacientes com suspeita clínica de dengue ou febre amarela, mas com diagnóstico laboratorial negativo para ambos os vírus. Um dos principais desafios no campo da metagenômica é a identificação precisa dos organismos presentes em uma amostra, o que normalmente é feito com base na análise taxonômica de *reads* curtas. A combinação de bases de dados incompletas, a semelhança entre genomas divergentes e erros das tecnologias de sequenciamento dificulta essa análise (36).

Em nosso estudo, utilizamos seis programas para classificar as *reads* recuperadas a partir de *pools* de amostras clínicas humanas, e uma vez que cada um deles apresenta algoritmo e bases de dados distintas, também gerou resultados distintos. Os programas Metaphlan2 e GOTTCHA foram os que identificaram menos organismos em comparação com os outros programas utilizados. Métodos que utilizam bases de dados formadas por genes marcadores ou segmentos únicos de genomas, como as utilizadas pelo Metaplan2 e GOTTCHA, geralmente são restritos na identificação de organismos, uma vez que a diversidade existente na natureza e na amostra, podem não estar contemplados naquela base de dados (74). Mas por outro lado, quando um organismo é identificado por esses programas o grau de confiança é elevado. O programa GOTTCHA, além de utilizar uma base de dados formada de segmentos únicos de genomas de vírus e bactérias, filtra os organismos seguindo diversos parâmetros, como os organismos com menos de 10 *hits* e cobertura abaixo de 0,5% (36). Por isso, apresenta restrições para identificar organismos que apresentam uma carga baixa, assim como organismos que podem estar emergindo (4). Patógenos que estão pouco representados na amostra é um problema para resultados que se baseiam unicamente em análises *in silico*. Entretanto, nossos resultados mostram que nem sempre a baixa carga impede a identificação do organismo, uma vez que a análise *in silico* é uma evidência preliminar que pode ser comprovada *in vitro.* Por exemplo, em um estudo no qual foram realizadas três repetições de cinco agentes patogênicos em diferentes cargas, com três ordens de grandeza cada, GOTTCHA foi capaz de identificar todos os organismos em média e alta concentração. No entanto, dos cinco organismos,

apenas um em baixo título foi identificado (4). Em nossos resultados, o GOTTCHA não identificou HAV, HGV e TTV.

Com o Kraken e o pipeline SURPI, utilizamos uma base de dados a partir da última versão do GenBank, o banco de dados mais atualizado e abrangente, e isso implicou na identificação de mais organismos. Além disso, muitos resultados foram concordantes entre eles, o que aumenta a confiança nos organismos identificados. Já o Taxoner, disponibiliza uma base de dados pré-indexada que foi criada em 2014 (34). O Blastn identificou menos vírus que o Kraken, isso pode ser consequência, mais uma vez, da base de dados utilizada, já que com o Blastn, utilizamos uma base de dados limitada aos vírus referência do NCBI, enquanto o Kraken utiliza o banco completo. Blastn é um algoritmo de busca local, e por gerar *seeds* de apenas 11 bases, alcança uma maior sensibilidade, mas perdendo especificidade. O programa Kraken, por exemplo, que requer similaridade de pelo menos 31 bases, é mais específico (32,33). Além disso, SURPI e BLAST não aceitam *reads* pareadas, não tendo a capacidade de mapear as *reads* a uma referência utilizando a informação da distância entre elas. Essa informação influencia dramaticamente, por exemplo, na resolução de rearranjos estruturais (inserções, deleções, inversões), bem como a montagem de regiões repetitivas.

A utilização de diversos programas de análise taxonômica é essencial, já que cada um deles apresenta vantagens e desvantagens. Além disso, como não há uma amplitude ou profundidade de cobertura necessária para fazer uma "identificação", a identificação de uma série de *reads* específicas de um organismo é importante para considerar que um resultado seja verdadeiro. Porém, quantas *reads* são necessárias para ter essa confiança pode variar de acordo com tipo de amostra ou organismo (4). Por exemplo, em nossos resultados, menos de 15 *reads* com similaridade ao HAV foram identificadas no *pool* 20 e confirmamos que duas amostras deste *pool* de fato continham esse vírus. O mesmo ocorreu quando o RNA total de cortes de biópsia de tecido hepático infectados naturalmente pelo vírus da hepatite C (HCV) foi sequenciado por Illumina, e apenas quatro *reads* virais por milhão foram identificadas (75).

Determinar a etiologia de doenças humanas com sintomas similares à infecção por dengue e febre amarela é fundamental, já que alguns desses

patógenos podem ter o potencial de causar surtos e/ou epidemias, mas não estão sendo considerados pela saúde pública. Nesse estudo, identificamos *reads* do DENV-4 em apenas um pool e nenhuma *read* similar ao YFV. No entanto, diversos patógenos que poderiam ser os determinantes das manifestações clínicas apresentadas pelos casos fatais suspeitos de dengue. Dentre esses, estão: B19V, *N. meningitidis* e *S. pneumoniae*, enquanto que o HAV pode ter sido o vírus causador das manifestações clínicas que levaram a suspeita de febre amarela. É importante ressaltar que a identificação *in silico,* assim como a confirmação da presença *in vitro* de certo organismo numa amostra clínica não é suficiente para afirmar que este foi a causa do quadro e do desfecho clínico. Contudo, se o organismo identificado é um patógeno e evidências clínico-epidemiológicas indicam essa associação, o diagnóstico pode ser definitivo.

Três dos vírus identificados nesse estudo, B19V, HGV e TTV, também foram identificados por metagenômica em amostras de soro de casos suspeitos de dengue da Nicarágua (40). B19V está associado a casos de doenças exantemáticas leves que afetaria crianças, em adultos, causa dor e inchados nas articulações e anemia grave, e pode causar morte de fetos. Esse vírus é classificado em três genótipos distintos e os três já foram identificados no Brasil (76). Neste estudo, por metagenômica, recuperamos o primeiro genoma completo de um B19V genótipo 1 do Brasil (77). A amostra a qual estava associado é de um dos casos fatais de 2011, de um paciente do sexo masculino com 12 anos de idade que apresentou sintomas como febre, hemorragia, trombocitopenia, hematócrito baixo (26%), extravasamento plasmático evidenciado por derrames cavitários, edema periorbital, hepatomegalia, oligúria e choque. Nos dados de metagenômica do *pool* que continha essa amostra, o patógeno com o maior número de *reads* foi o B19V.

A infecção pelo HGV ou TTV tem sido associada a muitas doenças. Por metagenômica, identificamos três amostras de um mesmo *pool* positivas para HGV. Essas amostras eram do Rio de Janeiro dos anos de 2010, 2012 e 2013 e os pacientes apresentaram febre, mialgia e hemorragias, sendo que um evoluiu para insuficiência renal aguda. Genótipos 1 e 2 do HGV estão presentes no sudeste e nordeste do Brasil, sendo o tipo 2 relativamente predominante em ambos os casos (78). Os HGV presentes nas três amostras pertencem ao genótipo 2. Quanto ao

TTV, duas amostras de um mesmo *pool foram* positivas. Ambas são de pacientes mulheres do Espírito Santo que apresentaram insuficiência hepática, e uma delas, de 19 anos, apresentou pulmão de choque, edema cerebral e sepse. Enquanto que a paciente de 50 anos manifestou sintomas de mialgia, artralgia, náusea/vomito, icterícia, hematúria, choque séptico. Esses vírus, assim como essas sintomatologias, estão presentes na literatura, mas em nenhum dos casos, se fez a relação causal entre doenças clínicas específicas e infecção por um dos dois vírus (79,80).

Já o HAV, é uma causa importante de hepatite aguda transmitida principalmente por via fecal-oral através de alimentos ou água contaminada ou por contatos pessoais. A infecção tem um caráter endêmico no Brasil e em 2012, ano da coleta das amostras sequenciadas em nosso estudo, foram registrados aproximadamente 600 casos de HAV no Brasil, e inclusive, essa virose está relacionada a casos de hepatite fulminante (81). No nosso estudo, no *pool* de amostras de pacientes com suspeita de febre amarela, duas amostras foram positivas para HAV, embora tenhamos recuperado menos de vinte *reads* correspondente a esse vírus nesse *pool.* HAV é classificado em três genótipos e ambas amostras pertencem ao sub genótipo 1A, único genótipo já relatado na América do Sul (82). Os pacientes relataram náusea e desconforto abdominal, um deles, com cinco dias de doença, apresentou plaquetopenia, leucopenia, hepatoesplenomegalia, enquanto a outra paciente, com um dia de doença, apresentou sintomas como febre, dor retro-orbitaria e mialgia. Ambos não evoluíram para óbito.

A metagenômica também já permitiu a recuperação de diversas bactérias patogênicas a partir de diferentes tipos de amostras clínicas humanas (83). Essa abordagem foi usada, por exemplo, para identificação da bactéria responsável por um surto de diarréia na Alemanha analisando amostras de fezes dos pacientes (48). Em nosso estudo, quatro amostras foram positivas para bactérias, duas para *N. meningitidis* e outras duas para *S. pneumoniae.* Essas bactérias são patógenos associados a surtos e epidemias, onde quadros graves, como pneumonia, septicemia e meningite, podem evoluir para morte (84–86). As duas amostras positivas para *N. meningitidis* foram identificadas por metagenômica, nos *pools* 09 e 22, sendo que no *pool* 09 foram recuperadas ~15.000 *reads,* e no outro foram

recuperadas ~1.500 *reads. In vitro,* foi determinado que *a*mbas pertenciam ao sorogrupo C, sorogrupo prevalente nos surtos e epidemias no Brasil, desde 2006, atingindo ínidices alarmantes entre 2010 e 2014 (87). Um dos casos era de um paciente de 20 anos procedente de Vitória-ES, que foi a óbito no ano de 2014, horas após ter apresentado sintomas de uma síndrome febril aguda e edema agudo pulmonar. O outro caso foi de um paciente do Rio de Janeiro, que foi a óbito em 2009 e apresentou sintomas como febre, hemorragia digestiva, choque hipovolêmico, petéquias generalizadas e mialgia. Essas sintomatologias estão associadas a infecções por *N. meningitidis*, que, portanto*,* pode ter sido o agente etiológico determinante dos óbitos.

Quanto aos dois casos positivos para *S. pneumoniae*, um identificado no *pool 03* e o outro no *pool 23,* são de óbitos que ocorreram em 2009 e 2013. No caso de 2009, a paciente foi a óbito dois dias após o aparecimento dos sintomas: febre, trombocitopenia, vômitos, diarréia e sua declaração de óbito relata parada cardio-respiratória, enquanto que no segundo caso, o paciente apresentou febre, prostração, cefaléia e sudorese intensa. Essas sintomatologias estão associadas a infecções por *S. pneumoniae* e, portanto*,* esse pode ter sido o agente etiológico determinante dos óbitos.

Outra parte das *reads* que mapearam tanto com bactérias quanto vírus, tinha a seguinte característica: Organismos que estariam presentes em todos os *pools.* Entre esses organismoss estão bactérias dos gêneros *Acinetobacter, Bradyrhizobium, Burkholderia, Corynebacterium, Klebsiella, Legionella*, *Methylobacterium, Micrococcus, Mycobacterium, Neisseria, Propionibacterium, Pseudomonas, Ralstonia, Serratia, Sphingomonas*, *Stenotrophomonas e Vibrio.* Alguns deles apresentaram uma prevalência similar de *reads,* por exemplo, de ~10 a 100 *reads* similares a *Corynebacterium* foram identificadas em cada *pool* da primeira corrida. Portanto, no nosso trabalho, não realizamos a comprovação *in vitro* dos mesmos. Alguns motivos podem ser erros de sequenciamento e/ou erro de análise de imagem durante a fase de sequenciamento do índice, assim, os *amplicons* de outras bibliotecas que foram sequenciadas nas mesmas corridas que nossas amostras foram incorretamente atribuídos ao índice correspondente as nossas (88). Além disso, também identificamos *reads* similares a bactérias comuns de

contaminarem sequenciamentos de alto-desempenho, que poderiam estar na água utilizada no laboratório e em reagentes utilizados no processamento das amostras.

A presença de uma grande porção de *reads* nos nossos resultados que não mapearam contra nenhuma sequência presente nas bases de dados não é algo incomum de ocorrer em projetos de metagenômica. Por exemplo, Oh e colaboradores aplicaram metagenômica em amostras de pele humana, onde obtiveram de 2 a 96% de *reads* não mapeadas (89). Assim como 92% dos c*ontigs* gerados a partir de amostras de pacientes com doença respiratória e febril, não puderam ser classificados (90). A diversidade das amostras é revelada pelos metagenomas e como sabemos, apenas 0,1-10% dos microrganismos visualizados no microscópio podem ser isolados em meios artificiais (91). Em relação aos possíveis patógenos, a mesma realidade se dá, principalmente se são organismos eventuais e/ou emergentes (92–94). Portanto, as *reads* não mapeadas podem representar organismos que ainda não foram caracterizados ou não apresentam seu genoma completo sequenciado (95).

6 CONCLUSÕES

- Pela abordagem de metagenômica, não foi identificado a presença dos vírus dengue e febre amarela na maioria das amostras suspeitas e não confirmadas para estas infecções.
- A abordagem de metagenômica mostrou-se robusta para a pesquisa de agentes infecciosos em amostras de pacientes com suspeita clínica de dengue e febre amarela, mas com diagnóstico laboratorial negativo.
- Em alguns dos casos fatais com suspeita clínica de dengue ou febre amarela foi identificada a presença dos seguintes agentes infecciosos: Parvovirus B19, vírus da Hepatite A, *N. meningitidis e S. pneumoniae.*

7 REFERÊNCIAS BIBLIOGRÁFICAS

1. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998 Oct;5(10):R245–9.
2. Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson R V. Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon Cenarchaeum symbiosum. J Bacteriol. 1998 Oct;180(19):5003–9.
3. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004 Dec;68(4):669–85.
4. Frey KG, Herrera-Galeano JE, Redden CL, Luu T V, Servetas SL, Mateczun AJ, et al. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. BMC Genomics. 2014 Jan;15:96.
5. Yang J, Yang F, Ren L, Xiong Z, Wu Z, Dong J, et al. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. J Clin Microbiol. 2011 Oct;49(10):3463–9.
6. Kim M, Lee K-H, Yoon S-W, Kim B-S, Chun J, Yi H. Analytical tools and databases for metagenomics in the next-generation sequencing era. Genomics Inform. 2013 Sep;11(3):102–13.
7. Garza DR, Dutilh BE. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Cell Mol Life Sci. 2015 Aug 9;72(22):4287–308.
8. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci. 2014 Jan;5:209.
9. Hall RJ, Draper JL, Nielsen FGG, Dutilh BE. Beyond research: a primer for considerations on using viral metagenomics in the field and clinic. Front

Microbiol. 2015 Jan;6:224.

10. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012 Jan;2(1):3.

11. Teeling H, Glöckner FO. Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. Brief Bioinform. 2012 Dec;13(6):728–42.

12. Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. PLoS One. 2013 Jan;8(10):e77910.

13. Illumina. HiSeq 2500 Sequencing System. [Internet]. [cited 2015 Nov 15]. Available from: http://www.illumina.com/documents/products/datasheets/datasheet_hiseq2500.pdf

14. Peng X, Wang J, Zhang Z, Xiao Q, Li M, Pan Y. Re-alignment of the unmapped reads with base quality score. BMC Bioinformatics. 2015 Jan;16 Suppl 5:S8.

15. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, et al. Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Syst. 2015 Feb 29;1(1):72–87.

16. Sangwan N, Lata P, Dwivedi V, Singh A, Niharika N, Kaur J, et al. Comparative metagenomic analysis of soil microbial communities across three hexachlorocyclohexane contamination levels. PLoS One. Public Library of Science; 2012 Jan;7(9):e46219.

17. Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code Biol Med. 2014 Jan;9:8.

18. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011. p. pp. 10–2.

19. FASTX-Toolkit [Internet]. [cited 2016 Apr 9]. Available from: http://hannonlab.cshl.edu/fastx_toolkit/

20. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic

datasets. Bioinformatics. 2011 Mar 15;27(6):863–4.

21. Schmieder R, Lim YW, Rohwer F, Edwards R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics. 2010 Jan;11:341.

22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Apr 28;30(15):2114–20.

23. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. 2014;24(7):1180–92.

24. Moore RA, Warren RL, Freeman JD, Gustavsen JA, Chénard C, Friedman JM, et al. The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. PLoS One. 2011 Jan;6(5):e19838.

25. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol. 2011 May;29(5):393–6.

26. Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int. 2014 Jan;2014:309650.

27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

28. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.

29. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. Faster and More Accurate Sequence Alignment with SNAP. 2011 Nov 23;

30. Neelakanta G, Sultana H. The use of metagenomic approaches to analyze changes in microbial communities. Microbiol insights. 2013 Jan;6:37–48.

31. Higashi S, Barreto A da MS, Cantão ME, de Vasconcelos ATR. Analysis of composition-based metagenomic classification. BMC Genomics. 2012 Jan;13

Suppl 5:S1.

32. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Jan;15(3):R46.

33. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. BioMed Central Ltd; 2009 Jan 15;10(1):421.

34. Pongor LS, Vera R, Ligeti B. Fast and sensitive alignment of microbial whole genome sequencing reads to large sequence datasets on a desktop PC: application to metagenomic datasets and pathogen identification. PLoS One. 2014 Jan;9(7):e103441.

35. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015 Sep 29;12(10):902–3.

36. Freitas TAK, Li P-E, Scholz MB, Chain PSG. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. 2015 Mar 12;gkv180 – .

37. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 2011 Jan;12 Suppl 2:S4.

38. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics. 2013 Sep 15;29(18):2253–60.

39. MacDonald NJ, Parks DH, Beiko RG. Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res. 2012 Aug;40(14):e111.

40. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis. 2012 Jan;6(2):e1485.

41. Xu B, Liu L, Huang X, Ma H, Zhang Y, Du Y, et al. Metagenomic analysis of

fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. PLoS Pathog. 2011 Nov;7(11):e1002369.

42. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, et al. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. PLoS Pathog. 2012 Sep;8(9):e1002924.

43. Smits SL, Zijlstra EE, van Hellemond JJ, Schapendonk CME, Bodewes R, Schürch AC, et al. Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. Emerg Infect Dis. 2013 Jan;19(9).

44. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med. 2008 Mar 6;358(10):991–8.

45. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med. 2014 Jun 19;370(25):2408–17.

46. Naccache SN, Peggs KS, Mattes FM, Phadke R, Garson JA, Grant P, et al. Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. Clin Infect Dis. 2015 Mar 15;60(6):919–23.

47. Hoffmann B, Tappe D, Höper D, Herden C, Boldt A, Mawrin C, et al. A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis. N Engl J Med. 2015 Jul 9;373(2):154–62.

48. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA. American Medical Association; 2013 Apr 10;309(14):1502–10.

49. Greninger AL, Chen EC, Sittler T, Scheinerman A, Roubinian N, Yu G, et al. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PLoS One. 2010 Jan;5(10):e13381.

50. Moore NE, Wang J, Hewitt J, Croucher D, Williamson DA, Paine S, et al.

Metagenomic analysis of viruses in feces from unsolved outbreaks of gastroenteritis in humans. J Clin Microbiol. 2015 Jan;53(1):15–21.

51. Gao D, Yu Q, Wang G, Wang G, Xiong F. Diagnosis of a malayan filariasis case using a shotgun diagnostic metagenomics assay. Parasit Vectors. 2016 Jan;9(1):86.

52. Tumioto GL, Gregianini TS, Dambros BP, Cestari BC, Alves Nunes ZM, Veiga ABG. Laboratory surveillance of dengue in Rio Grande do Sul, Brazil, from 2007 to 2013. PLoS One. Public Library of Science; 2014 Jan 8;9(8):e104394.

53. WHO | Dengue guidelines for diagnosis, treatment, prevention and control: new edition. World Health Organization;

54. SOPER FL. The elimination of urban yellow fever in the Americas through the eradication of Aedes aegypti. Am J Public Health Nations Health. 1963 Jan;53:7–16.

55. Almeida MAB, Cardoso J da C, Dos Santos E, da Fonseca DF, Cruz LL, Faraco FJC, et al. Surveillance for yellow Fever virus in non-human primates in southern Brazil, 2001-2011: a tool for prioritizing human populations for vaccination. PLoS Negl Trop Dis. 2014 Mar;8(3):e2741.

56. Vasconcelos PF da C. Febre amarela. Rev Soc Bras Med Trop. SBMT; 2003 Apr;36(2):275–93.

57. Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P. Metagenomics for pathogen detection in public health. Genome Med. 2013 Jan;5(9):81.

58. Ministério da Saúde - Secretaria de Vigilância em Saúde.Balanço Dengue Janeiro a Julho de 2007. [Internet]. [cited 2015 Nov 15]. Available from: http://ww2.prefeitura.sp.gov.br//arquivos/secretarias/saude/vigilancia_saude/doenca_agravo/0057/balanco_dengue_jan_jul_2007.pdf

59. Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, et al. A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res. 2009 Dec;37(22):e148.

60. Hang J, Forshey BM, Kochel TJ, Li T, Solórzano VF, Halsey ES, et al. Random amplification and pyrosequencing for identification of novel viral genome

sequences. J Biomol Tech. 2012 Apr;23(1):4–10.

61. Cassinotti P, Weitz M, Siegl G. Human parvovirus B19 infections: routine diagnosis by a new nested polymerase chain reaction assay. J Med Virol. 1993 Jul;40(3):228–34.

62. Endo K, Inoue J, Takahashi M, Mitsui T, Masuko K, Akahane Y, et al. Analysis of the full-length genome of a subgenotype IIIB hepatitis A virus isolate: primers for broadly reactive PCR and genotypic analysis. J Med Virol. 2007 Jan;79(1):8–17.

63. Jarvis L, Davidson F, Hanley J, Yap P, Ludlam C, Simmonds P. Infection with hepatitis G virus among recipients of plasma products. Lancet. Elsevier; 1996 Nov 16;348(9038):1352–5.

64. Kenar Koohi A, Ravanshad M, Rasouli M, Falahi S, Baghban A. Phylogenetic analysis of torque teno virus in hepatitis C virus infected patients in shiraz. Hepat Mon. Kowsar; 2012 Jul 1;12(7):437–41.

65. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. BioMed Central; 2010 Jan 10;11(1):595.

66. Yu X, Guda K, Willis J, Veigl M, Wang Z, Markowitz S, et al. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? BioData Min. 2012 Jan;5(1):6.

67. Malboeuf CM, Yang X, Charlebois P, Qu J, Berlin AM, Casali M, et al. Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. Nucleic Acids Res. 2013 Jan 7;41(1):e13.

68. Ninomiya M, Ueno Y, Funayama R, Nagashima T, Nishida Y, Kondo Y, et al. Use of illumina deep sequencing technology to differentiate hepatitis C virus variants. J Clin Microbiol. 2012 Mar 1;50(3):857–66.

69. Kohl C, Brinkmann A, Dabrowski PW, Radonić A, Nitsche A, Kurth A. Protocol for metagenomic virus detection in clinical specimens. Emerg Infect Dis. 2015 Jan;21(1):48–57.

70. Stremlau MH, Andersen KG, Folarin OA, Grove JN, Odia I, Ehiane PE, et al.

Discovery of novel rhabdoviruses in the blood of healthy individuals from West Africa. PLoS Negl Trop Dis. 2015 Mar;9(3):e0003631.

71. Xu B, Zhi N, Hu G, Wan Z, Zheng X, Liu X, et al. Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing. Proc Natl Acad Sci U S A. 2013 Jun 18;110(25):10264–9.

72. Illumina. Nextera XT Library Prep: Tips and Troubleshooting. 2015 [Internet]. [cited 2016 Jan 17]. Available from: http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-troubleshooting-guide.pdf

73. Zirkel F, Roth H, Kurth A, Drosten C, Ziebuhr J, Junglen S. Identification and characterization of genetically divergent members of the newly established family Mesoniviridae. J Virol. 2013 Jun 27;87(11):6346–58.

74. Bazinet AL, Cummings MP. A comparative evaluation of sequence classification programs. BMC Bioinformatics. 2012 Jan;13:92.

75. Daly GM, Bexfield N, Heaney J, Stubbs S, Mayer AP, Palser A, et al. A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. PLoS One. 2011 Jan;6(12):e28879.

76. Sanabani S, Neto WK, Pereira J, Sabino EC. Sequence variability of human erythroviruses present in bone marrow of Brazilian patients with various parvovirus B19-related hematological symptoms. J Clin Microbiol. 2006 Feb;44(2):604–6.

77. Conteville LC, Zanella L, Marín MA, Filippis AMB de, Nogueira RMR, Vicente ACP, et al. Parvovirus B19 1A complete genome from a fatal case in Brazil. Memórias do Inst Oswaldo Cruz. 2015 Sep;110(6):820–1.

78. Nishiya AS, Ribeiro-Dos-Santos G, Bassit L, Focaccia R, Chamone DF, Sabino EC. Genotype distribution of the GB virus C in citizens of São Paulo City, Brazil. Rev Inst Med Trop Sao Paulo. Instituto de Medicina Tropical de São Paulo; 2003 Aug;45(4):213–6.

79. Brajão de Oliveira K. Torque teno virus: a ubiquitous virus. Rev Bras Hematol

Hemoter. Jan;37(6):357–8.

80. Bhattarai N, Stapleton JT. GB virus C: the good boy virus? Trends Microbiol. 2012 Mar;20(3):124–30.

81. Ministério da Saúde - Secretaria de Vigilância em Saúde - Departamento de DST, Aids e Hepatites Virais. Boletim Epidemiológico - Hepatites Virais. 2015 [Internet]. [cited 2016 Jan 26]. Available from: http://www.aids.gov.br/sites/default/files/anexos/publicacao/2015/58210/_p_boletim_hepatites_final_web_pdf_p__16377.pdf

82. Blanco Fernández MD, Torres C, Riviello-López G, Poma HR, Rajal VB, Nates S, et al. Analysis of the circulation of hepatitis A virus in Argentina since vaccine introduction. Clin Microbiol Infect. 2012 Dec;18(12):E548–51.

83. Pallen MJ. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. Parasitology. 2014 Dec;141(14):1856–62.

84. Armstrong SK. Bacterial Metabolism in the Host Environment: Pathogen Growth and Nutrient Assimilation in the Mammalian Upper Respiratory Tract. Microbiol Spectr. asm Pub2Web; 2015 Jun 25;3(3).

85. Centers for Disease Control and Prevention (CDC). Pneumococcal Disease - Streptococcus pneumoniae. 2013 [Internet]. [cited 2016 Feb 9]. Available from: http://www.cdc.gov/pneumococcal/clinicians/index.html

86. Centers for Disease Control and Prevention (CDC). Meningococcal Disease - Diagnosis and Treatment. 2015 [Internet]. [cited 2016 Feb 9]. Available from: http://www.cdc.gov/meningococcal/about/diagnosis-treatment.html

87. Sáfadi MAP, Carvalhanas TRMP, Paula de Lemos A, Gorla MCO, Salgado M, Fukasawa LO, et al. Carriage rate and effects of vaccination after outbreaks of serogroup C meningococcal disease, Brazil, 2010. Emerg Infect Dis. 2014 May;20(5):806–11.

88. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PLoS One. Public Library of Science; 2014 Jan 4;9(4):e94249.

89. Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA. Biogeography and

individuality shape function in the human skin metagenome. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 Oct 2;514(7520):59–64.

90. Mokili JL, Dutilh BE, Lim YW, Schneider BS, Taylor T, Haynes MR, et al. Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. PLoS One. 2013 Jan;8(3):e58404.

91. Head I, Saunders J, Pickup R. Microbial Evolution, Diversity, and Ecology: A Decade of Ribosomal RNA Analysis of Uncultivated Microorganisms. Microb Ecol. 1998 Jan;35(1):1–21.

92. Lipkin WI. Pathogen discovery. PLoS Pathog. 2008 Apr;4(4):e1000002.

93. Relman DA. The search for unrecognized pathogens. Science. 1999 May 21;284(5418):1308–10.

94. Lipkin WI. Microbe hunting. Microbiol Mol Biol Rev. 2010 Sep;74(3):363–77.

95. Daly GM, Leggett RM, Rowe W, Stubbs S, Wilkinson M, Ramirez-Gonzalez RH, et al. Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data. PLoS One. Public Library of Science; 2015 Jan;10(6):e0129059.

8 MATERIAL SUPLEMENTAR

Tabela S1: Quantidade de *reads* geradas e processadas de cada *pool*.

Pools	Geradas	Pré-processamento			Não Humanas	
		Pares	*Singletons*	Perdidas	Pares	*Singletons*
01	1.264.388	278.340	29.312	924.583	65.288	40.253
02	2.609.322	168.552	16.415	2.343.354	40.880	24.842
03	5.080.650	3.221.912	392.506	1.407.607	399.692	275.188
04	2.265.210	57.236	7.126	2.128.488	16.560	10.539
05	1.905.406	169.220	15.674	1.661.035	36.636	22.183
06	5.721.162	3.173.264	458.033	1.992.934	403.446	275.808
07	1.890.950	536.186	38.415	1.274.856	83.518	47.992
08	4.023.394	2.891.494	221.742	875.824	397.986	239.085
09	1.883.538	548.830	45.346	1.243.838	124.372	72.451
10	1.670.910	62.226	6.209	1.555.348	10.454	6.169
11	2.561.630	1.207.546	119.462	1.194.482	140.590	92.795
12	2.214.166	1.024.654	78.274	1.074.003	176.010	103.698
13	1.601.974	210.148	18.075	1.324.774	61.514	36.185
14	1.073.918	388.638	35.194	627.581	48.370	31.779
15	2.446.786	257.636	32.458	2.084.278	52.036	33.977
16	1.242.532	98.054	13.688	1.090.800	18.752	12.446
17	2.042.270	1.271.100	198.069	547.195	133.638	99.861
18	2.261.434	214.310	43.022	1.923.547	53.500	38.376
19	1.746.966	60.466	5.529	1.626.528	13.866	8.352
20	1.314.064	821.282	64.317	411.555	181.162	102.782
21	81.562.986	7.894.412	1.350.840	63.968.749	3.697.084	363.999
22	92.415.792	23.593.952	5.905.941	57.525.483	19.420.714	4.192.036
23	87.532.378	26.912.396	4.614.215	50.619.152	19.396.938	3.457.782
24	91.196.508	25.410.608	4.237.625	56.329.237	8.205.454	1.280.122
25	81.607.852	21.582.828	3.985.854	51.055.550	5.243.136	765.523

Tabela S2: Gêneros de bactérias identificados por cada programa em cada *pool* da primeira corrida. . (M): Metaphlan2; (G): Gottcha; (K): Kraken; (S): Surpi; (T): Taxoner.

Bactérias	**M**	**G**	**K**	**S**	**T**
Acholeplasma		16	16	16	16
Achromobacter			3,20	3,19,20	1,5,6,8,11,13
Acidaminococcus				20	
Acidiphilium			1-4,6-9,12,14,17,20		
Acidithiobacillus				3,11,20	
Acidovorax			2,3,5,8,11,19,20	2,3,5,7-9,11, 19,20	3,8,11,20
Acinetobacter		1,8,11, 12,15,20	1,5-8,11,12,15,20	1-13,15-20	1,5-8,11,12,15,20
Actinobacillus			2,3,12,20		
Actinoplanes				1-3	3
Actinosynnema				1	
Adlercreutzia			1		
Aeromonas			3,6,8,20	2,3,12,13,20	3,20
Aggregatibacter					6
Agrobacterium			3,11	2,3,11,20	
Alcaligenes				3	12
Alcanivorax				3	
Alicycliphilus		2	2,7,11,20	2,12,15	2,3,8,11,19,20
Aliivibrio			3,13,16,20	3,13,15-17	
Alistipes			9		
Alkalilimnicola			3	3	
Allochromatium				3	3
Alteromonas			1-20	6-8,11,13,16, 20	
Amycolatopsis			12	2,3,12	
Anabaena			1,3,6,8-12,17,18		
Anaerococcus			3		3
Anaeromyxobacter			1	3,6	
Anaplasma			1-6,8,10-18		
Arcanobacterium				12,17	12
Arcobacter			20		
Aromatoleum			3	3	
Arthrobacter			1,3	1-3,17	
Arthrospira			17		
Asticcacaulis			3	2,20	
Azoarcus			3	3,15,20	
Azobacteroides			1,3,6-8,11-18		
Azorhizobium				16	
Azospirillum			3	3	20
Azotobacter			3,2	3,2	
Bacillus		2,8,14	1-3,7,8,11,12,14-18, 20	1,2,4-9,11-15, 18-20	1-3,8,14,20

Bacteroides		9	1,3,5,6,8-13,16,17, 20	9	9
Beijerinckia			3,6-8,11,13,17,20		
Belliella			3,6,11,17		
Beutenbergia				3	3
Bifidobacterium		3,20	3,12,20	3,20	3
Blastococcus			8,17	3,15	
Blattabacterium			1	5	
Bordetella			3,8,20		3
Borrelia			1-4,6,8-17	15	6,9
Brachybacterium			2-4,6,8	3,8	2,6
Brachyspira			1,3,5,6,8,9,17		3
Brevibacillus			1-8,10-12,14-19	2-5,7,8,11-13, 15,19,20	3
Brevundimonas			3,20	3,20	3,7,20
Buchnera			1		
Caldanaerobacter			14,15		
Caldicellulosiruptor			1-16,18-20		
Calothrix			8,12		
Campylobacter			12,16	5,9	7,18
Candidatus_Accumulibacter				3	
Candidatus_Carsonella				3	
Candidatus_Methylomirabilis				3	
Candidatus_Phytoplasma				3,8,17	
Carboxydothermus			3		
Cardinium			3,8,11		
Carnobacterium				20	16
Castellaniella				3	12
Catenulispora				3	
Caulobacter			3,6,20	3,6,20	3,20
Cellulomonas			2,11	1	
Cellulophaga			1,17		
Cellvibrio				3	
Chamaesiphon			1,3,6-9,11-14,17,20		
Chlamydia			2		
Chlorobium			1-3,5-9,11-17,20		
Chloroherpeton			3,8		
Chromobacterium			3,8	3,15	
Chromohalobacter		3	3	3	3
Chryseobacterium				2,4,5	
Citrobacter			1,3,5,9,20	3,8,9,20	5,9,20
Clavibacter				1,3	
Clostridium			3,8,12,17,20	6,8	8,12,16
Collimonas			3	3	
Comamonas				3,7	7
Conexibacter				2,8,18	
Coraliomargarita			12		

Corallococcus				1,12	
Crinalium			20		
Cronobacter		8	8,11,20	8,9,20	8,20
Cupriavidus			1,3,8,20	2-4,6,7,9,10,12, 17,20	3,8
Cyclobacterium				20	
Cylindrospermum			3,6,8,14,17		
Cytophaga			3		
Dechloromonas				3,14	
Dehalobacter			1,3,6,8,11,12,17		
Deinococcus			3	3	
Delftia			3,8,18,20	3	2,3,18,20
Dermacoccus				1,3	
Desulfarculus			9,12		
Desulfatibacillum			1,2,5,7,11-13,15, 17,19		
Desulfitobacterium				20	
Desulfobacterium			3,11		
Desulfosporosinus			20	20	
Desulfotomaculum			20		20
Desulfovibrio			3	3	3
Desulfurobacterium			3,6,8,9,11-14,17	7	
Dickeya			3,5-8,11,12,16,17	3,20	6,20
Dietzia				1-3,11	1
Dokdonia			20		
Dyella				3,20	
Edwardsiella			3,20	3,9,15,20	3,9
Eggerthella			3		
Ehrlichia			1-3,9-15,18,19		
Ensifer				3	3
Enterobacter		8,9,20	3,7-9,16,20	1-3,8,9,12,19,20	1-5,8,9,14-16,20
Enterococcus		9,20	8,9,20	3,8,9,20	3,8,9,16,20
Erwinia			1,3,20	1-3,20	1,2
Escherichia			1-15,17-20	1-3,5,7-11,13, 15,18,20	2,3,5,8-10,12, 18,20
Faecalibacterium					3
Ferrimonas				3	
Flavobacterium			1-4,6-9,11,12,14,15, 17,18,20	2	
Frankia				1-3	3
Fusobacterium				20	
Gallibacterium			3,6		
Gallionella			3	3	3
Geobacillus			3,7,8,11	3,7,8	8
Geobacter				3	
Geodermatophilus			3	1,3,8	
Gloeobacter				3	
Gloeocapsa			3		

Gluconacetobacter				20	
Gluconobacter			20	20	
Gordonia			1-3	1-3,5	1,20
Haemophilus		6,8,17	1-18	6,8,17	6,8,17,20
Haererehalobacter				3	
Hahella			3,6		
Haliangium			1,3,8		11
Haloferax			12		
Halogeometricum			12		
Halomonas			3,11	3	3,11,15
Haloquadratum			1,3,8,9,14,15,19		
Halorhodospira			3	3	
Halothece			3,6,8,12,15,17		
Halothiobacillus			3,17	3	
Hamiltonella			3,6,8,9,11,12,15,17		
Helicobacter			1,3,6,8,9,12,17,18, 20	6	6
Herbaspirillum				3,8	20
Histophilus			3,9,12,17		
Hydrogenophilus				8,20	8
Hymenobacter				3	
Idiomarina				20	
Isoptericola			2		
Janthinobacterium				3	
Jonesia				4	
Kineococcus				3	
Kitasatospora				1-3	
Kocuria			5,13		13
Kosakonia				20	
Kribbella				3	
Kutzneria				1,1	
Kytococcus				1	
Lactobacillus	20	20	1,2,5,6,7,10-13,17-20	3,8,9,19,20	3,9,20
Lactococcus		9,2	9,2	9,17,19,20	9,20
Laribacter			10	3	18
Lawsonia			1-20		
Legionella			1,3,5-8,11,12,14-20		
Leifsonia			3,18		
Lelliottia				20	
Leptospira		7,10-13	1-3,5-7,9-16,18,19	12	7,10,12,13
Leptothrix			3	3	
Leuconostoc		20	9,20	9,20	9,20
Liberibacter			2,14,17,19,20		
Listeria			17,20		
Lysinibacillus				8,20	
Mageeibacillus			1,9		

Magnetospirillum				3	
Mannheimia			1-3,6-8,12,14,17		
Marichromatium				3	3
Marinobacter			3,20	3	3,15
Megamonas					6
Mesorhizobium			3	3,20	
Methanobrevibacter			3,8,9,11,12,17		
Methanocaldococcus			3,8,17		
Methanococcoides			3,5,7,8		
Methanococcus			1,3,6-9,11,12,14,17		
Methanomethylovorans			3,8,17		
Methanosalsum			1,6,8,20		
Methanosarcina			3,6-11,12,14,15,17,20		
Methylacidiphilum			11		
Methylibium			3,20		
Methylobacterium		5,9,13,20	2-9,11-13,17,18,20	2,3,5-9,12,13, 15,18-20	2-9,15,20
Methylocystis			3,6,8,9,12-14,17		
Methylophaga			3		
Methylotenera			20		
Micavibrio				3	
Microbacterium			13		13
Micrococcus		13-15	1,5-7,9,12-16,18	3,8,14,15,17,18	1,5-7,9,12-16,18,20
Microcyclus					9
Microcystis			3,9	9	
Microlunatus			6		
Micromonospora			4	2,4	4,6
Modestobacter			1,2	3	
Moraxella				3,9,17,18	8,20
Morganella			1,11	3,2	
Mycoplasma			1-20	6,16	6,12,17,18
Myxococcus					20
Nakamurella				3,12	
Natrialba			3,6-8,11-13,17		
Natronococcus			3,6-8,11,12,17,18		
Neisseria	9,16-20	1-20	1-20	1-20	1-20
Neorhizobium				3	
Nitrosomonas			20	8	
Nitrosospira				9	
Nocardia			2,3,15,18	1-3,11,18	18
Nocardioides				1,2	
Nocardiopsis			3	1,3,15	3
Oceanimonas			2,3	3,15	
Oceanithermus				2	
Ochrobactrum		2	2,3	2	2
Oleomonas				3	

Olsenella			20		
Ornithobacterium			3,6-9,12		
Oscillatoria			1,3,5,6,8,9,11,12, 17,20	17	
Paenibacillus		20	3,2	3,11,20	20
Pandoraea				3	
Pantoea			3,8	3,20	3,9,20
Parachlamydia			1,2,11,14,15		
Paracoccus			3	3	
Paraglaciecola			11,17		
Pararhodospirillum			4		
Pasteurella			4		
Pectobacterium			3,8,20	2,3,20	
Pediococcus		20	20	20	
Pelagibacterium				3	
Peptoclostridium				2,3,9,13,15-20	
Phenylobacterium				18	
Photobacterium		3	3,8,11,12	2	
Photorhabdus			3,9,11,12,17,20		
Phycisphaera			12		
Phytoplasma			3	3,8,17	
Pimelobacter				3	
Polaromonas			6,11	3,11	
Polymorphum			8		
Porphyromonas			6,17		
Portiera			3,8,9,12,17		
Prevotella			3,8,9,12,16,20		
Prochlorococcus			3		
Propionibacteriaceae	1				
Proteus		11,12	1-3,6-8,11,12,14	2,5,11,12,20	11
Providencia			20		
Pseudoalteromonas				6	
Pseudogulbenkiania				9	
Pseudomonas		3,6-9, 13-18,20	2-15,17-20	2-20	3--20
Pseudoxanthomonas			3,8	3,8	3
Psychrobacter		13	3,13	13,2	13
Psychroflexus			18		
Psychromonas			8,14,17,20	3	
Pusillimonas			12		
Pyrolobus			3-5,9,12-15,18,19		
Rahnella		9	9,20	2,9,20	20
Ralstonia		20	11,20	3,9,11,20	9,20
Ramlibacter				3	3
Raoultella		9	3,9,20	9,20	
Renibacterium			8		
Rhizobium			3,6,8,10,13,20	3	

Rhodanobacter			3	3	
Rhodobacter				3,8	
Rhodococcus			2,3,9,17	1-3,5,11,12	1,2
Rhodoferax				3	
Rhodopseudomonas			1,3,6-9,11,12,14-18	3,16,20	16
Rhodospirillum			3,6,20	2,3	
Rhodothermus			1,2,4-7,11-13,15	3,8	
Rickettsia			1		
Riemerella			1,3,4,6,8-14,16,17,20		7,12,18
Rivularia			20		
Roseobacter			2		
Rothia		20	20	9	9,20
Rubrivivax			3	3,15,20	
Rubrobacter				2,17	
Ruegeria			3	3	
Ruminococcus			20		
Runella			6,15		
Saccharomonospora			1	1,3	
Saccharopolyspora			2,3	3	
Salinarchaeum			8,17		
Salinicola					3
Salmonella		20	1,3,6,8,9,12,14,17,20	1,3-5,7-9,20	3,8,9,20
Sanguibacter				2,6,15	
Saprospira			1		
Segniliparus			2	1,17	
Selenomonas				20	8
Serpula					3
Shewanella			1,3,4,6,8,11,12, 15-17,20	4,5,16,20	3,4
Shigella			2,3,8,20	20	
Shimwellia			3,8	20	
Siccibacter			3,4,8	2,3,20	
Sideroxydans				3	
Simkania			9		
Sinorhizobium			3,8,12,15,17	3	
Sodalis			3,8	3,20	6
Solibacillus			6,8	20	
Sorangium			3		
Sphaerochaeta			3,8,9		3,8
Sphingobium			3,8	3,8,12,15,20	3,7
Sphingomonas			3,4,8	3,4,15,19,20	20
Sphingopyxis				20	
Spiribacter				3	
Spiroplasma			12,17		16
Staphylococcus		6-8,15, 18-20	3,6-8,15,18-20	2-9,12,13,15, 19,20	3,6,7,15,18-20
Starkeya			3,20	3	7

Stenotrophomonas		3,20	3-6,20	3,4,8,15,20	3,6,20
Streptococcus	3	3,4,6,8,9, 20	1,3,4,6-12,14,16-18, 20	1-11,13-16, 18-20	1,3,4,6-9,16,18,20
Streptomyces			2,3,6,8,9,11,17,20	1-3,12,18,20	1,3
Streptosporangium				3	
Sulfolobus			1,2,12-15		
Sulfuricella			12		
Sulfurihydrogenibium		14	14	14	
Sulfuritalea				3	
Symbiobacter			20	3	
Symbiobacterium				3	
Synechococcus			3	20	
Syntrophothermus			17		
Taylorella			12		
Teredinibacter			3,8	3	
Tetragenococcus				20	
Thalassolituus			3		
Thauera				3,20	
Thermaerobacter					3
Thermoanaerobacter			1-9,11-18	6	13,16,18
Thermoanaerobacterium				6	
Thermobifida			3		
Thermodesulfobacterium			1		
Thermomonospora				1	
Thermus		3	3,13	3,13	3,13
Thioalkalivibrio			3,15	2,3,15,20	3
Thiobacillus				3	
Thiocystis			3,6		
Thioflavicoccus			8,14		
Thiomonas				3	
Tistrella			3	3	
Tolumonas				3	
Treponema			3		
Trichodesmium			1,3,6,8,11,12,14,17		
Trueperella				2	
Tsukamurella				2,3,12	
Variovorax			3,15	3,15	3
Veillonella		18	18	18	18
Verminephrobacter			1,3,5-9,11,12,14, 17,18,20		
Waddlia					11
Weeksella					20
Weissella				20	
Wolbachia					16
Xanthobacter			3,6		
Xanthomonas			3,6,8	3,15	
Xenorhabdus			1,3,6,8,11,15	11,20	

Xylanimonas					1,2
Xylella			3,6,7,10,20	17	7
Yersinia		1-3,6,10-13	1-3,5,6,8,10-15,17	1-20	1-3,5,6, 10-15,20
Zymobacter					20
Zymomonas			20		

Tabela S3: Vírus identificados por cada programa em cada *pool* da primeira corrida. . (M): Metaphlan2; (G): Gottcha; (K): Kraken; (S): Surpi; (T): Taxoner; (B): Blastn.

Virus	**M**	**G**	**K**	**S**	**T**	**B**
Abalone herpesvirus			3,6,8,11-13,17,20			
Acanthamoeba polyphaga mimivirus					1-4,7,8,12, 15,16,18	
Acidithiobacillus phage					3,11,20	
Acinetobacter phage		8				8
Agrotis segetum nucleopolyhedrovirus			3,9,12,13			
Alcelaphine herpesvirus 1			3,8			
Alcelaphine herpesvirus 2			1,6,8			
Alphamesonivirus		16	16	16,17	16	
Anguillid herpesvirus 1			3,6,8,11,14,17,19		8	
Aotine herpesvirus 1					8	
Aspergillus foetidus dsRNA mycovirus			12			
Bacillus phage			1,5,7,8,15	8		8
Ball python nidovirus			8			
Bean necrotic mosaic virus			1,3,12			
Bovine adenovirus 2					3	
Bovine herpesvirus 6			3			
Broad bean mottle virus			1,2,5,7		14	
Brochothrix phage			8,14			
Cachoeira Porteira virus					12	
Campoletis sonorensis ichnovirus			4,15			
Canarypox virus			7,8,14			
Casuarina virus			16	16		16
Cavally virus					16	16
Caviid herpesvirus 2			1-19			
Cercopithecine herpesvirus 2			6,8,11,12,17,18		12	
Cercopithecine herpesvirus 5			1,3,8,11,14			
Choristoneura occidentalis granulovirus			8,9,11	3		8,9,11,20
Coccolithovirus						
Cotesia congregata bracovirus			1-12,14,15,17,18		3,8,12,14	
Cotesia sesamiae Kitale bracovirus					1,3,8	

Cotesia sesamiae Mombasa bracovirus					8	
Cowpox virus					1,6,11	
Cronobacter phage						2
Cyanobacteria phage					4	
Cyanophage				8		
Cynomolgus Epstein-Barr Virus					11	
Cyprinid herpesvirus 1			1-6,8,12,14,17,20		8	3,8,12
Cyprinid herpesvirus 2			1			3,8
Cyprinid herpesvirus 3			1-3,5,6,8,9,11, 12,14,17,18,20		7,17	20
Dak Nong virus				16	16	
Deerpox virus W-1170-84			3,9			
Dill cryptic virus 1			3,5,8,12,14			
Dill cryptic virus 2			11,12			
Dinoroseobacter phage						7
Dulcamara mottle virus			1-20			
Ectocarpus siliculosus virus			3,8,11,17,20			
Ectromelia virus			8			
Elephant endotheliotropic herpesvirus					6	
Elephantid herpesvirus			1,3,6-8,11,12, 15,17		3,6	
Emiliania huxleyi virus			2,3,6-9,11-14,17		1,3,6,12	1
Enterobacteria phage		2,8,15	2,8,15	1-18	2,8,15	2,8,9,15,20
Epinotia aporema granulovirus			3,8			
Equid herpesvirus						3,8
Falconid herpesvirus						3,8
Fowl aviadenovirus E			11,18			
Frog virus 3			8			
Gallid herpesvirus 2			3			3,8
Gallid herpesvirus 3						3,8
Gentian ovary ring-spot virus			1-19			
Glypta fumiferanae ichnovirus			1-4,6-9,11-14,17, 18,20			
Hana virus			16	16	16	16
HCBI8.215 virus						3,8
Hemileuca sp nucleopolyhedrovirus			17			
Hepatitis A virus	20		20		20	20
Hepatitis C virus			1-20		2,3,5,6,8,10, 11,13,15-18	
Hibiscus latent Fort Pierce virus			1,2,4,5,7,9,10, 12-16,18			
Hibiscus latent Singapore virus			1-4,6,8,9,12,15			
Hop trefoil cryptic virus 2			2			
Human endogenous retrovirus			1-3,6,8,9,11-13, 15-17,20		1,3,6-9,12, 13,17	1-3,5-9, 11-13, 15-17,20
Human herpesvirus 1					3,6-9,11,14, 17	

Human herpesvirus 4	17	6,14,17	6,14,17	6,14,17	3,6-8,12,14, 17	6,14,17
Human herpesvirus 5					3	
Human herpesvirus 6						3,8
Human herpesvirus 7						3,8
Human herpesvirus 8			6,12			
Human immunodeficiency virus 1				8		
Human papillomavirus					1,3,6,8,12, 17,18	
Human parvovirus B19	4	4	4	4	4	4
Human T-lymphotropic virus 1					6,8,15	
Hyposoter fugitivus ichnovirus			1-3,6-9,11-15, 17,20		1,7,8,12	
Ictalurid herpesvirus 1			1-4,6- 9,11,12, 14,15,17,18,20		6-8,12	6
Impatiens necrotic spot virus			3,6,8			
Lactobacillus phage		3,2	3	3		3,2
Lactococcus phage		20	20			20
Leuconostoc phage						20
Macacine herpesvirus 1			1-5,7-10,11,12, 14,15,18,19			17
Magnaporthe oryzae chrysovirus 1			3,14			
Mamestra configurata nucleopolyhedrovirus A			17			
Maruca vitrata nucleopolyhedrovirus			3,8,11			
Megavirus chiliensis			8			
Meleagrid herpesvirus 1						3,8
Meno virus						16
Molluscipoxvirus			3,6,8			
Molluscum contagiosum virus			3,6,8			
Mongoose feces-associated gemycircularvirus						3,8
Monkeypox virus			12,20			
Moumouvirus			3,8			
Murid herpesvirus 2			3			
Murid herpesvirus 8			1-3,6,7,9,11,12, 14,15			
Musca hytrovirus			6,12			
Nam Dinh virus					16	16
Nse virus strain						16
Orgyia pseudotsugata multiple nucleopolyhedrovirus			3,6,8,12,17			
Oryctes rhinoceros nudivirus			3,6			
Oryctes rhinoceros virus					3	14
Ovine adenovirus A						3
Ovine herpesvirus 2			11		8	3,8
Ovine mastadenovirus A		3	3	3		
Pandoravirus dulcis			1-9,11-15,17-20		3,6,9,12,18	2,4
Pandoravirus salinus			1-9,11-15,17-20		1,3,6,8,13, 17	

Penaeus monodon nudivirus			3,6,8,17			
Penicillium chrysogenum virus			8			
Phaeocystis globosa virus			3,6,8,11,12,15			
Pigeon aviadenovirus A			1-3,6-8,11,12,14, 17,18			
Pleurotus ostreatus virus 1			5,7,14-16,18			
Plutella xylostella granulovirus			8			
Propionibacterium phage			6	6		3,5,6
Pseudaletia unipuncta granulovirus			3			
Pseudomonas phage						3,8
Rat cytomegalovirus Maastricht					20	
Roseolovirus					11	
Saccharomyces cerevisiae killer virus			1-18,20			
Saimiriine herpesvirus 1			1,3,6,11			
Saimiriine herpesvirus 4			8			
Salmonella phage			20		20	20
Sclerotinia sclerotiorum partitivirus			2			
Shamonda virus			11			3,11,20
Shigella phage						8,9,20
Simbu virus			11			3,11,20
Simian immunodeficiency virus					6	
Simian virus					12	
Sorghum chlorotic spot virus			1,15,17,18			
Spodoptera frugiperda ascovirus 1a			9,17			
Spodoptera frugiperda multiple nucleopolyhedrovirus			20			
Spodoptera litura nucleopolyhedrovirus			3,17			
Staphylococcus phage						8,2
Stealth virus 1					3,8,9,11,12, 14,17	
Streptococcus phage		9		9		3,9
Synechococcus phage						9
Tick-borne encephalitis virus			1-19		18	
Tomato aspermy virus			3,8			
Tomato spotted wilt virus			6			
Trichoplusia ni ascovirus 2c			8,17,20			
Tupaiid herpesvirus 1			8,13			
Tursiops truncatus papillomavirus 1					3	
unclassified Coccolithovirus					3,6-8,12	
Upsilonpapillomavirus 1			8			
Vaccinia virus					8	
Vibrio phage			2,4,9,15-18			8
Vicia cryptic virus			1-20			
White clover cryptic virus 2			1,3,7,9,11,12,19			
White spot syndrome virus			8			
Yaba monkey tumor virus			20			
Yersinia phage			6,8,11,14,15,17			

Tabela S4: Gêneros de bactérias identificados por cada programa em cada *pool* da segunda corrida. Os *pools* 21 ao 25 estão apresentados como *pools 1* ao 5. . (M): Metaphlan2; (G): Gottcha; (K): Kraken; (S): Surpi; (T): Taxoner.

Bactérias	**M**	**G**	**K**	**S**	**T**
Acetobacter			3,4		
Acetobacterium				2-5	
Acholeplasma				1,3,5	
Achromobacter		3	1-5	1-5	1-5
Acidimicrobium			1	2	
Acidiphilium		3	1-5	1-5	1-5
Acidithiobacillus			2	2,3	
Acidobacterium			1-4	1-5	
Acidothermus			4	1	
Acidovorax		4	1-5	1-5	1-5
Acinetobacter		1,3-5	1-5	1-5	1-5
Actinobacillus			4	1,4,5	
Actinomadura				1	
Actinomyces					2
Actinoplanes		3	1-5	1-5	2-4
Actinosynnema			1,2,4,5	1,3-5	2,4
Adlercreutzia			2	1,5	
Advenella			1,2	1,2	
Aequorivita				3	
Aeromonas		2,3	1-5	1-5	1-3
Afifella					3
Afipia				1-5	1-4
Aggregatibacter			1,4,5	1,4,5	4,5
Agrobacterium		1-5	1-5	1-5	
Akkermansia				3	
Alcaligenes					2
Alcanivorax			1,2,4,5	1-5	
Alicycliphilus		1,3	1-5	1-5	1-4
Alicyclobacillus			4	3	
Aliivibrio			4		
Alistipes		1	1	1,2	1
Alkalilimnicola			2-5	1,2,5	
Alkaliphilus				1	
Allochromatium			1-4	1,2,5	1
Alteromonas			1-5	1-5	

Aminobacter				2,5	
Ammonifex			1		
Amphibacillus				3	
Amycolatopsis			1-5	1-5	2-4
Amycolicicoccus			2	1	
Anaerococcus		3	1,3	1,3	1,3
Anaerolinea				1	
Anaeromyxobacter			1-5	1-5	
Anaerostipes				1,2,5	
Anaplasma			2,4		
Anoxybacillus		5	4,5	5	
Aquincola				1,2,5	
Arcanobacterium				2	
Arcobacter			4		4
Arsenophonus				1	1
Arthrobacter			1-5	1-5	1-3,5
Asticcacaulis		4,5	3-5	1-5	2,5
Atopobium				1,5	
Aurantimonas				2,5	2,5
Avibacterium				1	
Azoarcus		2	1-5	1-5	
Azorhizobium		1,2	1-5	1-5	
Azospira			1,2,5	1-5	
Azospirillum		3-5	1-5	1-5	1-5
Azotobacter			1-5	1-5	
Bacillus			1-5	1-5	2,4,5
Bacteriovorax					4
Barnesiella				2	
Bartonella			2	2,5	
Basfia				1	
Bdellovibrio			3,5		
Beijerinckia		2,4,5	1-5	1-5	3,5
Belliella			2,3	1,4	
Beutenbergia			1,3,4	1-5	1,3,4
Bibersteinia		5		5	5
Bifidobacterium		1-5	1-5	1-5	1-5
Blastochloris				1,2,4	
Blastococcus		1	1-5	1-5	
Blattabacterium		5	5	2,3,5	5
Blautia				1,2,5	

Bordetella			1-5	1-5	1-5
Borrelia			2-5	1,3,5	2-5
Bosea				1,3-5	1,3,4
Brachybacterium		2	1-5	1-5	1-4
Brachymonas				1,2	
Brachyspira			2-5	2,3,5	
Bradyrhizobium	1-5	1-5	1-5	1-5	
Brevibacillus			2,4		
Brevundimonas		2	1-5	1-5	1-4
Brucella			1-5	1-5	1-5
Buchnera				2,3,5	
Burkholderia		1,3	1-5	1-5	1-5
Butyrivibrio				1,2,5	
Caldanaerobacter			2,4		
Caldicellulosiruptor			2,4		
Caldilinea			2	1,3,5	
Caldisericum			3		
Campylobacter			3-5	1-5	1-5
Capnocytophaga		1,3	1,3	1,3,5	1,3
Carboxydothermus			5	1,5	
Carnobacterium			4	1,3-5	4
Castellaniella				1-5	2-4
Catenulispora			1-4	1-5	3-5
Caulobacter		2,3	1-5	1-5	1-5
Cellulomonas		4	1-5	1-5	1-4
Cellulophaga			2,4	1,3	
Cellulosilyticum				3	
Chelativorans		3-5	1-5	1-5	1,3
Chitinophaga			1	1,2,5	
Chlamydia				1	
Chlorobaculum			1,5	1-3	
Chlorobium			2-5	1,5	
Chloroflexus				1,3	
Chloroherpeton			2,3,5	1,5	
Chromobacterium			1-5	1-5	
Chromohalobacter			1-5	1-5	
Chryseobacterium				1,2,5	
Citrobacter		4	2-4	1-5	1-4
Clavibacter			1,2,4,5	1-5	
Clostridium		1,2	1,2,4	1-5	1-4

Collimonas		1,3	1-5	1-5	
Comamonas			1-5	1-3,5	1,2,5
Conexibacter		1,2	1-5	1-5	1-4
Coprococcus				1,2,5	1,5
Coraliomargarita			2,3,5		
Corallococcus		5	1,4,5	1-5	
Coriobacterium			2		
Corynebacterium		1-5	1-5	1-5	1-5
Crinalium				3,5	
Croceibacter			4		
Cronobacter			2-4	1-5	3,4
Cupriavidus		1,5	1-5	1-5	1-5
Curvibacter				1,2,4	3
Cyanobacterium			2	1,2	
Cytophaga				1,3	
Dactylosporangium				3	
Dechloromonas			1,2,5	1-3,5	
Deferribacter				1,3	
Dehalobacter			2-5		
Deinococcus			1-5	1-5	1,3,4
Delftia			1-5	1-5	1-5
Dermacoccus				1-5	
Desulfarculus			1-3,5	1	
Desulfatibacillum			2,5		
Desulfobacterium			2	5	5
Desulfobulbus				4,5	
Desulfococcus			3	2,3,5	
Desulfomicrobium			3	1	
Desulfomonile			2		
Desulfosporosinus				3	
Desulfotalea			2		
Desulfotomaculum				1	2
Desulfovibrio			1-5	1-3,5	
Desulfurivibrio				2	
Desulfurobacterium			2,3		
Dichelobacter				5	
Dickeya			1-5	1,2,4,5	3,4
Dietzia				1,3,5	
Dinoroseobacter			1-5	1-5	
Dokdonia			2-5	1	

Dyadobacter			1,4	1	1
Dyella				1-5	
Echinicola				1	
Edwardsiella			2,4,5	1,3-5	4
Eggerthella			2	1,2	
Ehrlichia			2,5	3	
Elizabethkingia				1,5	
Emticicia			2	5	
Ensifer				1-5	1-5
Enterobacter		2-5	1-5	1-5	1-5
Enterococcus		4	1,2,4,5	1-5	4
Erwinia		4	4	1,2,4,5	3,4
Erysipelothrix				2	
Erythrobacter		4	1-5	1-5	
Escherichia		3-5	1-5	1-5	1-5
Ethanoligenens				1	1,3,4
Eubacterium		1,5	1,2,5	1-3,5	1,3,5
Exiguobacterium		1	1	1,3	1
Faecalibacterium				1,2,5	1,5
Ferrimonas			1,2,4	1	2
Fervidobacterium			2		
Fibrella			2	1,3,5	
Filifactor		3	3	3,5	3
Fimbriimonas				1-5	
Finegoldia		1,4	1,3,4	1,3-5	
Flavobacterium		4	2-5	1-5	
Flexibacter				3,4	
Fluviicola				2	
Francisella				5	
Frankia			1-5	1-5	1-3
Frateuria			1,2	1-3,5	
Fulvimarina				5	
Fusobacterium		1,3,5	1-5	1-5	1-3,5
Gallibacterium				5	
Gallionella				1,2,5	
Gardnerella			2	1,2,4	
Gemmatimonas		2	2-5	1-5	
Geobacillus			2,3,5	2,3,5	2,5
Geobacter			1-4	1-3,5	
Geodermatophilus		1	1-4	1-5	1-5

Glaciecola				1,5	4
Gluconacetobacter		1	1-5	1-5	3
Gluconobacter			1,3-5	1-3,5	
Gordonia		2,3	1-5	1-5	1-4
Gordonibacter			2	1,3,5	
Gramella				1	
Granulibacter			1-3	1-5	2,3,5
Granulicella			1-4	1-3,5	5
Haemophilus		1-5	1-5	1-5	1-5
Hahella			2,3,5	3,5	
Halanaerobium			3	3,5	
Haliangium			2-5	1-5	1,3
Haliscomenobacter			2,4		
Haloarcula			2-5		
Halobacterium			1		
Halobacteroides				5	
Haloferax			3,5		
Halogeometricum			2,3		
Halomonas		1,2	1,2,4,5	1-3,5	3,5
Haloquadratum			2-5		
Halorhodospira			2-4	1-3,5	5
Halothermothrix				1	
Halothiobacillus			1-5	1,5	
Helicobacter			2-5	2,3,5	2,3,5
Heliobacterium				5	
Herbaspirillum	1	1	1-4	1-5	1-5
Herminiimonas			1,4	1-5	3
Hippea				1,5	
Hirschia			3	3	
Histophilus			2-5	1,5	
Hydrogenophaga				1	
Hymenobacter				1-3,5	
Hyphomicrobium		1-3	1-5	1-5	1-5
Hyphomonas			2-5	1-5	
Idiomarina				1	
Ignavibacterium				1,3	
Ignicoccus			2,3,5		
Ilumatobacter			2,4	1,2,4,5	
Ilyobacter				2,5	
Inquilinus				2	

Intrasporangium			1,2,4,5	1-5	1,4
Isoptericola			1,2,4	1-5	1,4
Isosphaera			1-3	1,2	
Jahnella				2	
Janibacter				1	
Jannaschia				1-5	
Janthinobacterium			1-3	1-5	1-5
Jonesia				1	
Kangiella			4	3,5	
Ketogulonicigenium			1-5	1-5	3-5
Kineococcus			1-4	1-5	
Kitasatospora			4	1-5	2-4
Klebsiella		1,3-5	1-5	1-5	1-5
Kocuria		2	1,2,4,5	1,3-5	2
Komagataeibacter		3	2	1-5	
Kosmotoga			5		
Kribbella			1-5	1-3,5	1,4
Kutzneria				1-5	1,3
Kytococcus		1	1,3	1,3-5	1
Lacinutrix				5	
Lactobacillus		1-5	1-5	1-5	1-5
Lactococcus		1-4	1-4	1-5	1-5
Laribacter			1-5	1-5	
Lawsonia			1-5		
Leadbetterella				1	
Legionella			2-5	2,5	3
Leifsonia			1-4	1,2,4,5	1,3
Leisingera			1-5	1-5	2-5
Leptospira		5	2,5	5	5
Leptospirillum				2	
Leptothrix		2	1-5	1-5	
Leptotrichia		1	1	1,4,5	1
Leuconostoc		1,4	1,4	1,3,4	1,4
Listeria			2-5	1,4,5	4
Lysobacter				1-3,5	
Macrococcus		2	1-3	2,3,5	
Magnetococcus				2,5	
Magnetospira				1,5	
Magnetospirillum		1,2	1-5	1-5	1-5
Mannheimia			1-5	1,3-5	

Maribacter				1	
Maricaulis			1-4	1-5	
Marichromatium				1-5	1-4
Marinithermus				1	
Marinitoga				3,5	
Marinobacter			2-5	1-5	3
Marinomonas			3	1,3,5	
Marivirga				1,2	
Marmoricola					2
Massilia				1,5	
Meiothermus		1	1-3	1-3,5	3
Melissococcus				1	
Mesorhizobium		1,3	1-5	1-5	1-5
Mesotoga				2	
Methanobrevibacter			5		
Methanocaldococcus			2-5		
Methanococcoides			4		
Methanococcus			2-4		
Methanocorpusculum			2,3		
Methanohalobium			4		
Methanomethylovorans			2-5		
Methanosaeta			2,3		
Methanosalsum			2-4		
Methanosarcina			2-5		
Methanosphaerula			3		
Methanothermococcus			4		
Methanotorris			2,5		
Methylacidiphilum			1		
Methylibium			1-5	1-5	3,4
Methylobacillus			2	1,3-5	
Methylobacterium		1-5	1-5	1-5	1-5
Methylocella		1,2,4	1-5	1-5	2,3,5
Methylococcus			2,3	1-3,5	
Methylocystis		1-5	1-5	1-5	
Methylomicrobium			2-4	2,5	
Methylomonas				2	4
Methylophaga			3	5	
Methylovorus				1-4	2
Micavibrio			3,4	1-4	
Microbacterium		1	1-4	1-5	1-4

Micrococcus		1-5	1-5	1-5	1-5
Microcystis		5	1-3,5	1-3,5	3,5
Microlunatus			1-4	1-3,5	1,3-5
Micromonospora			1-5	1-5	
Mobiluncus			2	1,5	
Modestobacter		2	1-4	1-5	
Moraxella			2,5	1,5	3,5
Morganella			2,4		
Muricauda				1	
Mycobacterium			1-5	1-5	1-4
Mycoplasma			2-5	1,3-5	2-5
Myxococcus		5	1-5	1-5	2,3
Nakamurella			1-4	1-5	5
Natrialba			2-5		
Natronococcus			2-5		
Nautilia				3	
Neisseria	2	1-5	1-5	1-5	1-5
Niabella					2
Niastella		1	1,2	1,3,5	
Nitratifractor				2	4
Nitrobacter		1-5	1-5	1-5	2-5
Nitrosococcus			2-4	5	
Nitrosomonas			3	1,5	
Nitrosospira			3	1,2	
Nitrospira				1,3-5	
Nocardia			1-5	1-5	1,3,4
Nocardioides		1,3	1-5	1-5	
Nocardiopsis			1-4	1-5	4,5
Nonlabens			1	1	
Nonomuraea				1,3-5	
Novosphingobium		2	1-5	1-5	2-5
Oceanimonas			3-5	1	
Oceanithermus				1,2,4	1,3
Oceanobacillus				3	
Ochrobactrum			1-5	1-5	5
Octadecabacter			1,3-5	2,3,5	
Odoribacter				3,5	
Oligotropha		1-5	1-5	1-5	1-5
Olsenella				2,4	
Opitutus			1-5	1-5	

Ornithobacterium			1-5	1	
Oscillibacter			2,3,5		
Owenweeksia				2,5	
Paenibacillus			1-3,5	1-5	2
Pandoraea			2-5	1-5	2-4
Pantoea			1-4	1-5	3,4
Parabacteroides				1,5	
Paracoccus		3,4	1-5	1-5	1,2
Parvibaculum		1-4	1-5	1-5	5
Parvularcula			1-3	1-5	2,4
Pasteurella			3,4	3	3
Paucimonas				1,5	
Pectobacterium			2,4	1-5	3
Pediococcus				1,2,5	
Pedobacter				1,2	2
Pelagibacterium			1-4	1-5	
Pelobacter			1-3,5	1,3	
Pelodictyon				1-3,5	
Pelomonas				1,3	
Peptoniphilus				2,5	
Phaeobacter			1-5	1-5	
Phenylobacterium		1,2	1-5	1-5	2
Photobacterium			2	2,5	
Photorhabdus			2,3		
Phycisphaera			2-4	2,3	
Pimelobacter				1	
Pirellula			2	1,5	
Planctomyces					2
Polaribacter			5		
Polaromonas		2-5	1-5	1-5	
Porphyromonas		1	1,2	1,5	
Prevotella		1-5	1-5	1-5	2-5
Prochlorococcus		1	1,4,5	1,5	
Propionibacterium	1,3-5	1-5	1-5	1-5	1-5
Prosthecochloris			2-4	1	
Proteus		2,5	2-5	2-5	2,5
Providencia			5	1,4,5	
Pseudoalteromonas				2,5	
Pseudogulbenkiania			1-5	1,3-5	
Pseudomonas		1-5	1-5	1-5	1-5

Pseudonocardia			1-5	1-5	3,5
Pseudovibrio				1,2	
Pseudoxanthomonas			1-5	1-5	1-5
Psychrobacter		1	1,3-5	1-5	
Psychroflexus			2	2	
Psychromonas			1-3		
Pusillimonas				1,5	3
Pyrolobus			2,3,5		
Rahnella		4	4,5	1-5	
Ralstonia		1-4	1-5	1-5	1-5
Ramlibacter		2	1-5	1-5	1-5
Raoultella		1,4	1,3,4	1,3-5	4
Renibacterium				1-3	
Rhizobium		2,3	1-5	1-5	
Rhodanobacter			2-5	1-5	
Rhodobacter		5	1-5	1-5	1-5
Rhodoblastus				2	
Rhodococcus		3,4	1-5	1-5	1-5
Rhodoferax			1-4	1-5	
Rhodomicrobium			1-5	1-5	2-4
Rhodopseudomonas		1-5	1-5	1-5	1-5
Rhodospirillum			1-5	1-5	1-5
Rhodothermus			1,2,4,5	1-3,5	
Rhodovulum				3	
Rickettsia		5	1,5	1-3,5	5
Riemerella			2-5	1	1-5
Roseburia		4,5	2-5	1,2,5	5
Roseibacterium				1-3,5	1-5
Roseiflexus			2-5	1-3	
Roseobacter		4	1,2,4,5	1-5	2
Rothia		1-5	1-5	1-5	1-5
Rubrivivax			1-5	1-5	
Rubrobacter		1-5	1-5	1-5	
Ruegeria			1-5	1-5	1,3,4
Ruminococcus				1-3,5	1,3,5
Runella			2,3		
Saccharomonospora			1,2,4	1,5	
Saccharopolyspora			1,2,4,5	1-5	
Saccharothrix			1-4	1-3,5	
Salinibacter			2,4	1	

Salinispora			1-5	1-3,5	3
Salmonella		1,4	1-5	1-5	1,3,4
Sandarakinorhabdus				1	
Sanguibacter			1,2,4	1,2,4,5	4
Saprospira			2	1	
Sebaldella				5	
Segniliparus			3	1,3,5	
Selenomonas			1,2,5	1,3	1,2
Serratia		4,5	1-5	1-5	1,3-5
Shewanella		1	1-5	1-3,5	
Shigella			2-5	3-5	
Shimwellia			4	1,5	
Simiduia				4,5	
Singulisphaera				1-3,5	
Sinorhizobium		2,3,5	1-5	1-5	1-5
Slackia			1		
Snodgrassella				1	
Sodalis				1-3,5	
Solitalea				1	
Sorangium		1,4	1-5	1-5	
Sphaerobacter		4	1,4	1-3,5	4
Sphaerochaeta			2-5	1,5	2,3,5
Sphingobacterium		4	4	1,3,4	4
Sphingobium		3-5	1-5	1-5	1-5
Sphingomonas		1,2,4,5	1-5	1-5	1-5
Sphingopyxis		2	1-5	1-5	2
Spirochaeta			1,2	1,5	3
Spiroplasma				3	
Spirosoma		2	2	2,4	2
Stackebrandtia			1-4	1,2,5	2,4
Staphylococcus		1-5	1-5	1-5	1-5
Stappia				5	4
Starkeya		1-4	1-5	1-5	1-5
Stenotrophomonas		1-5	1-5	1-5	1,3-5
Stigmatella			2,3,5	1-5	2-4
Streptobacillus				2,5	
Streptococcus	3	1-5	1-5	1-5	1-5
Streptomyces			1-5	1-5	
Streptosporangium			1-4	1-5	2-4
Subtercola				1	

Sulfuricella			1,3,4	4,5	
Sulfuricurvum			2	2,3	
Sulfurihydrogenibium				3	
Sulfurimonas				3	
Sulfuritalea				1-5	2-5
Sulfurospirillum				2	
Sulfurovum				5	
Symbiobacterium			3,4	1,5	
Syntrophobacter			1-3		
Syntrophus			4	1	
Tannerella				3	
Teredinibacter			2-5		
Terrabacter				1	
Terriglobus				1,2,4,5	
Tessaracoccus					2
Tetragenococcus			1,4	1,3	
Tetrasphaera				4,5	
Thauera			1-5	1-5	1-5
Thermaerobacter			2,4,5	1,3-5	
Thermincola			2-5		
Thermoanaerobacter				1,5	4
Thermoanaerobacterium			1		
Thermobacillus			3	1,4	
Thermobifida			2,5	1-3	
Thermobispora			1,4,5	1-5	1-5
Thermocrinis				2	
Thermodesulfatator				3	
Thermodesulfobium				2,5	
Thermomicrobium				1,4	
Thermomonospora		1	1-5	1,3-5	1-3,5
Thermotoga				2	
Thermus		2	2,4,5	1,2,5	2,5
Thioalkalivibrio			1-5	1-5	1,3-5
Thiobacillus			1-5	1-5	
Thiocapsa				1,3-5	
Thiocystis			1-5	1-5	
Thioflavicoccus			2,4,5	1,3,5	
Thiomonas			1-3	1-5	1,3,5
Thioploca				3	
Tistrella		3,4	1-5	1-5	

Tolumonas				2	
Tomitella					2
Treponema		1	1,3-5	1-3,5	
Truepera			3-5	1,4	
Trueperella				1	2
Tsukamurella			2-4	1-3,5	1,3,4
Ureaplasma				1,3	
Variovorax		3,4	1-5	1-5	1-5
Veillonella		1-5	1-5	1-5	1-5
Verminephrobacter			1-5	1-5	
Verrucosispora			1,4	1,2,5	
Vibrio			2-4	1-3,5	2,3,5
Virgibacillus				3	
Waddlia					2
Weeksella		1,4	1,2,4	1,4	
Wigglesworthia				3	
Wolbachia			1	1,3	
Xanthobacter		1-5	1-5	1-5	2-4
Xanthomonas		1,3,4	1-5	1-5	1-5
Xenorhabdus			3,5	1,4,5	
Xylanimonas			1-4	1,2,4,5	1,2,4
Xylella			2-5	3,5	
Yersinia			2-5	1-5	2-4
Zobellia			2	3	
Zymomonas			1-3	1-5	2-4

Tabela S5:Vírus identificados por cada programa em cada *pool* da segunda corrida. Os *pools* 21 ao 25 estão representados como *pools 1* ao 5. . (M): Metaphlan2; (G): Gottcha; (K): Kraken; (S): Surpi; (T): Taxoner; (B): Blastn.

Vírus	M	G	K	S	T	B
Abalone herpesvirus			2-5		2,3,5	
Acanthamoeba polyphaga mimivirus					4	
Acanthocystis turfacea Chlorella						2
Acidithiobacillus phage					2	
Adoxophyes honmai entomopoxvirus				5		
Aeromonas phage						3
Agrotis ipsilon multiple nucleopolyhedrovirus				3		
Agrotis segetum nucleopolyhedrovirus			2-5			
Alcelaphine herpesvirus 1			2,4			
Alcelaphine herpesvirus 2			2-5			
Alphapapillomavirus				2,3		
Andrias davidianus ranavirus					1-5	
Anguillid herpesvirus 1			2-4		3	
Antheraea pernyi nucleopolyhedrovirus			2			
Aotine herpesvirus 1					2-5	
Armillaria luteovirens					5	
Autonomous rat parvovirus				2-5		
Baboon endogenous virus						2
Bacillus phage				3		
Bebaru virus						1
Beet soil-borne mosaic virus				5		
Betaherpesvirinae					2,4	
Betapapillomavirus 2			4			
Bracovirus					2,5	
Bufavirus		2	2	2		2
Cafeteria roenbergensis virus			3	3		
Campoletis sonorensis ichnovirus			2,3	2		
Canarypox virus			2			
Carnation etched ring virus				1		
Caulobacter phage						1-4
Caviid herpesvirus 2			2-5			3
Cercopithecine herpesvirus 2			2-5		2,3,5	
Cercopithecine herpesvirus 5			1-5	1-3,5	2-4	
Chelonid herpesvirus 5					3,4	
Chikungunya virus	1	1	1	1,3,5	1	1

Chinese giant salamander iridovirus					1-5	
Choristoneura occidentalis granulovirus			2,3			2,3,5
Circoviridae 2 LDMD-2013						5
Clostridium phage		2	2	2	2	2
Coccolithovirus						
Colobus guereza papillomavirus type 1					5	
Colwellia phage 9A			4			
Common midwife toad ranavirus					2-5	
Cotesia congregata bracovirus			1-5		1-5	
Cotesia plutellae polydnavirus				5		
Cotesia sesamiae bracovirus				3	3	
Cotesia sesamiae Kitale bracovirus					1-5	
Cotesia sesamiae Mombasa bracovirus					1-5	
Cotesia vestalis bracovirus				5		
Cowpox virus			2,3		1-5	
Cucumber mosaic virus					2	
Cyanophage				1		1
Cyanophage P-RSM6				1		
Cynomolgus macaque cytomegalovirus					5	
Cyprinid herpesvirus 1			1-5		1,2,4,5	2,3
Cyprinid herpesvirus 2			2,3		3	2,3
Cyprinid herpesvirus 3			1-5		1-5	4
Cyprinivirus					2	
Deerpox virus			2			
Dengue virus		2	2	2	2	2
Dill cryptic virus 1			3			
Dulcamara mottle virus			2-5			
Bromoviridae			2-5			
Ectocarpus siliculosus virus 1			2-5		2-4	
Elephant endotheliotropic herpesvirus 1A					2,3	
Elephantid herpesvirus 1			1-5		2,3	
Emiliania huxleyi virus			2-5	2,3,5	2-5	
Enterobacteria phage		1,3, 4		1-5	1,3,5	1,3-5
Enterovirus B				5		
Epinotia aporema granulovirus			4			
Equid herpesvirus 3						2,3
Equid herpesvirus 5						2,3
Erwinia phage						5
Escherichia phage						1
Falconid herpesvirus 1						2,3

Feline leukemia virus					3	
Felis catus gammaherpesvirus 1						2
Foot-and-mouth disease virus				2		
Frog virus 3			2-5		2,3,5	
Gallid herpesvirus 2			2-4		3	2,3
Gallid herpesvirus 3			2-5		5	2,3,5
Gayfeather mild mottle virus			2-5	2		
GB virus C	2			2	2	
Gemycircularvirus SL1			2,3			2,3
Gentian ovary ring-spot virus			2,4			
Glypta fumiferanae ichnovirus			1-5		1-5	
Glyptapanteles flavicoxis bracovirus					1-5	
Groundnut ringspot and Tomato chlorotic spot virus			4			
Haloarcula hispanica pleomorphic virus 2				5		
HCBI8.215 virus						2,3
Helicoverpa armigera nucleopolyhedrovirus				1		
Heliothis armigera entomopoxvirus				1		
Hemileuca sp nucleopolyhedrovirus			2,3			
Hepatitis B virus				2,3		
Hepatitis C virus			2-5	3	2,4	
Hibiscus latent Fort Pierce virus			3			
Hirudovirus				3		
Human endogenous retrovirus			2-5	2,3,5	2-5	2-5
Human herpesvirus 1					3	
Human herpesvirus 4	3	3	3	3,5	2-5	3
Human herpesvirus 5		5	5	2,4,5	2-5	5
Human herpesvirus 6			2,3		3	2,3
Human herpesvirus 7			2,3			2,3,5
Human herpesvirus 8			2-4			
Human papillomavirus					2-5	1,2
Human T-lymphotropic virus 1					3	
Hyphantria cunea nucleopolyhedrovirus			3			
Hyposoter fugitivus ichnovirus			1-5		2-5	
Ictalurid herpesvirus 1			1-5		2-5	3
Impatiens necrotic spot virus			2-4			
Influenza A virus			3,5	3		2-5
Invertebrate iridovirus			2,3			
Lactococcus prophage		4				3,4
Listeria phage			4			
Lunk virus NKS-1			2-5		2,5	

Lymantria dispar multiple nucleopolyhedrovirus					3	
Lymphocryptovirus					3	
Lymphocytic choriomeningitis virus					3	
Macacine herpesvirus 1			4			
Macacine herpesvirus 4			2,3,5			3
Magnaporthe oryzae chrysovirus 1			2,3			
Mamestra configurata nucleopolyhedrovirus			3,5			
Mayaro virus						1
Measles virus					2	
Megavirus chiliensis			2,3			
Megavirus lba			2,3			
Meleagrid herpesvirus 1					3	2,3
Microbacterium phage						1
Micromonas pusilla reovirus			2			2
Mimivirus terra2 genome						3
Mollivirus sibericum isolate P1084-T						2
Molluscum contagiosum virus			2-5		2,3	
Mongoose feces-associated gemycircularvirus						2,3
Monkeypox virus			4		2,3	
Moumouvirus			3			
Murid herpesvirus 1				2		3
Murid herpesvirus 2				3		
Musca hytrovirus			2,3,5			
Orgyia leucostigma NPV			2-5			
Orgyia pseudotsugata multiple nucleopolyhedrovirus			2-5		1,2,4,5	
Oryctes rhinoceros nudivirus			2-5	3		
Oryctes rhinoceros virus					2	3
Ostreococcus lucimarinus virus						2
Ovine herpesvirus 2			2,3		2	2,3
Pandoravirus dulcis			1-5		2-5	3
Pandoravirus inopinatum isolate						3
Pandoravirus salinus			1-5	3	1-5	2,4
Paramecium bursaria Chlorella virus				2		
Pectobacterium phage ZF40			3,5			
Penaeus monodon nudivirus			2,4,5			
Penguinpox virus			3			
Penicillium chrysogenum virus			2,3			
Phaeocystis globosa virus			1-5		1-5	2,3,5
Pigeon aviadenovirus A			2-5			
Planaria asexual strain-specific virus-like element			2,3			

type 1						
Polydnaviridae					2-5	
Porcine endogenous retrovirus					3	2
Primate lentivirus group					1	
Prochlorococcus phage				1,4		
Pronghorn antelope pestivirus			4			
Pseudocowpox virus				5		
Pseudomonas phage		4	4	2,4,5	2,4	2-4
Rana grylio iridovirus					2,3,5	
Recombinant HCV viruses					2	
Rhizoctonia solani dsRNA virus 2			3			
Rhodococcus phage						2-4
Saccharomyces cerevisiae killer virus M1			2,3,5			
Saimiriine herpesvirus 1			2-5		2	
Saimiriine herpesvirus 2					3	
Saimiriine herpesvirus 4			2,4			
Salinivibrio phage						4
Sauropus leaf curl disease associated DNA beta			3,5			
Shamonda virus			2,3			2,3
Shigella phage						4
Simbu virus			2,3			2,3
Simian-Human immunodeficiency virus				1,4,5		
Simian immunodeficiency virus				1	1	
Simian virus					3	
Soft-shelled turtle iridovirus					2	
Sorghum chlorotic spot virus			2			
Spodoptera frugiperda ascovirus 1a			2-5			
Spodoptera frugiperda multiple nucleopolyhedrovirus			2,3,5			
Spodoptera littoralis nucleopolyhedrovirus					2,3	
Spodoptera litura granulovirus					5	
Spodoptera litura nucleopolyhedrovirus			2-5			2,3,5
Staphylococcus phage		1		1,3,5	1	1,3
Stealth virus					2-5	
Stenotrophomonas phage						3
Streptococcus phage		3	3	3	3	3
Streptomyces phage						3
Stx2-converting phage						3
Suid herpesvirus 2			3			
Synechococcus phage			1	1		1,2
Tadarida brasiliensis circovirus						2,3

Tembusu virus						2
Thermus phage		5	5	5	5	5
Tick-borne encephalitis virus			2-4			
Tomato aspermy virus			2-5		2,3,5	
Tomato spotted wilt virus			3			
Torque teno virus				4		
Trichoderma atroviride					2,5	
Trichoplusia ni ascovirus 2c			2,3,5			
Trichoplusia ni single nucleopolyhedrovirus				2		
Tunisvirus fontaine2				5		
Tupaiid herpesvirus 1			2,3		2-5	
unclassified Coccolithovirus					2-5	
unclassified Siphoviridae					1-5	
uncultured cyanophage				4		
Upsilonpapillomavirus 1			2,3,5			
Vicia cryptic virus			2-5			
Villosiclava virens					2	
Weeksella virosa DSM 16922					1,2,4	
White spot syndrome virus			2,3			
Woolly monkey sarcoma virus					2,3,5	
Xanthomonas phage					2	
Xestia c-nigrum granulovirus			2			
Xylella phage				1		
Yellowstone lake mimivirus						3

9 ANEXOS

Phylogenetic analyses of chikungunya virus among travelers in Rio de Janeiro, Brazil, 2014-2015

Liliane Costa Conteville[1,2], Louise Zanella[1], Michel Abanto Marín[1], Ana Maria Bispo de Filippis[2], Rita Maria Ribeiro Nogueira[2], Ana Carolina Paulo Vicente[1]/+, Marcos César Lima de Mendonça[2]

[1]Fundação Oswaldo Cruz, Instituto Oswaldo Cruz, Laboratório de Genética Molecular de Microrganismos, Rio de Janeiro, RJ, Brasil
[2]Fundação Oswaldo Cruz, Instituto Oswaldo Cruz, Laboratório de Flavivírus, Rio de Janeiro, RJ, Brasil

Chikungunya virus (CHIKV) is a mosquito-borne pathogen that emerged in Brazil by late 2014. In the country, two CHIKV foci characterized by the East/Central/South Africa and Asian genotypes, were established in North and Northeast regions. We characterized, by phylogenetic analyses of full and partial genomes, CHIKV from Rio de Janeiro state (2014-2015). These CHIKV strains belong to the Asian genotype, which is the determinant of the current Northern Brazilian focus, even though the genome sequence presents particular single nucleotide variations. This study provides the first genetic characterisation of CHIKV in Rio de Janeiro and highlights the potential impact of human mobility in the spread of an arthropod-borne virus.

Key words: chikungunya virus - Asian genotype genome - Brazil

Chikungunya virus (CHIKV) is a mosquito-borne pathogen that belongs to the genus *Alphavirus*, family *Togaviridae*, endemic in parts of Africa, Southeast Asia and on the Indian subcontinent. It usually produces a non-fatal febrile illness in humans, associated with rash and severe arthralgia (Powers & Logue 2007), and occasional neurological manifestations in children (Robin et al. 2008).

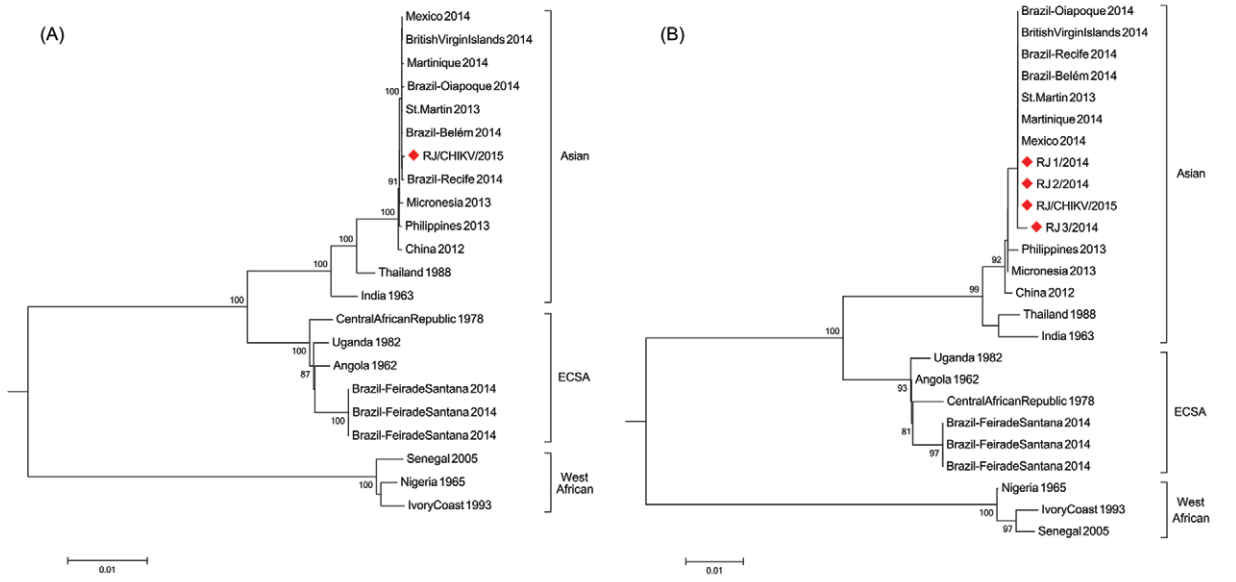
The first autochthonous CHIKV case in the Americas occurred in the Caribbean (Island of Saint Martin) in late 2013 (CDC 2014). After this event, the presence of competent vectors and the intense travel of people led to the establishment of autochthonous CHIKV cases in South American countries, besides Argentina, Chile and Uruguay (Carbajo & Vezzani 2015, PAHO/WHO 2015). In Brazil, imported cases have been reported since June 2014. By September 2014, local transmission of the Asian genotype, the one circulating in the Caribbean, was confirmed in Amapá, northern edge of Brazil. A week later, the East/Central/South African (ECSA) genotype, previously undetected in the Americas, emerged in Bahia state, Northeastern Brazil. Since then, more than 25 thousand suspected CHIKV cases were registered in Brazil (MS 2016).

Until November 2015, Rio de Janeiro state, located in the Southeast of Brazil, had only registered imported CHIKV cases (MS 2016). It is 3,000 km and 1,200 km apart from Amapá and Bahia states, respectively; the current foci of CHIKV in the country. Rio de Janeiro state was predicted to be one of the 35 municipalities with higher risk of CHIKV establishment due to importation from the North and Northeast Brazilian foci (Nunes et al. 2015). By December 2015, the first autochthonous cases were detected in the state (MS 2016).

Here, we performed a phylogenetic analysis of four CHIKV identified in the Rio de Janeiro state in 2014-2015. These CHIKV strains were from individuals with recent travel history to the Caribbean region (three of them were Brazilians with recent travel history to Curacao, Barbados and Dominican Republic, while the other patient is a Venezuelan who came to Brazil). Their main clinical manifestations were fever, arthralgia and exanthema. Whole-genome sequencing was performed on an Illumina HiSeq 2500 system (Oswaldo Cruz Foundation, high-throughput sequencing platform) using 2 x 100 bp paired-end reads generated with Nextera XT libraries. Bioinformatic analyses allowed the recovery of nearly complete genome of the CHIKV virus from the 2015 case (RJ/CHIKV/2015). E1 gene sequences were recovered by PCR and Sanger sequencing from other three 2014 cases. The sequences were submitted to GenBank under accession number KU355832-KU355835. Phylogenetic trees were constructed using Neighbor-Joining method and was evaluated by thousand bootstrap replicates.

The phylogenetic analysis of full-length genomes reveals that the RJ/CHIKV/2015 belongs to the Asian genotype (97-99% identity) and clusters together with other Brazilian imported cases - Guadalupe to Belém, Pará and Dominican Republic to Recife, Pernambuco - and an autochthone case identified in Amapá (middle 2014), as well as with strains from the Caribbean and Mexico (Figure, panel A). Considering this set of genomes, RJ/CHIKV/2015 presents eight unique single nucleotide variations. Four are nonsynonymous: P156S in the methyl-transferase domain and R1307I in the nsP1 C-terminal domain; R1806Q in the nsP3 hypervariable region; and K546R in the B-cell epitope of the E2 protein. As most CHIKV strains, RJ/CHIKV/2015 possess the opal stop codon (TGA) located at the C-terminal of

doi: 10.1590/0074-02760160004
Financial support: CNPq, FAPERJ grants (Project: E-25/010.001558/2014).
+ Corresponding author: anapaulo@fiocruz.br
Received 8 January 2016
Accepted 11 March 2016



Phylogenetic trees based on full-length genome (A) and partial E1 gene (B) constructed with Neighbor-Joining approach. Sequences derived from this study are labeled by gray diamond symbols. Numbers besides internal branches indicate bootstrap values based on 1,000 replicates. Scale bar means base substitutions per site. RJ 1/2014: Venezuelan patient who came to Brazil and had the onset symptoms in 26 September 2014. RJ 2/2014: Brazilian patient with recent travel history to Barbados who had the onset symptoms in 30 September 2014. RJ 3/2015: Brazilian patient with recent travel history to Dominican Republic and had the onset symptoms in 22 June 2014. RJ/CHIKV/2015 is from a Brazilian patient who had the onset symptoms in 3 January 2015, after returning from the Curacao Island.

the nsP3 protein, which has been associated with enhanced CHIKV replication (Chen et al. 2013).

Phylogenetic analysis using partial E1 (435 bp) of the all four imported Rio de Janeiro CHIKV showed that the other three strains also belong to the same Asian genotype cluster (Figure, panel B). The E1 sequences from the Rio de Janeiro travelers are identical, except by T/C synonymous substitution in the RJ_3/2014 strain. Moreover, all of them present Alanine in the position E1-226, as the E1 gene from Asian genotype strains analysed so far, and therefore does not have the mutation that increases CHIKV transmission by *Aedes albopictus* mosquitoes (Tsetsarkin et al. 2007).

This study provides original genomic information of non-autochthonous CHIKV strains identified in travelers coming from the Caribbean region to Rio de Janeiro. This is the second most populous metropolitan area in Brazil and the primary national and international tourist attraction city of the country. Severe outbreaks caused by other arboviruses, Dengue and Zika virus, sharing the same mosquito vector as CHIKV have been occurring in the country as well in Rio de Janeiro (MS 2016). Our results highlight the importance of a genetic surveillance system. So far more than 25 thousand cases have been reported in Brazilian regions, and both the Asian and ECSA genotypes could be circulating in the country (Faria et al. 2016, MS 2016).

ACKNOWLEDGEMENTS

We thank the IOC/FIOCRUZ high-throughput sequencing platform.

REFERENCES

Carbajo AE, Vezzani D. Waiting for chikungunya fever in Argentina: spatio-temporal risk maps. Mem Inst Oswaldo Cruz. 2015; 110(2): 259-262.

CDC - Centers for Disease Control and Prevention. Chikungunya virus. Atlanta: US Department of Health and Human Services; 2014. Available from: cdc.gov/chikungunya.

Chen KC, Kam Y-W, Lin RTP, Ng MM-L, Ng LF, Chu JJH. Comparative analysis of the genome sequences and replication profiles of chikungunya virus isolates within the East, Central and South African (ECSA) lineage. Virol J. 2013; 10: 169.

Faria NR, Lourenço J, de Cerqueira EM, de Lima MM, Pybus O, Alcantara LC. Epidemiology of chikungunya virus in Bahia, Brazil, 2014-2015. PLoS Currents Outbreaks. 2016; Feb 1. Edition 1. doi:10.1371/currents.outbreaks.c97507e3e48efb946401755d468c28b2.

MS - Ministério da Saúde. Boletim Epidemiológico. 2016. Available from: portalsaude.saude.gov.br/.

Nunes MRT, Faria NR, de Vasconcelos JM, Golding N, Kraemer MU, de Oliveira LF, et al. Emergence and potential for spread of chikungunya virus in Brazil. BMC Med. 2015; 13: 102.

PAHO/WHO - Pan American Health Organization/World Health Organization. Number of reported cases of chikungunya fever in the Americas. 2015. Available from: paho.org/chikungunya.

Powers AM, Logue CH. Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus. J Gen Virol. 2007; 88: 2363-2377.

Robin S, Ramful D, Le Seach F, Jaffar-Bandjee MC, Rigou G, Alessandri JL. Neurologic manifestations of pediatric chikungunya infection. J Child Neurol. 2008; 23: 1028-1035.

Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S. A single mutation in chikungunya virus affects vector specificity and epidemic potential. PLoS Pathog. 2007; 3(12): e201.

Parvovirus B19 1A complete genome from a fatal case in Brazil

Liliane Costa Conteville[1,2]/+, Louise Zanella[1], Michel Abanto Marín[1], Ana Maria Bispo de Filippis[2], Rita Maria Ribeiro Nogueira[2], Ana Carolina Paulo Vicente[1], Marcos César Lima de Mendonça[2]

[1]Fundação Oswaldo Cruz, Instituto Oswaldo Cruz, Laboratório de Genética Molecular de Microrganismos, Rio de Janeiro, RJ, Brasil
[2]Fundação Oswaldo Cruz, Instituto Oswaldo Cruz, Laboratório de Flavivírus, Rio de Janeiro, RJ, Brasil

Parvovirus B19 (B19V) infects individuals worldwide and is associated with an ample range of pathologies and clinical manifestations. B19V is classified into three distinct genotypes, all identified in Brazil. Here, we report a complete sequence of a B19V genotype 1A that was obtained by high-throughput metagenomic sequencing. This genome provides information that will contribute to the studies on B19V epidemiology and evolution.

Key words: parvovirus B19 - genotype 1 genome - fatal case

Primate erythroparvovirus 1, previously referred to as Parvovirus B19 (B19V), is a single-stranded linear DNA nonenveloped virus that belongs to the family Parvoviridae and genus Erythroparvovirus (Adams et al. 2014). B19V infects individuals worldwide and is the etiological agent associated with erythema infectiosum, aplastic crisis, hydrops faetalis and arthritis; in rare cases it has been associated to co-infections in human immunodeficiency virus-positive patients, acute leukaemias in children and generalised oedema in adults (Kerr et al. 2003, Pereira et al. 2014, Vlaara et al. 2014).

The genome of B19V is about 5.6 kb with two major open reading frames (ORFs) flanked by two inverted terminal repeats (ITRs), that can be folded into hairpins and are involved in virus replication (Cotmore & Tattersall 2005). One ORF encodes a nonstructural protein (NS1) and the other one, two capsid proteins (VP1 and VP2). VP1 and VP2 share the same amino acid (aa) sequence, but VP1 has an unique region (VP1u) at the amino terminus represented by an additional 227 aas. Besides these major ORFs, there are three minor ORFs that encode NSs: 7.5 kDa, X and 11 kDa. All transcripts are expressed from a single promoter, the p6 promoter (Ozawa et al. 1987, Zhi et al. 2006).

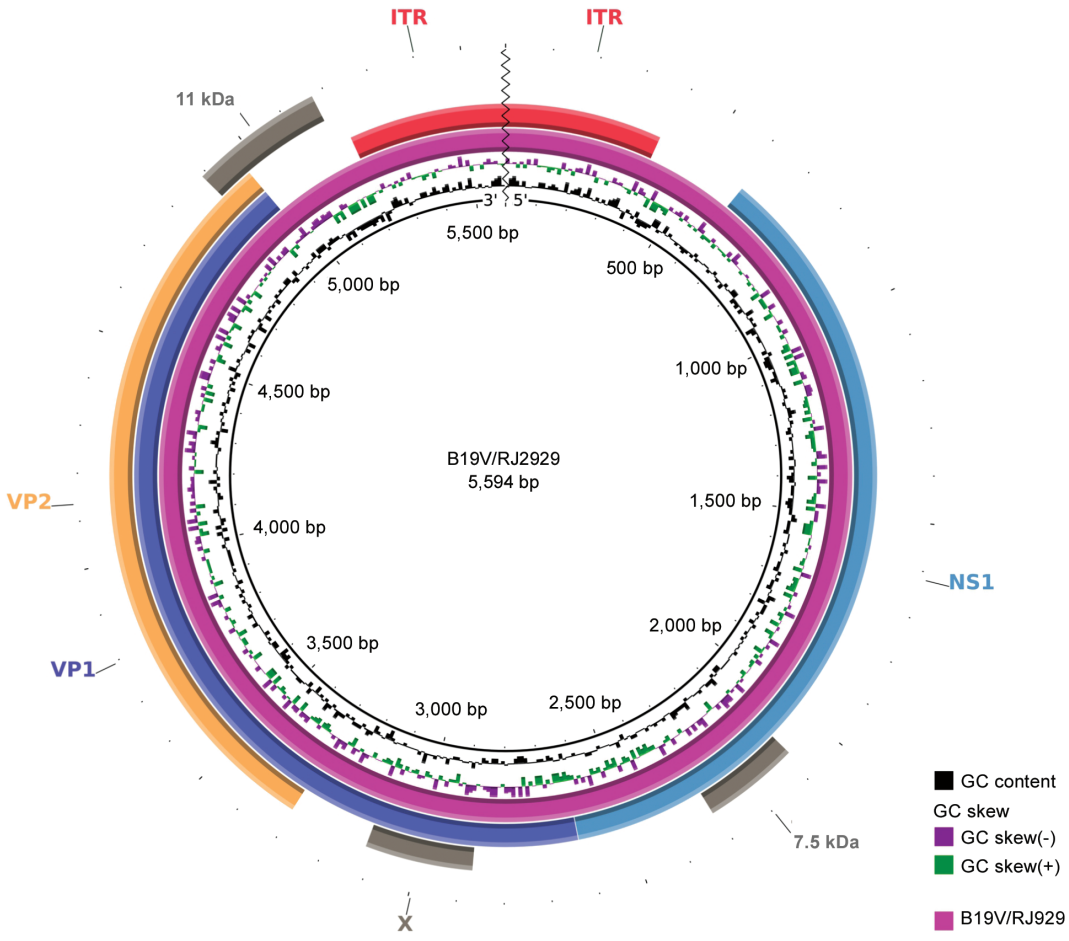
B19V was classified into three distinct genotypes (1, 2 and 3) based on NS1-VP1u region. Genotype 1 was segregated into subtypes 1a and 1b and genotype 3 into subtypes 3a and 3b (Servant et al. 2002, Toan et al. 2006, Parsyan et al. 2007). All three genotypes have been identified in Brazil (Sanabani et al. 2006) but, so far, there are only nearly full-length genome sequences of B19V, most of them are from patients in São Paulo, with different types of leukaemia (da Costa et al. 2013).

In this study we revealed the first full genome of a B19V genotype 1A from a fatal case of a 12-year-old boy from Rio de Janeiro, Brazil with suspected dengue infection. This genome was recovered from a serum sample by metagenomic approach using high-throughput sequencing performed in Illumina HiSeq 2500 platform. Taxonomic profiling programs found hits with similarity to B19V. *de novo* assembly was performed with SPAdes 3.5.0. Specific PCR and Sanger sequencing confirmed the presence of B19V in the sample. Phylogenetic analysis was performed using NS1-VP1-VP2 regions and showed that the B19V/RJ2929 strain belongs to genotype 1A (data not shown).

The B19V/RJ2929 genome is 5,594 bp in length with overall 43.92% GC content (Figure). Contrasting, the ITRs (inferred from published sequence FN598217) have higher GC content (57.85%), resulting in a stable hairpins formation used as a self-primer to start genome replication. All binding sites for transcriptional factors of the p6 promoter are conserved. The comparison of B19V/RJ2929 with B19V 1A sequences available in GenBank revealed some aa substitutions in the major and minor proteins: in NS1, F444C and M452I, two conservative substitutions, in VP1-VP2, two nonconservative substitutions P740R and T741P and in 11 kDa there was one conservative substitution D65N.

This complete genome has been deposited in GenBank under accession KT268312.

doi: 10.1590/0074-02760150261
Financial support: CNPq, FAPERJ (E-25/010.001558/2014)
+ Corresponding author: lilianeconteville@gmail.com
Received 13 July 2015
Accepted 26 August 2015



Genomic map of Parvovirus B19 (B19V). The inner circle represents 5'-3' sequence sense followed by percentual GC content and GC skew. B19V/RJ2929 genome is the purple circle. Major and minor open reading frames and inverted terminal repeats (ITRs) are labelled. Figure was performed using Blast Ring Image Generator (sourceforge.net/projects/brig). NS: nonstructural protein.

ACKNOWLEDGEMENTS

To the Oswaldo Cruz Institute/Oswaldo Cruz Foundation high-throughput sequencing platform.

REFERENCES

Adams MJ, Lefkowitz EJ, King AM, Carstens EB 2014. Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses. *Arch Virol 159*: 2831-2841.

Cotmore SF, Tattersall P 2005. A rolling-hairpin strategy: basic mechanisms of DNA replication in the parvoviruses. In J Kerr, SF Cotmore, ME Bloom, RM Linden, CR Parrish (eds.), *Parvoviruses*, Hodder Arnold, London, p. 171-181.

da Costa AC, Bendit I, de Oliveira AC, Kallas EG, Sabino EC, Sanabani SS 2013. Investigation of human parvovirus B19 occurrence and genetic variability in different leukaemia entities. *Clin Microbiol Infect 19*: e31-e43.

Kerr JR, Barah F, Cunniffe VS, Smith J, Vallely PJ, Will AM, Wynn RF, Stevens RF, Taylor GM, Cleator GM, Eden OB 2003. Association of acute parvovirus B19 infection with new onset of acute lymphoblastic and myeloblastic leukaemia. *J Clin Pathol 56*: 873-875.

Ozawa K, Ayub J, Hao YS, Kurtzman G, Shimada T, Young N 1987. Novel transcription map for the B19 (human) pathogenic parvovirus. *J Virol 61*: 2395-2406.

Parsyan A, Szmaragd C, Allain JP, Candotti D 2007. Identification and genetic diversity of two human parvovirus B19 genotype 3 subtypes. *J Gen Virol 88*: 428-431.

Pereira RFA, Garcia RCNC, de Azevedo KML, Setúbal S, de Siqueira MAMT, de Oliveira SA 2014. Clinical features and laboratory findings of human parvovirus B19 in human immunodeficiency virus-infected patients. *Mem Inst Oswaldo Cruz 109*: 168-173.

Sanabani S, Neto WK, Pereira J, Sabino EC 2006. Sequence variability of human erythroviruses present in bone marrow of Brazilian patients with various parvovirus B19-related hematological symptoms. *J Clin Microbiol 44*: 604-606.

Servant A, Laperche S, Lallemand F, Marinho V, de Saint Maur G, Meritet JF, Garbarg-Chenon A 2002. Genetic diversity within human erythroviruses: identification of three genotypes. *J Virol 76*: 9124-9134.

Toan NL, Duechting A, Kremsner PG, Song LH, Ebinger M, Aberle S, Binh VQ, Duy DN, Torresi J, Kandolf R, Bock CT 2006. Phylogenetic analysis of human parvovirus B19, indicating two subgroups of genotype 1 in Vietnamese patients. *J Gen Virol 87*: 2941-2949.

Vlaara PJ, Mithoeb G, Janssen WM 2014. Generalized edema associated with parvovirus B19 infection. *Int J Infect Dis 29*: 40-41.

Zhi N, Mills IP, Lu J, Wong S, Filippone C, Brown KE 2006. Molecular and functional analyses of a human parvovirus B19 infectious clone demonstrates essential roles for NS1, VP1, and the 11-kilodalton protein in virus replication and infectivity. *J Virol 80*: 5941-5950.