UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
HEMOCENTRO DE RIBEIRÃO PRETO
Programa de Pós-Graduação: Oncologia-Clínica, Células-tronco e Terapia Celular

IAN NUNES VALENÇA

**Desenvolvimento de uma pipeline de bioinformática para a identificação de
doenças infecciosas emergentes em pacientes com regime de transfusão
crônica**

**Ribeirão Preto – SP
2021**

IAN NUNES VALENÇA

Desenvolvimento de uma pipeline de bioinformática para a identificação de doenças infecciosas emergentes em pacientes com regime de transfusão crônica

Versão corrigida. A versão original encontra-se disponível tanto na Biblioteca da Unidade que aloja o Programa, quanto na Biblioteca Digital de Teses e Dissertações da USP (BDTD)

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Oncologia Clínica Células Tronco e Terapia Celular.

Área de Concentração: Bioinformática

Orientador: Prof. Dr. Svetoslav Nanev Slavov

Ribeirão Preto
2021

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Valença, Ian Nunes

Desenvolvimento de uma pipeline de bioinformática para a identificação de doenças infecciosas emergentes em pacientes com regime de transfusão crônica. Ribeirão Preto, 2021.

68 p.: il.; 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Bioinformática.

Orientador: Slavov, Svetoslav Nanev.

1. Metagenômica. 2. Viroma. 3. Hemoterapia. 4. Doenças emergentes. 5. Bioinformática.

Dissertação de autoria de Ian Nunes Valença, sob o título " **Desenvolvimento de uma pipeline de bioinformática para a identificação de doenças infecciosas emergentes em pacientes com regime de transfusão crônica**", apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação de Oncologia Clínica, Células-tronco e Terapia Celular, na área de concentração Bioinformática, aprovada em **14 de Junho de 2021** pela comissão julgadora constituída pelos doutores:

Dr.: Svetoslav Nanev Slavov

Instituição: FUNDHERP (FMRP)

Julgamento: APROVADO

Presidente

Dr^a. Flávia Leite dos Santos

Instituição: HCFMRP

Julgamento: APROVADO

Dr^a.: Marta Giovanetti

Instituição: FIOCRUZ

Julgamento: APROVADO

Dedico esse trabalho aos meus pais, Maria Virgínia Oliveira
Nunes e Sebastião Filgueiras Bastos, pelos esforços e sacrifícios que
me permitiram chegar até aqui.

Agradecimentos

Ao meu orientador **Dr. Svetoslav Nanev Slavov**, por ter feito possível a realização desse trabalho, pelos seus ensinamentos e contribuição para o meu desenvolvimento científico e pessoal

A **Universidade de São Paulo** e à **Faculdade de Medicina de Ribeirão Preto** por me proporcionarem um local de aperfeiçoamento de excelência.

À **Fundação Hemocentro de Ribeirão Preto** pelo fornecimento de um ambiente de extrema qualidade para o desenvolvimento do meu projeto.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (**CAPES**).

À **Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP** pelo apoio financeiro através do processo de número **2019-07861-8**.

“O aspecto mais triste da vida atualmente é que a ciência acumula conhecimento mais rápido do que a sociedade adquire sabedoria.”

- Isaac Asimov

Resumo

VALENÇA, Ian Nunes. **Desenvolvimento de uma pipeline de bioinformática para a identificação de doenças infecciosas emergentes em pacientes com regime de transfusão crônica**. 2021. 65 f. Dissertação (Mestrado em Ciências) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2021.

O risco do surgimento de novas doenças infecciosas tem sido um fator de preocupação há várias décadas. Atualmente, diversas condições como urbanização, desflorestamento, aumento da mobilidade aérea e alta densidade de criação de animais para abate são alguns dos fatores que corroboram para surgimento de doenças infecciosas emergentes em um espaço de tempo cada vez menor. Esse risco potencial é ainda mais acentuado nas regiões equatoriais e tropicais do globo, onde os vírus de origem zoonótica possuem maiores chances de contato com a população humana, desafiando medidas de biossegurança de diversas áreas da saúde, inclusive de bancos de sangue, devido à possibilidade de transmissão transfusional desses agentes. Portanto, esse estudo visa a análise do viroma de pacientes com hemoglobinopatias submetidos a um regime transfusional crônico e hemofilia utilizando uma pipeline de bioinformática para a análise computacional de dados de Sequenciamento de Nova Geração (SNG). Esses pacientes possuem alto risco de aquisição de doenças transmissíveis via transfusão devido ao alto volume de sangue que recebem durante o tratamento clínico. A pipeline utilizada demonstrou êxito na análise bioinformática contendo os seguintes programas para cada um dos respectivos processos: Controle de qualidade (FastQC), Filtragem (Trimmomatic, Prinseq, Deconseq), Classificação e montagem (Kraken 2, SPAdes, BLASTn, BLASTx). Os agentes mais importantes identificados foram classificados utilizando a análise Filogenética. Nesse projeto foram avaliadas 75 amostras, provenientes de pacientes com hemoglobinopatias (30 de pacientes com beta-talassemia e 45 com anemia falciforme) submetidos a regime de transfusão crônica e 5 amostras de pacientes com hemofilia tratados com fatores plasmáticos. Como controle testamos 76 amostras de doadores de sangue. A análise metagenômica identificou vírus pertencentes a duas famílias virais principais: *Anelloviridae* e *Flaviviridae*. Os anelovírus foram os vírus mais abundantes nos pacientes com beta-talassemia. Em segundo lugar nós conseguimos identificar o flavivírus pegivirus humano 1 (HPgV-1, GBV-C ou vírus da hepatite G). No grupo de pacientes com beta-talassemia conseguimos identificar viremia por Hepatite B e Hepatite C. No grupo de pacientes com anemia falciforme, o vírus HPgV-1 foi o mais representativo. Também em pacientes com anemia falciforme conseguimos identificar contaminantes do sequenciamento como plasmídeos, correspondendo majoritariamente a

vetores de clonagem utilizados na prática laboratorial. A abundância viral foi semelhante entre doadores de sangue e pacientes que recebem hemocomponentes o que sugere que o surgimento de vírus que podem ameaçar a segurança transfusional é um evento de baixa probabilidade, porém a análise metagenômica pode ser utilizada para identificar futuros agentes que podem ter impacto na hemoterapia.

Palavras-chave: metagenômica, viroma, hemoterapia.

Abstract

VALENÇA, Ian Nunes. **Development of a bioinformatics pipeline for the identification of emerging infectious diseases in patients submitted to chronic transfusion.** 2021. Dissertação (Mestrado em Ciências) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2021.

The emergence of new infectious diseases has been a factor of concern for several decades. Currently, many conditions, such as the high density of animals for meat consumption, facilitated aerial mobility and the rate of deforestation, corroborate to a reality where new diseases may emerge in increasingly short periods of time, especially in countries located in the tropical regions, where viruses of zoonotic origin have even more favorable conditions to emerge due to the very close contact between the population and the environment. In this scenario, the threat of the emergence of these viruses defies biosafety measures in health systems, including blood banks, due to the possibility of transfusion transmission of these agents. This study aims to develop and improve a pipeline for computational analysis of Next Generation Sequencing (NGS) data from patients with hemoglobinopathies undergoing chronic transfusion regimen. The pipeline introduced demonstrated success in fulfilling three main steps for bioinformatics analysis with the subsequent programs used: Quality control (FastQC), Filtration (Trimmomatic, Prinseq, Deconseq), Classification and assembly (Kraken 2, CLARK, SPAdes, BLASTn, BLASTx) and Phylogenetic Analysis (MAFFT, IQ-Tree, Tree-Puzzle, Bioedit, FigTree). In this project, 75 samples were evaluated, from patients with hemoglobinopathies (30 from patients with beta-thalassemia and 45 with sickle cell anemia) who underwent a chronic transfusion regimen, 76 samples from blood donors as a control group, and 5 samples from patients with hemophilia treated with plasma factors. The implemented pipeline was able to identify two predominant viral families: *Anelloviridae* and *Flaviviridae*. Firstly, anelloviruses were the most representative viruses in patients with beta-thalassemia. Second, we were able to identify the human pegivirus flavivirus 1 or HPgV-1 (formerly referred to as GBV-C or hepatitis G virus). In the group of patients with beta-thalassemia, we were able to identify Hepatitis B and Hepatitis C viruses. In the group of patients with sickle cell anemia, the HPgV-1 virus was the most representative. It was possible to identify agents of contaminating origin as plasmids, mostly corresponding to cloning vectors used in the laboratory. Phylogenetic analysis confirmed the most prevalent genotypes of each of the identified viral groups. There were no significant differences between the viral families identified in the groups of poly-transfused patients and blood donors.

Keywords: Metagenomics, virome, hemotherapy.

Lista de figuras

Figura 1 – Sacerdote do antigo Egito e rosto da múmia do faraó Ramsés V	14
Figura 2 – Vítimas da Peste Negra	15
Figura 3 – Sequenciamento Illumina e Ion torrent	18
Figura 4 – Sequenciamento PacBio	20
Figura 5 – Sequenciamento <i>nanopore</i>	21
Figura 6 – Fluxograma metagenômica	22
Figura 7 – Controle de qualidade FastQC	25
Figura 8 – Algoritmos de montagem de sequências	26
Figura 9 – Pipeline inicial <i>workflow</i>	38
Figura 10 – Pipeline modificada <i>workflow</i>	39
Figura 11 – Árvores filogenéticas anelovírus e HPgV-1 (anemia falciforme)	47
Figura 12 – Árvore filogenética HCV	48
Figura 13 – Árvores filogenéticas HPgV-1 e HBV (beta-talassemia)	49
Figura 14 - Árvore HCV do grupo de hemofilia	

Sumário

1	Introdução	14
1.1	Doenças emergentes.....	14
1.2	Sequenciamento de nova geração.....	16
1.2.1	<i>Sequenciamento de primeira geração.....</i>	<i>17</i>
1.2.2	<i>Sequenciamento de segunda geração</i>	<i>17</i>
1.2.2.1	<i>Sequenciamento HiSeq e MiSeq Illumina</i>	<i>18</i>
1.2.2.2	<i>Sequenciamento Ion Torrent.....</i>	<i>19</i>
1.2.3	<i>Sequenciamento de terceira geração</i>	<i>19</i>
1.2.3.1	<i>Pacific Biosciences (PacBio).....</i>	<i>20</i>
1.2.3.2	<i>Oxford Nanopore</i>	<i>21</i>
1.3	Análise metagenômica.....	22
1.4	Análise bioinformática.....	23
1.4.1	<i>Controle de qualidade e pré-processamento das sequências</i>	<i>24</i>
1.4.2	<i>Montagem de sequências</i>	<i>25</i>
1.4.3	<i>Classificadores taxonômicos.....</i>	<i>26</i>
1.5	Análise filogenética	27
1.5.1	<i>Métodos baseados em distância.....</i>	<i>28</i>
1.5.2	<i>Métodos probabilísticos</i>	<i>28</i>
1.5.3	<i>Análise filogenética aplicada a virologia.....</i>	<i>29</i>
1.6	Metagenômica viral aplicada a hemoterapia.....	30
1.6.1	<i>Terapia transfusional em pacientes com Anemia Falciforme</i>	<i>31</i>
1.6.2	<i>Terapia transfusional em pacientes com Beta-talassemia</i>	<i>31</i>
1.6.2	<i>Terapia transfusional em pacientes com Hemofilia.....</i>	<i>32</i>
2	Objetivos	34
2.1	Objetivos gerais	34
2.2	Objetivos específicos	34
3	Materiais e Métodos.....	35
3.1	Amostras clínicas	35
3.1.1	Pacientes com Beta-talassemia submetidos a transfusões múltiplas	35
3.1.2	Pacientes com Anemia Falciforme submetidos a transfusões múltiplas	35

3.1.3	Pacientes com hemofilia submetidos a tratamento com fatores plasmáticos	36
3.1.4	Critérios de inclusão e exclusão.....	36
3.2	Sequenciamento de nova geração (SNG)	36
3.3	Análise bioinformática dos dados de sequenciamento	37
3.3.1	<i>Pipeline para identificação de vírus emergentes</i>	37
3.3.2	<i>Análise filogenética dos contigs virais mais relevantes</i>	39
4	Resultados	40
4.1	Análise quantitativa de doadores de sangue	40
4.2	Análise qualitativa de amostras de doadores de sangue	40
4.3	Análise quantitativa do viroma em pacientes com hemoglobinopatias.....	41
4.4	Análise qualitativa dos vírus comensais	42
4.5	Análise quantitativa dos vírus de interesse clínico	43
4.6	Análise qualitativa e quantitativa dos pacientes com hemofilia	44
4.7	Artefatos do sequenciamento.....	45
4.7	Análise Filogenética.....	46
4.8	Prevalência de RNA de HPgV-1 entre pacientes de anemia falciforme.....	49
4.9	Prevalência de RNA de HPgV-1 entre pacientes de Beta-talassemia.....	50
4.10	Prevalência de RNA de HCV entre os pacientes com hemofilia.....	50
5	Discussão	52
6	Conclusão	59
	Referências	60

1. INTRODUÇÃO

1.1 Doenças emergentes

Doenças causadas por infecções virais e bacterianas têm acometido a civilização humana desde os primórdios. O primeiro registro documentado de um quadro clínico atribuído a infecção supostamente viral trata-se de um afresco encontrado em Mênfis, capital do antigo Egito, feito aproximadamente no ano de 3700 a.C. Este afresco retrata um sacerdote com sintomas típicos de poliomielite parálitica (Figura 1). Outras evidências sugerem que o faraó Ramsés V (XX dinastia raméssida), morreu em 1196 a.C. em decorrência de uma infecção por varíola cujas lesões podem ser observadas até hoje no rosto da sua múmia preservada em excelente estado de conservação (HOPKINS DR., 1980).

Figura 1. (À esquerda) Representação de um sacerdote do antigo Egito com sintoma característico de poliomielite parálitica. (À direita); O rosto da múmia do faraó Ramsés V (museu nacional egípcio, Cairo), contendo lesões indicando uma infecção pelo vírus da varíola. (Fonte: HOPKINS DR., 1980)

À medida que os conglomerados urbanos foram se formando, a proliferação dessas doenças infecciosas foi facilitada pelo contato mais frequente e intenso entre as pessoas, tornando as epidemias cada vez mais comuns. Em relação a algumas dessas epidemias ainda não se possui consenso sobre o agente causador, como é o caso da Peste Antonina (Império Romano, 165-180 a.C) e a Peste de Atenas (Grécia, Líbia, Egito e Etiópia, 429-426 a.C.) onde se estima que tenham morrido entre 5-10 milhões de pessoas e 75-100 mil pessoas, respectivamente. Pouco tempo depois o mundo viria a testemunhar o que ficou conhecido como a primeira das três grandes ondas de peste bubônica da história, a Peste de Justiniano (541-543

d.C.), responsável pela morte de cerca de 25 milhões de pessoas e influenciando a decadência do império romano. Acredita-se que o causador deste surto foi a *Yersinia pestis*, resultado de um estudo que analisou amostras de material biológico extraído de dentes de duas vítimas enterradas em um cemitério medieval na Alemanha (WAGNER et al., 2014).

A segunda, e maior, epidemia de *Yersinia pestis* ocorreu alguns séculos depois e ficou conhecida como a Peste Negra, ceifando a vida de metade da população européia entre 1347 e 1353.

Figura 2. Cidadãos de Tournai, Bélgica, enterrando os corpos das vítimas da Peste Negra, 1349 d.c. (Fonte: Biblioteca Real da Bélgica).

No século XVI, doenças que circularam na Europa, África e Ásia durante séculos começaram a se disseminar pelo mundo através do movimento populacional que se intensificou com a era das grandes navegações. A varíola, que já era considerada uma doença severa entre os brancos, foi devastadora para os americanos nativos, que nunca haviam sido expostos a esta doença. Dessa forma, a varíola, juntamente com o vírus da influenza e da rubéola, foi determinante para o declínio das populações nativas do Novo Mundo (PATTERSON KB et al., 2002). Na primavera de 1918, relatos de uma nova enfermidade começaram a aparecer em Madri, alertando para uma estranha doença sem vítimas fatais até então. Essa primeira onda,

relativamente branda, viria a infectar aproximadamente 500 milhões de pessoas (cerca de 1/3 da população global na época) e se tornaria conhecida como um dos eventos históricos mais letal que já existiu em termos de contagens de óbitos com cerca de 50 milhões de mortos: a gripe espanhola (FLECKNOE et al., 2018).

Na última década pelo menos oito surtos virais foram responsáveis por centenas de milhares de mortes em diversos países. Atualmente enfrenta-se a maior pandemia de um agente viral da família *Coronaviridae* (SARS-CoV-2) e provavelmente a pandemia com o maior impacto econômico da história humana. A alta densidade da criação de animais para abate, a taxa de destruição das florestas e a intensidade de movimentação através dos transportes aéreos criam um ambiente altamente propício para o surgimento de doenças virais emergentes e re-emergentes (DELWART, 2007). Além disso, atualmente temos um conhecimento muito limitado a respeito da diversidade viral existente habitando o nosso planeta, considerando que apenas 6.590 espécies de vírus foram catalogadas de acordo com o último relatório oficial do Comitê Internacional de Taxonomia Viral, *International Committee on Taxonomy of Viruses* (ICTV) (<https://talk.ictvonline.org/taxonomy>). No entanto, estima-se que existam pelo menos 87 milhões de espécies de vírus que infectam eucariotos (GEOGHEGAN JL et al., 2017) .

Enquanto antigamente a caracterização dos agentes causadores de surtos era inexistente ou muito demorada, atualmente, graças ao surgimento das tecnologias de Sequenciamento de Nova-Geração e o avanço na capacidade computacional de processamento de dados, é possível a rápida e ampla identificação de agentes virais emergentes em praticamente qualquer ambiente.

1.2 Sequenciamento de Nova-Geração

Sequenciamento de Nova-Geração (SNG, do inglês: *Next-generation sequencing, NGS*) refere-se ao conjunto de tecnologias de sequenciamento de DNA de alto desempenho que surgiram algumas décadas atrás como uma consequência natural do sequenciamento de Frederick Sanger introduzido em 1977. Diferentemente do sequenciamento de DNA pelo método Sanger, as tecnologias de SNG são mais rápidas e geram milhões à bilhões de bases em uma única corrida. Como exemplo podemos citar o projeto Genoma Humano, no qual o sequenciamento completo levou 15 anos para ser finalizado a um custo de aproximadamente 3 bilhões de dólares utilizando o método de Sanger (VENTER *et al.*, 2003). Por outro lado, através da aplicação de um dos primeiros equipamentos de SNG i.e. Roche 454 o sequenciamento do genoma do pesquisador norte americano J.D Watson demorou cerca de 2 meses e custou aproximadamente um centésimo do preço do Projeto Genoma Humano. Esses exemplos demonstram a grande importância do sequenciamento de última geração levando-se

em conta, principalmente, a quantidade de dados gerados e o custo mais baixo quando comparado com as técnicas clássicas de sequenciamento. Por estes motivos, e por dar informação em alvos gênicos não pré-definidos, o SNG permite diversas aplicações inclusive na área de metagenômica.

1.2.1 Sequenciamento de primeira geração

As tecnologias de sequenciamento de Sanger e Maxam-Gilbert foram classificadas como as primeiras de sua geração, e iniciaram o campo de sequenciamento de DNA com sua publicação em 1977. O sequenciamento de Sanger é conhecido como o método de terminação de cadeia. Esse método consiste em usar uma das fitas de uma molécula de DNA como um molde de sequenciamento. Esse sequenciamento é feito utilizando nucleotídeos quimicamente modificados chamados de dideoxi-nucleotídeos (ddNTPs) nos quais faltam todos os grupos de hidroxila. Uma vez que os ddNTPs são incorporados a fita, a polimerase é bloqueada e assim são gerados fragmentos de DNA de diversos tamanhos. Os fragmentos são então separados de acordo com o tamanho em um gel e podem ser visualizados através de um sistema de imagem (raios-X, luz UV, laser). O sequenciamento de Sanger foi amplamente utilizado desde sua criação, e até hoje é utilizado para o sequenciamento de regiões específicas de DNA em diversas linhas de pesquisa genômicas (KCHOUK M *et al*, 2017).

1.2.2 Sequenciamento de segunda geração

Podem ser destacadas as seguintes características da tecnologia de sequenciamento de segunda geração: o sequenciamento shotgun de fragmentos genômicos aleatórios de DNA ou cDNA é realizado com o uso de sequências denominadas adaptadores que são ligados aos fragmentos de DNA ou cDNA para construção das bibliotecas genômicas. A amplificação da biblioteca é realizada em uma superfície sólida (*flowcell* da Illumina) ou em emulsão (tecnologia *Ion Torrent*). A incorporação de nucleotídeos é monitorada diretamente pela detecção de fluorescência (Illumina) ou por mudanças na carga elétrica (liberação de H⁺) durante o procedimento de sequenciamento (Ion Torrent). O SNG de segunda geração gera bilhões de leituras de tamanho geralmente curto (100-400 pb) em um menor tempo com relação às tecnologias de primeira geração.

Figura 3. À esquerda representação esquemática do sequenciamento em ponte com adaptadores ancorando as sequências em uma *flow-cell* pelo sequenciador Illumina. À direita representação dos poços utilizados no Ion Torrent onde ocorre a liberação dos prótons. (Fonte: adaptado de *Illumina* e *Thermo Fisher Scientific*)

1.2.2.1 Sequenciamento Illumina (NovaSeq 6000, MiSeq, NextSeq)

O primeiro sequenciador da marca Illumina i.e Solexa foi comercializado em 2007. Atualmente, Illumina é o sistema de sequenciamento mais difundido no mundo e domina a maior parte do mercado, particularmente com as plataformas MiSeq e NovaSeq 6000. O sequenciador Illumina apresenta a tecnologia de sequenciamento por síntese usando terminação de cadeia com nucleotídeos fluorescentes. As bibliotecas de DNA são produzidas na base de fragmentação enzimática (ou tagmentação) e o sequenciamento é realizado via amplificação em ponte gerando assim clusters de sequências. A corrida de sequenciamento Illumina gera uma informação de até 6000 Gb (sequenciador Illumina NextSeq 6000) porém o comprimento das sequências é relativamente curto, aproximadamente 100 bp. Atualmente a Illumina fornece sete máquinas de sequenciamento (iSeq 100, MiniSeq, MiSeq, NextSeq 550, NextSeq 1000/2000, NovaSeq 6000) com alta geração de dados (até 6000 Gb). Os sequenciadores das linhas iSeq 100, MiniSeq e MiSeq são apropriados para sequenciamento de genomas de pequeno porte como de vírus e bactérias. Os sequenciadores NextSeq 550, 1000/2000 e especialmente NovaSeq 6000 são denominados também sequenciadores de produção e geram uma enorme quantidade de bases, aplicável para estudos do transcriptoma, *single-cell*, metilação, metagenômica e biópsia líquida. (BESEER J et al, 2018)

1.2.2.2 Sequenciamento Ion Torrent

A tecnologia Ion Torrent (<http://www.iontorrent.com>) foi desenvolvida pelo grupo de pesquisa que introduziu o sequenciamento 454, introduzindo novas abordagens na área de SNG. No Ion Torrent, os nucleotídeos incorporados são detectados eletronicamente por alterações no pH da solução circundante e, portanto, essa leitura não é dependente da presença de componentes ópticos. A reação de sequenciamento é realizada dentro de um microchip composto de uma *flowcell* que possui sensores eletrônicos na parte basal. A incorporação do nucleotídeo libera prótons que são convertidos em sinais eletrônicos detectados pelos sensores. O chip Ion Torrent possui um sensor de campo elétrico sensível aos íons (H^+) e é projetado para detectar prótons produzidos durante a reação de sequenciamento. O chip é posicionado dentro de uma *flowcell* e é sequencialmente lavado com dNTPs individuais sem marcação na presença da DNA polimerase. A incorporação do respectivo nucleotídeo na cadeia de DNA libera H^+ e altera o pH da solução circundante, o que é proporcional ao número de nucleotídeos incorporados. (KUMAR KR et al, 2019).

No mercado, atualmente, existem dois sistemas de sequenciamento Ion Torrent: Ion Gene Studio S5 e Ion Torrent Genexus System. Os sistemas Ion Gene Studio S5 foram desenvolvidos para abranger várias aplicações de SNG, priorizando velocidade e escalabilidade. Cinco Ion S5 *chips* (*Chips* Ion 510-550) permitem que o rendimento do sequenciamento seja escolhido entre um intervalo de 2 milhões a 130 milhões de sequências por corrida de acordo com o chip utilizado. Já o sistema Ion Torrent Genexus propõe um fluxo de análise integrado, combinando e automatizando todas as etapas do sequenciamento em dois equipamentos (*Purification system* e *Integrated Sequencer*). A análise completa inclui a extração dos ácidos nucléicos, purificação, quantificação, preparação de bibliotecas, sequenciamento, análise dos dados e um relatório com um tempo de processamento total estimado em 24 horas.

1.2.3. Sequenciamento de terceira geração

As tecnologias de sequenciamento de terceira geração revolucionaram o sequenciamento de genomas de alta qualidade. Diferente das tecnologias de segunda geração, que produzem sequências relativamente curtas de algumas centenas de pares de base, a terceira geração, baseada no sequenciamento de moléculas únicas, gera sequências que ultrapassam os 10.000 pb e chegam a moléculas de 100.000 pb. As duas principais plataformas

comercialmente disponíveis que utilizam a tecnologia de terceira geração são a Pacific Biosciences SMRT (Single Molecule Real Time) e a Oxford Nanopore.

1.2.3.1 Pacific Biosciences (PacBio)

A plataforma Sequel da PacBio, que utiliza a tecnologia SMRT, se baseia na detecção em tempo real dos eventos de incorporação de bases nucleotídicas durante o processo de alongamento da fita de DNA molde. Os nucleotídeos utilizados contêm marcadores fluorescentes ligados a cadeia fosfato e são detectados com base no comprimento das ondas emitidas dos diferentes tipos de fluoróforos liberados durante a quebra dessa cadeia. Essa detecção em tempo real ocorre em uma escala nanométrica dentro de várias cavidades com algumas dezenas de nanômetros de diâmetro chamada de ZMW (Zero Mode Waveguide) fabricadas em uma placa de metal depositada sobre uma superfície de vidro. Uma DNA polimerase, ligada a fita molde de DNA, é ancorada na superfície de vidro no fundo de cada ZMW (RHOADS A, 2015).

Figura 4. Método de sequenciamento em tempo real da tecnologia SMRT e o design do *zero mode waveguide* (ZMW) à esquerda (Fonte: adaptado de Eid J, 2009 Science).

As ondas de luz que atravessam a superfície de vidro em cada ZMW iluminam apenas os 30 nm mais profundos de cada poço, uma vez que as dimensões do ZMW são menores que o comprimento de onda da luz. Isso permite que cada nucleotídeo incorporado seja excitado e identificado individualmente em frações de segundos. Sinais de luz das bases que ainda não foram incorporadas e que circundam os poços são incorporados como ruídos de fundo (RHOADS A, 2015).

1.2.3.2 Oxford Nanopore

A tecnologia *nanopore* também se baseia na construção de “buracos” em escala nanométrica. Em seus dispositivos, uma corrente iônica passa através de um nanoporo inserido

em uma membrana com uma determinada voltagem. Quando um analito passa através dessa membrana cria uma distorção na corrente. Como a distorção é característica para cada analito, isso permite uma medição dessa variação para cada base da fita única de DNA ou RNA que passa através do nanoporo, gerando uma leitura de aproximadamente 400 bases por segundo. (MIKHEYEV AS et al, 2014)

Figura 5. Ilustração do princípio de funcionamento de um sequenciador de DNA *nano*pore. Simplificadamente, uma fita de DNA passa através de um nanoporo, a corrente iônica é medida e traduzida na sequência nucleotídica do DNA original. (Fonte: adaptado de www.scienceinschool.org).

O sequenciamento com a plataforma MinION pode ser feito tanto no laboratório como em atividade de campo. Com dimensões reduzidas, o MinION cabe na palma da mão, podendo se conectar a qualquer computador que tenha uma entrada USB e, diferentemente de outras plataformas, permite uma análise em tempo real. Embora o sequenciamento através de nanoporos seja capaz de produzir sequências de até 2 Mb de comprimento, duas das maiores limitações nas primeiras versões deste método eram o baixo rendimento relativo de dados e uma alta taxa de erros (aproximadamente 10%). Recentes avanços nas *flow-cells* (v. 10) sugerem diminuição na taxa de erros com melhorias na precisão do *base-calling* (até 99.99% de precisão) (PETERSEN LM et al., 2019).

1.3 Análise metagenômica

Os estudos metagenômicos têm como base a análise do genoma microbiano completo encontrado em comunidades microbiológicas ambientais, incluindo os vírus (HANDELSMAN J et al., 1998). Os primeiros estudos do viroma avaliavam amostras ambientais como água e esgoto (HANDELSMAN et al., 2004). O progresso do Sequenciamento de Nova Geração (SNG) fez com que houvesse um grande avanço na área de análise de dados de amostras ambientais e uma transferência dessa tecnologia para análise de amostras clínicas com origem humana (sangue, plasma, líquido cérebro-espinhal, e fezes) (FAHSBENDER E et al., 2020; SIQUEIRA JD et al., 2018). Portanto, o sequenciamento de nova-geração é uma ferramenta poderosa que não necessita de um alvo previamente determinado para a identificação de agentes infecciosos desconhecidos em amostras humanas (MINOT S et al., 2011; LYSHOLM et al., 2012, FOULONGNE V et al., 2012) ou de animais (PHAN TG et al., 2011; SHAN et al., 2011).

Figura 6. Fluxograma simplificado de uma análise metagenômica viral. (Fonte: Elaborado pelo autor)

A implementação de métodos moleculares de alta sensibilidade e especificidade para detecção dos agentes infecciosos mais comuns transmissíveis por transfusão sanguínea (HIV, HBV, HCV, HTLV, doença de Chagas e sífilis) melhoraram de uma maneira significativa a vigilância de doenças infecciosas nos bancos de sangue. No entanto, vírus desconhecidos ou não suspeitos podem estabelecer viremia persistente ou aguda assintomática e apresentam uma

ameaça à transfusão sanguínea, uma vez que podem ser transmitidos aos receptores de sangue. Na grande maioria das vezes os receptores de sangue são pacientes com certo grau de imunossupressão, portanto, o quadro clínico causado por agentes virais emergentes pode levar a consequências clínicas graves. Deste modo, a vigilância constante de agentes virais emergentes que podem impactar a segurança transfusional é relevante, especialmente em pacientes com hemoglobinopatias (anemia falciforme, beta-talassemia e hemofilia) que recebem hemocomponentes durante o tratamento e, portanto, apresentam maior risco de aquisição de infecções parenterais.

A aplicação de Sequenciamento de Nova-Geração (SNG) nos mais recentes casos de surtos virais como SARS (SALEMI M et al., 2004), gripe aviária (H5N1)(ZEHENDER G et al., 2012), Ebola (ILUNGA KALENGA O et al., 2019), Zika (FARIA N et al., 2017) e recentemente o SARS-CoV-2 (XAVIER J et al., 2020) tem permitido uma rápida identificação dos meios de transmissão, o conhecimento das taxas de evolução dos vírus e evidências da origem zoonótica desses surtos. Acesso a esse tipo de informação durante os estágios iniciais dessas epidemias pode oferecer informações importantes sobre quando, onde e como uma epidemia pode emergir, permitindo estratégias de intervenção e medidas preventivas mais eficientes (ROSE R et al., 2016). Nesse contexto, as tecnologias de sequenciamento e a análise bioinformática podem ser usadas para o monitoramento desses vírus em diversos tipos de ambientes, permitindo a análise de praticamente todos os vírus que possuam uma quantidade detectável de cópias presentes em uma amostra. Esse método também supera limitações de técnicas tradicionais do estudo de vírus, como o isolamento, cultivo, caracterização microscópica e detecção molecular que são laboriosas e restritas, uma vez que a maior parte dos vírus não pode ser cultivada em laboratório (SAUVAGE V et al, 2012; ZHANG YZ et al, 2018).

1.4 Análise bioinformática

Um passo crucial no uso de dados de sequenciamento de nova geração para a descoberta de novos vírus ou para a composição do viroma é o uso de ferramentas de bioinformática que são necessárias para identificar sequências virais entre a complexa composição genômica proveniente do hospedeiro. Estudos do viroma são mais complexos comparado com os estudos do microbioma bacteriano devido, principalmente, à falta de um gene único, universalmente compartilhado e filogeneticamente informativo, similar ao gene de RNA ribossomal 16S que é presente entre todos os filos de bactérias. Portanto, a identificação de sequências virais depende do alinhamento com sequências virais depositadas em diferentes bancos de dados. No entanto, os vírus demonstram taxas de mutação bastante elevadas (VIRGIN, 2014) e, portanto, os

referidos bancos nem sempre possuem correspondência à grande maioria das sequências presentes nas amostras analisadas (REYES et al., 2010; MOKILI et al., 2012).

O potencial das tecnologias de sequenciamento associado às novas ferramentas para manipulação de grande quantidade de dados (*Big Data*), providenciam oportunidades de novas descobertas e resoluções de questões fundamentais da virologia. Como em diversas áreas da ciência, cada vez mais o estudo da virologia tem se tornado dependente de análises bioinformáticas complexas e de uma cooperação mútua entre bioinformatas, especialistas de laboratório e virologistas. Isto se deve ao fato de que abordagens computacionais altamente específicas estão se tornando intrínsecas às diversas práticas da virologia. (MARZ M et al., 2014).

Comunidades científicas de bioinformática dedicadas ao estudo de vírus ainda são relativamente poucas comparado com os grupos de pesquisa que estudam bactérias (POLLOCK J et al., 2018). Por conta disso, muitas famílias virais ainda são pouco representadas em bancos de dados públicos e, dentre estas, uma parte ainda conta com erros, discrepâncias, redundâncias, ambiguidades e inconsistências com a literatura publicada (BOUADJENEK MR et al., 2017) o que evidencia a necessidade de mais estudos na área, especialmente no que se refere a vírus emergentes. Além disso, o tamanho relativamente pequeno dos genomas virais torna possível o sequenciamento de um grande número de unidades genômicas que requerem um fluxo de análise computacional sequencial para transformar os vários gigabytes de dados em informação biológica, o que leva a necessidade do constante desenvolvimento de *pipelines* de bioinformática para suprir a atual falta de ferramentas confiáveis e acessíveis.

1.4.1 Controle de qualidade e pré-processamento das sequências

A presença de sequências de baixa qualidade (com alta frequência de bases erradas) e de adaptadores utilizados em SNG podem dificultar a análise e a leitura dos resultados obtidos. Dessa forma, a filtragem (*Trimming*) ou o pré-processamento dos dados de sequenciamento pode ser considerada uma prática imprescindível para se evitar resultados que correspondam a padrões abaixo de um nível ótimo. Existem vários programas que realizam essa tarefa com alguns deles sendo específicos para determinadas plataformas de sequenciamento e variando na eficiência com que processam os dados. Alguns desses programas são o Trimmomatic, Trimgalore e o Prinseq.

Figura 7. Exemplo de um gráfico de qualidade gerado pelo FastQC. O eixo-y representa os escores de qualidade. Quanto maior esse escore melhor o *base call*. O fundo do gráfico divide o eixo y em *base calling* de boa qualidade (verde), média qualidade (laranja) e qualidade ruim (vermelho). Essa qualidade tende a diminuir à medida que a corrida do sequenciamento avança. A linha vermelha central representa o valor mediano. As colunas amarelas representam a variação dos interquartis (25-75%). A linha azul representa a qualidade média. (Fonte: FastQC Manual).

1.4.2 Montagem de sequências

Em bioinformática, a montagem de sequências se refere ao alinhamento e junção de fragmentos de uma sequência de DNA mais longa que se sobrepõem objetivando a reconstrução da sequência original. Isso se deve ao fato da maioria das tecnologias de SNG não serem capazes de gerar genomas inteiros em uma única leitura, mas apenas pequenos fragmentos de sequências, como é o caso do sequenciamento Illumina. Normalmente, esses fragmentos são chamados de *reads*. No que diz respeito às diferentes estratégias de montagem de sequências, dois tipos principais podem ser identificados: montagem *de novo* e montagem com genoma de referência. A primeira utiliza os fragmentos curtos para recriar sequências completas (e algumas vezes inéditas) sem a utilização de uma sequência de referência. A segunda alinha os fragmentos contra uma sequência de referência preexistente, construindo uma sequência que é similar, mas não necessariamente igual ao modelo de referência. Em termos de complexidade e tempo computacional, as montagens *de novo* são ordens de magnitude mais lentas e demandantes de memória do que montagens por referência. Isso se deve principalmente ao fato do algoritmo de montagem ter que comparar cada sequência com todas as outras (uma operação que tem um tempo de complexidade de $O(n^2)$). Alguns

programas utilizam abordagens baseadas em grafos de Brujin para superar essa barreira, tornando o processo de montagem *de novo* extremamente viável, como é o caso do SPAdes. O SPAdes inicia o processo de montagem através da utilização de grafos de Brujin de diversos tamanhos (diferentes tamanhos de *k-mers*) reconstruindo o grafo de montagem enquanto detecta e remove sequências quiméricas. Depois, as distâncias entre os *k-mers* são estimadas para determinar os limites do grafo de montagem. Por fim, o grafo de montagem é construído e o SPAdes gera um conjunto de sequências de DNA contíguas (*contigs*). (COMPEAU et al., 2011)

Figura 8. **a.** Representação de um genoma circular simplificado **b.** Sequências representadas como *nodes* em um grafo, e *edges* representados como alinhamentos em algoritmos tradicionais de sequenciamento de Sanger. **c.** Uma forma alternativa de montagem através de um ciclo Hamiltoniano, computacionalmente mais exigente. **d.** Algoritmos de montagem modernos utilizam o ciclo Euleriano. (Fonte: adaptado de Compeau, 2011 Nature Biotechnology).

1.4.3 Classificadores taxonômicos

Para muitas amostras de estudos de metagenômica as espécies, gêneros e até mesmo filos presentes nas amostras são em sua maioria desconhecidos até o momento do sequenciamento. Se um organismo é muito divergente de qualquer genoma previamente depositado nos bancos genômicos, então sua sequência de DNA pode ser considerada como uma sequência desconhecida. No entanto, muitas espécies possuem algum grau de similaridade com os genomas das espécies depositados nos bancos de dados, e essa similaridade pode servir

como um parâmetro para reconstituição do genoma estudado (por exemplo de um vírus emergente não conhecido). Um dos mais conhecidos algoritmos desse tipo e um dos melhores métodos para atribuir classificações taxonômicas a sequências desconhecidas é o programa BLAST, que pode classificar uma sequência através da busca pelo melhor alinhamento em um grande banco de dados de sequências genômicas.

Outros métodos de classificação têm sido propostos, utilizando alinhamento de sequências e técnicas de aprendizado de máquina em uma tentativa de se aperfeiçoarem através da precisão do BLAST. MEGAN, PhymmBL and Naive Bayes Classifier (NBC) são alguns exemplos de programas com diferentes estratégias (menor ancestral comum (LCA), modelos de Markov interpolados e regra Bayesiana, respectivamente). Contudo, um dos melhores programas que combina eficiência computacional e alta sensibilidade é conhecido como Kraken. Essa vantagem em velocidade é alcançada através de sua estratégia de utilização do pareamento exato de *k-mers*, que são subsequências únicas de tamanho *k* contidas em uma determinada sequência biológica. Este pareamento é realizado com as sequências alvos presentes nos bancos de dados, ao invés de um alinhamento inexato de sequências (por exemplo, alinhamento através do BLAST). (WOOD DE et al., 2014).

O Kraken utiliza um banco de dados próprio contendo informações de LCA de todos os organismos cujo genoma está fragmentado em *k-mers* previamente classificados. O banco de dados, construído a partir de uma biblioteca de genomas definida pelo usuário, permite que ocorra uma verificação do nó mais específico na árvore taxonômica que é associado com um *k-mer*. As sequências são classificadas através de uma busca no banco de dados para cada *k-mer* em uma sequência, e então os resultados obtidos do conjunto de dados taxonômicos do LCA são utilizados para determinar uma classificação apropriada para cada sequência. Sequências que não possuem nenhum *k-mer* no banco de dados são deixadas sem classificação pelo Kraken. (WOOD DE et al., 2014)

1.5 Análise filogenética

A análise filogenética se baseia na reconstrução hipotética das relações evolutivas de um grupo de organismos ou um conjunto de sequências (nucleotídeos ou aminoácidos). A rápida expansão dos bancos de dados de sequências nucleotídicas e protéicas, em grande parte devido aos estudos de metagenômica, tem forçado um rápido crescimento na área da metodologia filogenética com a criação de um grande número de ferramentas de bioinformática. Para a construção de uma árvore filogenética a partir das sequências nucleotídicas ou de aminoácidos usando um método que se fundamenta em uma hipótese evolucionária, é necessário construir uma matriz de distâncias baseado em um modelo de substituição nucleotídica. (SCHNEIDER H, 2003). Apesar da existência

de vários métodos de inferência evolutiva, a maioria pode ser classificada em dois grandes grupos: métodos baseados em distância e métodos probabilísticos (parcimônia).

1.5.1. Métodos baseados em distância

Os métodos de classificação evolutiva baseados na medição da distância (também conhecidos como métodos algorítmicos) implementam um algoritmo incorporado de um modelo de evolução (i.e. substituição nucleotídica) que computa uma matriz de distâncias a partir da qual uma árvore filogenética é calculada através de agrupamentos progressivos. Especificamente, as distâncias nesta matriz estão relacionadas com o número de diferenças entre cada par de sequências. O modelo de evolução determina de que forma as substituições nucleotídicas ocorreram a partir do último ancestral comum compartilhado pelas sequências analisadas. A árvore filogenética é reconstruída a partir dos dados da matriz, onde as sequências mais próximas ocupam posições comuns. Agrupamento de vizinhos (*Neighbour joining* (NJ)) e UPGMA (*Unweighted Pair Group Method Using Arithmetic averages*) são exemplos de métodos que utilizam abordagens baseadas em distância (BROCCHIERI, 2001). Apesar de relativamente rápidos, métodos baseados em distância possuem limitações significativas. Por exemplo, NJ gera apenas uma única árvore em contraste com métodos baseados em caracteres, onde uma árvore *consensus* é gerada a partir de várias árvores candidatas. Além disso, o método de NJ pode gerar árvores diferentes, diferentes topologias, dependendo da ordem em que as sequências que constituem o banco de dados foram adicionadas. Uma outra desvantagem se refere à impossibilidade de se identificar as mudanças de caracteres específicas que suportam um determinado ramo (SOLTIS AND SOLTIS, 2003).

1.5.2. Métodos probabilísticos

Métodos probabilísticos buscam pela árvore mais provável em um grupo de sequências de acordo com cada posição dos caracteres no alinhamento de sequências e com o modelo de evolução. Os métodos mais comuns baseados na abordagem de caracteres incluem: Máxima parcimônia (MP), Máxima verossimilhança (ML) e, de forma menos comum, métodos Bayesianos.

O método MP é fundamentado na suposição de que a maior probabilidade de se recuperar a árvore que, de fato, representa a filogenia do grupo é escolher a hipótese que pressupõe o menor número de hipóteses *ad hoc*, isto é, o menor número de mudanças para explicar toda a variação observada na matriz de caracteres (no caso de dados moleculares, as sequências alinhadas). O método baseia-se no princípio da homologia, ou seja, se dois táxons compartilham uma característica é porque esta foi herdada do último ancestral comum a ambos. Ainda que a evolução possa não ser sempre parcimoniosa, o método assume que o critério de parcimônia leva ao maior número total de acertos da árvore verdadeira, quando se minimiza o número de passos evolutivos

aceitos na árvore (SCHNEIDER H, 2003). Uma vantagem do MP é que ele pode gerar um conjunto específico de caracteres para cada clado e comprimento de ramo em termos do número de modificações de cada ramo da árvore.

Métodos de ML são baseados em modelos probabilísticos de evolução e buscam uma árvore com a máxima probabilidade de acordo com esses modelos. (BROCHIERI, 2001). Ao contrário do método de máxima parcimônia, a máxima verossimilhança considera todos os sítios indistintamente.

1.5.3 Análise filogenética aplicada a virologia

A análise filogenética é usada em uma grande variedade de estudos para responder questões fundamentais da pesquisa de vírus, incluindo epidemiologia, diagnóstico, estudos forenses, filogeografia, origem, evolução e taxonomia dos vírus. Uma das primeiras questões a ser respondida durante uma epidemia se trata da origem e identidade do vírus. Essa informação forma a base para a implementação de medidas práticas de contenção e planejamento de resposta ao vírus, permitindo que se tenha uma resposta rápida e específica ao agente viral e, possivelmente, contendo ou minimizando um avanço epidêmico, através do desenvolvimento de drogas antivirais e vacinas.

Entre as diferentes abordagens para a identificação dos vírus no início de uma epidemia, a caracterização filogenética é usada para determinar a relação do vírus recém-descoberto com outros vírus previamente caracterizados e sequenciados. Os resultados dessas análises podem ser suficientes para se obter a informação de interesse, como acontece regularmente com vírus humanos que são monitorados constantemente (i.e., Influenza, HIV, HCV etc.). Para esses vírus, existe um grande banco de dados de isolados previamente caracterizados e cepas que abrangem grande parte da diversidade natural desses vírus. À medida que um vírus identificado pertença a uma dessas espécies conhecidas, é bem provável que ele tenha evoluído de um isolado conhecido ou de uma variante próxima, o que se torna evidente no agrupamento desses vírus em *Clusters* na árvore filogenética (GORBALENYA AE et al., 2017)

Quando uma doença infecciosa emergente é causada por um vírus nunca antes detectado, a análise filogenética é um instrumento para a classificação desse vírus e, no caso de uma infecção zoonótica, é um instrumento para a determinação da dinâmica de introdução do vírus na população humana, o que se dá através da busca pelo reservatório natural do vírus. Esse foi o caso de muitas doenças infecciosas emergentes incluindo as causadas pelo SARS coronavírus (SARS-CoV), MERS coronavírus (MERS-CoV), vírus da Ebola, vírus da Zika (GORBALENYA AE et al., 2017) e atualmente com o SARS-CoV-2.

1.6 Metagenômica viral aplicada a hemoterapia

Um tipo particularmente importante de amostra, que possui potencial de revelar vírus emergentes e circulantes em uma população, envolve o sangue de pacientes com hemoglobinopatias hereditárias e que, por esse motivo, são submetidos à múltiplas transfusões e podem ser considerados recipientes de maior susceptibilidade às doenças infecciosas de transmissão parenteral (GOODNOUGH LT et al., 2017). Atualmente, os centros de transfusão ao redor do mundo aplicam testes de diagnóstico de alta sensibilidade para detectar alguns dos principais tipos de vírus que são conhecidamente transmissíveis pelo sangue, sendo eles os vírus HIV, Hepatite B (HBV), HCV e, em alguns países como Brasil, HTLV. Dessa forma, os procedimentos de hemoterapia, incluindo as transfusões de sangue, são consideradas seguras. No entanto, apesar do risco de transmissão via transfusão dos vírus rotineiramente testados ter sido dramaticamente reduzido, mesmo as técnicas de detecção mais sensíveis não podem anular esse risco completamente. Portanto, ocasionais transmissões desses agentes virais via transfusão podem ser observadas.

Além do risco de transmissão via transfusão dos vírus que são rotineiramente testados, existe um risco significativo de transmissão de praticamente qualquer agente viral capaz de estabelecer viremia. Nesse sentido, as infecções virais emergentes são especialmente importantes e podem ameaçar a segurança transfusional. A lista desses agentes virais é relativamente extensa e varia de arbovírus com viremia de curta duração como Dengue (DENV) (KARIM F et al., 2017; Sabino EC 2016) e Zika (ZIKV) (MOTTA IJ et al., 2016) à infecções com viremia prolongada como Hepatite E (HEV) (TICEHURST JR et al., 2019; RIVIERO-BARCIELA et al., 2017) e Parvovírus tipo 4 (PARV4) (MA YY et al., 2012).

Pacientes com hemoglobinopatias que recebem múltiplas transfusões são constantemente expostos aumentando o risco de adquirir vírus transmitidos por via parenteral. Historicamente, esses pacientes demonstram alta prevalência de HIV, HCV e HBV (Mansour AK et al, 2012). Além disso, devido ao alto volume de sangue transfundido, infecções virais que não são rotineiramente testadas podem estar presentes nesses pacientes. Adicionalmente, pelos métodos atuais, que basicamente envolvem o uso de PCR, é impraticável o teste de todos os vírus conhecidos que são transmitidos por via parenteral devido ao alto custo desses procedimentos e a impossibilidade de se implementar um teste de diagnóstico generalizado.

A caracterização da comunidade viral presente no sangue humano pode ser considerada essencial para a vigilância epidemiológica e para identificar possíveis ameaças à segurança transfusional. Alguns dos vírus comumente detectados no sangue são os Anelovírus (dos três gêneros que infectam humanos: alpha, beta e gama) e o Pegvirus Humano-1 (HPgV-1). Ambos são considerados não-patogênicos e comensais. Outras famílias virais já detectadas no sangue de indivíduos saudáveis incluem os herpesvírus, papilomavírus, poliomavírus, poxvírus e picornavírus

(FURUTA R.A., et al 2015) e provavelmente estes agentes virais representam contaminação viral da pele ocorrendo durante a retirada do sangue do doador. Alguns estudos metagenômicos em amostras de plasma também resultaram na descoberta de novos vírus como os rabdovírus: vírus Ekpoma-1 (EKV-1) e Ekpoma-2 (EKV-2), ambos encontrados em indivíduos saudáveis do Oeste Africano (STREMLAU MH et al., 2015). Dessa forma, uma quantidade crescente de estudos utilizando abordagens de metagenômica tem sido aplicada em diversos tipos de componentes do sangue e de pacientes, visando a caracterização da origem e da composição dos vírus presentes nessas amostras com o intuito de avaliar sua transmissão transfusional. (SAUVAGE V et al., 2017)

O aumento crescente na identificação de novos vírus de origem zoonótica ao redor do mundo, como o atual SARS-CoV-2, sugere que adicionais patógenos provavelmente irão cruzar a barreira entre as espécies tornando-se doenças infecciosas emergentes humanas no futuro. Em vista disso, o uso de SNG e a análise metagenômica viral para caracterizar o viroma de pacientes que recebem múltiplas transfusões pode revelar informações relevantes a respeito da abundância viral bem como a presença de infecções virais emergentes ou não suspeitas com potencial de ameaça para a segurança transfusional ao redor do mundo.

1.6.1 Terapia transfusional em pacientes com Anemia Falciforme

A Anemia falciforme é uma doença hereditária causada por uma mutação pontual no gene da β -globina (GTG $\rightarrow$ GAG) levando a substituição do ácido glutâmico pela valina na estrutura proteica da hemoglobina. Essa mutação produz uma hemoglobina que adquire uma conformação prolongada e rígida impossibilitando a adsorção de oxigênio e confere formas irregulares às células da série vermelha e o característico formato de foice (MAROUF, 2011).

A transfusão de sangue tem sido utilizada como uma opção de tratamento em pacientes com anemia falciforme visando aumentar a capacidade de transporte de oxigênio no sangue, diminuir a viscosidade do sangue e suprimir a falcização das células. Como esse tipo de tratamento é eficiente, com o tempo, a terapia transfusional tornou-se frequentemente utilizada como uma intervenção terapêutica para a profilaxia de complicações da anemia falciforme. No entanto, é possível que pacientes que recebem frequentes transfusões possam hospedar infecções virais emergentes ou não suspeitas.

1.6.2 Terapia transfusional em pacientes com Beta-talassemia

A beta-talassemia é uma das desordens autossômicas recessivas mais comuns ao redor do mundo, com cerca de 1.5% da população global sendo heterozigotos para essa patologia (DE SANCTIS V et al., 2017). Ela possui uma alta prevalência, principalmente em pessoas de origem mediterrânea do Sul e Sudeste da Europa, oriente-médio, transcaucásia, bem como de regiões

geográficas incluindo a Ásia Central, o Subcontinente Indiano, a Indochina e o Extremo Oriente. Os índices mais altos reportados estão no Chipre, Sardenha e Sudeste Asiático. A beta-talassemia também é comum em populações afrodescendentes (WEATHERALL DJ, 2002). A frequência gênica da beta-talassemia nessas regiões está provavelmente relacionada a uma pressão seletiva causada pelo *Plasmodium falciparum*, agente causador da malária, devido a uma correlação entre a presença endêmica da malária nessas regiões, e a distribuição de casos de beta-talassemia. Portadores de beta-talassemia estão relativamente protegidos contra a infecção pelo *Plasmodium falciparum*. (CAO A et al., 2010). Contudo, devido as migrações populacionais e, de forma mais limitada, ao tráfico de escravos do século XV ao XVIII, a beta-talassemia atualmente também se tornou comum no Noroeste Europeu, nas Américas do Norte e do Sul, Caribe e Oceania.

A beta-talassemia é causada por uma redução ou ausência da síntese de cadeias de beta globina no tetrâmero da hemoglobina, que é composta por duas cadeias alfa globina e duas cadeias beta globina. Três condições clínicas e hematológicas com um grau crescente de gravidade são reconhecidas: talassemia menor, talassemia intermediária e beta-talassemia major. A talassemia menor, que resulta da heterozigotidade para a beta-talassemia, é clinicamente assintomática e é identificada por características hematológicas específicas. A talassemia intermediária compreende um grupo, clinicamente e genotipicamente, muito heterogêneo, que abrange desde os pacientes assintomáticos até os que são dependentes de transfusão. Finalmente, a talassemia major compreende os pacientes que são dependentes de transfusão e que, por esse motivo, estão no escopo desse estudo como pacientes de alto risco de aquisição de doenças virais emergentes que podem ameaçar a segurança transfusional.

1.6.3. Terapia em pacientes com hemofilia

A hemofilia é uma desordem hereditária dos componentes do sangue caracterizada pela baixa concentração de fatores de coagulação específicos. Os tipos mais estudados são as deficiências de fator VIII (Hemofilia A) e fator IX (Hemofilia B), ambas ligadas ao cromossomo X. Esses dois tipos de hemofilia são clinicamente indistinguíveis entre si, e o diagnóstico deve ser confirmado através de ensaios específicos para cada fator de coagulação. A tendência de sangramento está relacionada com o nível de concentração do fator e é classificada em leve, moderada e severa.

Atualmente, não existe cura para a hemofilia A ou B. O foco dos tratamentos está em elevar a concentração do fator de coagulação afetado a níveis suficientes para evitar sangramentos traumáticos e espontâneos. Concentrados de fatores VIII e IX derivados de plasma se tornaram disponíveis na década de 70, resultando em um tratamento muito efetivo de episódios de sangramento com a possibilidade de auto infusão e programas de terapia domiciliar. No entanto, a aplicação de fatores de coagulação resultou em taxas muito altas de doença transmissíveis via

hemocomponentes. Com o tempo a produção desses fatores derivados de plasma foram aprimorados, resultando em produtos de maior pureza e consequentemente menos contaminantes. Processos de inativação viral foram introduzidos somente em 1986 para diminuir a transmissão de HIV e do vírus da Hepatite C. Concentrados de fator VIII recombinante (produzido através de engenharia genética) se tornou disponível a partir de 1992, enquanto o fator IX, em 1999. Os produtos recombinantes são preferíveis quando disponíveis, por que eles eliminam o risco de transmissão de doenças infecciosas (BOLTON-MAGGS PH et al., 2003).

OBJETIVO PRINCIPAL

Aplicação de pipeline de bioinformática para avaliar a presença de vírus emergentes, re-emergentes e não suspeitos em dados de sequenciamento de amostras de plasma de pacientes com doenças hematológicas obtidos através de SNG.

OBJETIVOS ESPECÍFICOS PROPOSTOS

- Aplicar pipeline de bioinformática para analisar a abundância viral em pacientes com doenças hematológicas (anemia falciforme, beta-talassemia, hemofilia) recebendo hemocomponentes bem como doadores de sangue sadios.
- Análise filogenética dos *contigs* e *consensus* virais de maior importância;
- Análise comparativa da diversidade viral entre os grupos testados.

3. MATERIAIS E MÉTODOS

3.1 Amostras clínicas

A coleta de amostras foi realizada nos ambulatórios de hemoglobinopatias e hemofilia do Hemocentro de Ribeirão Preto e foi aprovado pelo Comitê de Ética Institucional do Hospital Universitário da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (HCRP 12196/2018).

3.1.1 Doadores de sangue

Foi avaliado o viroma de 76 amostras de plasma de doadores que frequentaram o Hemocentro de Ribeirão Preto no período de junho a dezembro de 2020. A idade média dos doadores foi de 40,5 anos (intervalo 23-58 anos). Esse grupo se dividiu em 45 mulheres (59%) e 31 homens (41%).

3.1.1 Pacientes com Beta-Talassemia submetidos a transfusões múltiplas

Nesse estudo foi avaliado o viroma de 30 pacientes com beta-talassemia que foram atendidos pelo ambulatório do Hemocentro de Ribeirão Preto na Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo. Dos pacientes analisados, 20 são homens (66,7 %) e 10 são mulheres (33,3 %) com uma idade média de 26,6 anos (intervalo: 6-47 anos de idade). A análise sorológica dos vírus testados de rotina nos bancos de sangue revelou 23,3% (n = 7/30) de soroprevalência de anti-HCV IgG e 16,7% de prevalência de anti-Hbs Ag (n = 7/30).

3.1.2 Pacientes com Anemia Falciforme submetidos a transfusões múltiplas

Nós avaliamos o viroma de 45 pacientes com anemia falciforme que foram atendidos pelo ambulatório do Hemocentro de Ribeirão Preto, Faculdade de Medicina de Ribeirão Preto, USP. Desses pacientes, 23 são homens (51,1%) e 22 são mulheres (46.9%) com idade média de 26,2 anos (variando entre 8 a 47 anos de idade). A análise sorológica dos agentes infecciosos rotineiramente testados nos bancos de sangue revelou a presença de infecção por HBV em quatro casos (n=4/45, 8,9 %), um caso de HCV (n=1/45, 2,2 %) bem como um caso de doença de Chagas (n=1/45, 2,2 %).

3.1.3 Pacientes com hemofilia submetidos a tratamento com fatores plasmáticos

Foram avaliados o viroma de 5 pacientes com hemofilia tratados com fatores plasmáticos atendidos pelo ambulatório do Hemocentro de Ribeirão Preto. A idade média correspondeu a 35,8 anos de idade (intervalo: 28-42 anos) e todos os pacientes foram do sexo masculino.

3.1.4 Critérios de inclusão e exclusão

O principal critério de inclusão nos pacientes foi a existência de tratamento regular com bolsas de sangue. Não foram incluídos pacientes com anemia falciforme que não tiveram tratamento transfusional. No estudo de pacientes hemofílicos foram incluídos apenas indivíduos que recebem fatores plasmáticos. Considerando os doadores de sangue, foram incluídos todos os doadores aprovados na entrevista de rotina de doação de sangue.

3.2 Sequenciamento de nova geração (SNG)

Inicialmente 600 µL de plasma obtido de cada paciente (anemia falciforme e beta-talassemia) foi tratado com Turbo DNase (Turbo DNA-free kit, ThermoFisher Scientific, São Paulo, Brasil) para remoção do DNA humano ou bactérias nas amostras analisadas (incubação com DNase a 37°C por 30 min). Posteriormente, após inativação da DNase, as amostras tratadas foram misturadas na forma de pools. O volume total de cada *pool* (aproximadamente 4 mL) foi extraído utilizando *High Pure Viral Nucleic Acid Large Volume Kit* (Roche, São Paulo, Brasil) seguindo as instruções de uso do fabricante com pequenas alterações: o carreador de material genético compondo o kit foi substituído por um carreador de origem sintética derivado de poliacrilamida linear (*GenElute Linear Polyacrylamide carrier* (LPA), Merck, São Paulo, Brasil) e o Isopropanol foi adicionado ao *Ligation Buffer* antes da etapa de adição da amostra para extração na coluna de sílica. Todo processo de extração foi realizado em colunas de sílica conforme as instruções do fabricante e os ácidos nucleicos extraídos foram recuperados em 50 µL de água *nuclease free* pré-aquecida à 70°C e submetidos a transcrição reversa utilizando o *Superscript III First-Strand Synthesis System* (ThermoFisher Scientific, São Paulo, Brasil). A amplificação do cDNA obtido foi realizada utilizando o *QuantiTect Whole Transcriptome Kit* (QIAGEN, São Paulo, Brasil). A preparação das bibliotecas foi feita utilizando o *Nextera DNA Flex*

Library Preparation Kit (Illumina, San Diego, CA, USA) e o *Nextera DNA CD indexes kit* seguindo as instruções do fabricante. Foi realizado sequenciamento de 20 bibliotecas incluindo 6 pools do grupo de talassemia, 7 pools do grupo de anemia falciforme, 1 pool de hemofilia e 6 pools de doadores de sangue. Todos os procedimentos foram feitos utilizando sequenciador Illumina NextSeq 500 com o kit *NextSeq High Output Kit v. 2.5*, 300 ciclos (Illumina, San Diego, CA, USA) e kit NovaSeq SP (2x150) no equipamento NovaSeq 6000 seguindo instruções do fabricante.

3.3 Análise bioinformática dos dados de sequenciamento

3.3.1 Pipeline para identificação de vírus emergentes

A pipeline utilizada nesse projeto foi introduzida através da colaboração com o pesquisador Mathias Vandenbogaert da Unidade de Resposta Urgente contra Ameaças Biológicas do Instituto Pasteur, Paris, França. Inicialmente, as sequências foram processadas para a avaliação da qualidade do sequenciamento utilizando o software FastQC (v.0.11.8). As sequências de baixa qualidade (escore inferior a 30), sequências com bases ambíguas e sequências duplicadas foram removidas utilizando o programa PRINSEQ (v.0.20.4), que utiliza a qualidade atribuída individualmente a cada base nucleotídica como parâmetro de análise. Em seguida foi utilizado o software DeconSeq (v.0.4.3) para a remoção de sequências humanas usando como referência o genoma humano hg38 com limites de identidade e cobertura entre as sequências de 94% e 95%, respectivamente, de forma a reter o máximo possível de sequências virais semelhantes ao DNA humano.

As sequências remanescentes foram classificadas taxonomicamente utilizando o programa Kraken 2 (v.2.0.6), que utiliza uma abordagem baseada em um banco de dados pré-construído de *k*-mers e uma estratégia de detecção do menor ancestral comum (LCA) de cada sequência.

A escolha do comprimento dos *k*-mers foi de 31, alcançando uma alta precisão enquanto mantém uma sensibilidade satisfatória. O banco de dados utilizado para a etapa de classificação inclui genomas de archaea, bactéria, fungos, vírus e *homo sapiens* e foi construído individualmente para cada *software* de classificação. As sequências que não foram classificadas, e que podem representar sequências virais que não estão presentes no nosso banco de dados ou vírus emergentes que são distantemente relacionados com os conhecidos, foram primeiramente analisadas através do Blastx e Blastn contra um banco de dados viral. Grupos de sequências classificadas em táxons de interesse clínico foram submetidos a montagem *de novo* através do software SPAdes e foram analisadas manualmente. Todos os

contigs gerados foram analisados utilizando Blastn com o banco de sequências construído a partir dos dados disponíveis no NCBI utilizando um *cutoff* de 10^{-3} , de forma a preservar o máximo possível de sequências virais que possuem baixa similaridade com o banco de dados existente. Para a avaliação da similaridade protéica entre as sequências foi utilizado o programa Blastx. A pipeline inicial pode ser observada na figura 8.

Figura 9. Representação esquemática da pipeline utilizada na análise de metagenômica.

Objetivando reduzir a quantidade de dados incongruentes e falso-positivos, foram introduzidas algumas mudanças na pipeline original no decorrer da nossa análise. Foi verificado que na etapa de filtragem dos dados ocorria um gasto de tempo de processamento demasiadamente longo, que significava um maior risco de comprometimento da análise por possíveis interrupções no cluster utilizado. Além disso, o programa responsável por esse gasto computacional, Prinseq, não realizava a retirada de algumas sequências de forma eficiente. Portanto, houve a substituição pelo programa Trimmomatic 0.36. Ele utiliza o escore médio de cada sequência como parâmetro e, dessa forma, consegue manter um desempenho superior sem que ocorra perda de dados. De um outro lado o programa Deconseq retirava uma grande quantidade de sequências humanas mas nesse processo foram retiradas sequências de retrovírus endógenos o que poderia comprometer a avaliação do viroma nos nossos grupos de pacientes. Portanto, esse programa foi retirado da pipeline inicial. As modificações introduzidas na pipeline inicial podem ser observadas na figura 9.

Figura 10. Representação esquemática da pipeline utilizada na análise metagenômica viral. As mudanças como a utilização do Trimmomatic no lugar do Prinseq, a remoção do Deconseq, a utilização de um escore de qualidade acima de 30 e a verificação de qualidade antes e após o *trimming* (limpeza das sequencias) foram os principais melhoramentos que foram feitos até o momento.

3.3.2 Análise filogenética dos contigs virais mais relevantes

Para o alinhamento múltiplo dessas sequências, utilizamos o software MAFFT (KATO K et al., 2002) versão 7.450, através de uma seleção automática da estratégia de alinhamento (L-INS-i, FFT-NS-i e FFT-NS-2) de acordo com o tamanho do conjunto de dados submetido. Para verificação do sinal filogenético utilizamos o software TREE-PUZZLE v. 5.3 (SCHMIDT HA et al., 2002) versão 5.3, que utiliza uma representação gráfica de um mapeamento de verossimilhança das sequências alinhadas, indicando, através de uma distribuição de pontos, se o conjunto de dados utilizados possuem um sentido filogenético intrínseco. Em seguida, para reconstrução das árvores filogenéticas, utilizamos o IQ-TREE (Nguyen, Lam-Tung, et al., 2014) versão 1.6.8. A seleção do modelo de substituição nucleotídica utilizado foi feito através do *Bayesian Information Criteria* (BIC) implementado pelo próprio programa. Para suporte estatístico foram utilizados 10.000 *bootstrap replicates* e somente valores de *bootstrap* estatísticos superiores a 75% foram mantidos. Por fim, para visualização da árvore gerada utilizamos o FigTree (RAMBAUT A., 2012) versão 1.4.3.

4. RESULTADOS

4.1 Análise quantitativa de amostras de doadores de sangue

O grupo de amostras dos doadores de sangue teve como principal intuito a obtenção de um controle para comparação com os grupos de pacientes poli transfundidos. O grupo foi constituído de 74 amostras de plasma de doadores de sangue reunidos em um único pool. O sequenciamento Illumina gerou um total de 787.735.815 sequências das quais aproximadamente 0.0001% (105.385 *reads*) foram classificadas taxonomicamente como virais. O restante das sequências foram classificadas majoritariamente como *Homo sapiens* (97%) e bactérias (~2%). O processo de *trimming* removeu aproximadamente 39 milhões das sequências iniciais e um total de 5.050.160 sequências foram consideradas “não-classificadas”. (Tabela 1).

Tabela 1. Dados da análise de sequenciamento das amostras de doadores. Fluxo da análise das sequências partindo de dados brutos a classificação viral taxonômica.

		<i>Sequências totais</i>	<i>Seq. depois do trimming</i>	<i>Seq. classificadas</i>	<i>Seq. virais</i>	<i>Seq. virais (%)</i>	<i>Seq. não- classificadas</i>
Doadores	Pool 1*	159.160.059	123.110.043	118.059.883	11.442	0,007%	5.050.160
	Pool 1	77.037.956	75.137.945	55.110.741	15.202	0.02%	19.889.005
	Pool 2	113.749.616	111.578.345	83.749.512	16.836	0.01%	28.165.370
	Pool 3	96.732.192	94.632.897	69.911.637	14.968	0.01%	24.696.999
	Pool 4	118.266.616	116.786.656	86.690.502	13.227	0.01%	29.726.446
	Pool 5	120.235.932	117.935.987	88.906.258	7.518	0.001%	29.498.424
	Pool 6	102.553.444	100.653.678	75.513.175	26.192	0.02%	25.183.751

*Esse pool é constituído de 16 amostras e foi gerado através do sequenciamento Illumina NextSeq. Demais pools representam 10 amostras cada e foram gerados através do sequenciamento Illumina NovaSeq 6000.

4.2 Análise qualitativa de amostras de doadores de sangue

Como pode ser observado na Tabela 1, a análise das amostras do grupo de 76 doadores de sangue incluído nessa etapa do projeto como controle negativo revelou a presença de dois

principais grupos virais, pertencentes as famílias *Anelloviridae* e *Flaviviridae*. Nós encontramos as seguintes espécies da família *Anelloviridae*: TTV-1-3,5-9,10-20, 21,24,27,29; TTMDV-2,7,9 e TTMV-2. (Tabela 2).

Tabela 2. Número de sequências identificadas para cada agente viral no grupo de doadores de sangue.

Vírus	Grupo Doadores						
	Pool 1*	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6
<i>Anelloviridae</i>	910	14019	13468	14087	12137	7518	24874
Torque Teno vírus	779	12714	12554	11328	11980	4051	24134
Torque Teno midi vírus	30	31	2	9	13	313	49
Torque Teno mini vírus	-	1131	7	390	4	132	571
HPgV-1	9,530	302	2527	391	96	1822	54

*Esse pool é constituído de 16 amostras e foi gerado através do sequenciamento Illumina NextSeq. Demais pools representam 10 amostras cada e foram gerados através do sequenciamento Illumina NovaSeq 6000.

4.3 Análise quantitativa do viroma em pacientes com hemoglobinopatias

Para o grupo de amostras de pacientes com talassemia, o sequenciamento Illumina gerou um total de 808.427.182 sequências para as 30 amostras de plasma coletadas, agrupadas em 6 pools. Foi determinada uma média de 134 milhões de sequências por *pool* com comprimento nucleotídico de 150 pares de base (Tabela 3). Entre as sequências virais classificadas, o maior contig gerado alcançou um máximo de 3.786 pb pertencente ao vírus da Hepatite C (o mapeamento genético revelou que o contig ocupa a posição que se inicia em 3.441 pb e termina em 7.158 pb quando alinhado com o genoma de referência do HCV, NC_038882. Essa região corresponde aos três genes que codificam as proteínas NS3, NS4A e NS4B, e parcialmente a NS5).

Tabela 3. Dados da análise de sequenciamento. Fluxo da análise das sequências partindo de dados brutos a classificação viral taxonômica.

		<i>Sequências totais</i>	<i>Seq. depois do trimming</i>	<i>Seq. classificadas</i>	<i>Seq. virais</i>	<i>Seq. virais (%)</i>	<i>Seq. não- classificadas</i>
Beta-talassemia	Pool 1	126.241.206	112.680.449	106.827.862	2.642	0.002%	5,852,587
	Pool 2	140.189.184	120.822.545	113.278.558	141.841	0.11%	7,543,987
	Pool 3	123,039,880	109,175,346	102,493,334	26.556	0.02%	6,682,012
	Pool 4	119,051,092	107,960,664	101,267,675	3,414	0.003%	6,692,989
	Pool 5	139,336,406	128,896,726	119,770,867	3,673	0.003%	7,125,859
	Pool 6	160,569,414	78,111,788	73,753,074	3,555	0.002%	4,358,714
Anemia Falciforme	Pool 1	146,101,592	111,012,990	108,151,056	2,413	0.002%	2,861,934
	Pool 2	133,479,112	76,858,174	76,267,771	36,385	0.04%	590,403
	Pool 3	137,453,466	115,461,556	112,191,344	4,164	0.004%	3,270,212
	Pool 4	135,904,442	114,650,258	113,837,030	28,336	0.02%	813,228
	Pool 5	134,316,328	116,530,849	115,998,000	1,265	0.001%	532,849
	Pool 6	132,184,180	83,628,290	83,234,978	3,578	0.004%	393,312
	Pool 7	134,044,498	117,973,640	117,365,090	3,597	0.003%	608,55
Hemofilia	Pool 1	46.650.422	44.272.173	32.550.213	3.320	0,01%	2.378.247

No grupo de pacientes com anemia falciforme, o protocolo de sequenciamento gerou um total de 953.483.618 sequências a partir de 45 amostras de plasma que foram agrupadas em 7 pools, o número de amostras corresponde ao total de pacientes com anemia falciforme recebendo transfusões periódicas no Hemocentro de Ribeirão Preto. Nós obtivemos uma média de 136 milhões de sequências que foram geradas para cada pool analisado (variação: $132 - 146 \times 10^6$ sequências). O processo de filtragem das sequências removeu aproximadamente 31×10^6 sequências e as sequências restantes foram submetidas a classificação taxonômica através do Kraken 2. (Tabela 3).

4.4 Análise qualitativa dos vírus comensais

Entre os vírus detectados, três famílias virais principais foram representadas: *Anelloviridae*, *Flaviviridae* e *Hepadnaviridae*. A maior parte das sequências virais (> 95%)

correspondeu aos anellovírus e foram principalmente atribuídas à espécie Torque Teno Vírus (TTV) do gênero Alphatorquetenovírus, sendo detectadas em todos os *pools*. Outros gêneros da família *Anelloviridae*, incluindo Gammatorquevírus (Torque Teno Midi Vírus, TTMDV) e Betatorquevírus (Torque Teno Mini Vírus, TTMV) também foram representados (1,5 – 16,7% de todos os anellovírus presentes). HPgV-1 (conhecido como GB vírus C ou vírus da Hepatite G) também foi encontrado em três *pools*, com um número particularmente alto de sequências no pool 2 (158.790 reads) (tabela 2). A reconstrução filogenética da maior sequência consensus do HPgV-1 (1.360 pb, o alinhamento inicia na posição 2.228 pb e termina na posição 3.554 pb quando alinhado com o genoma de referência do HPgV-1 NC_001710 correspondendo ao gene p7-NS2) demonstrou que ele pertence ao genótipo 2 (subgenótipo 2A) (figura 12).

Nas amostras de pacientes com anemia falciforme, a composição do viroma demonstrou a presença de agentes virais comensais como representantes da família *Anelloviridae* bem como do vírus da hepatite G: HpGV-1 (*Flaviviridae*). Dentre os anellovírus, o gênero Alphatorquevirus foi o mais representativo e foi possível identificar várias espécies de Torque Teno vírus (TTV). De forma similar, porém menos abundantes que o TTV, dois outros gêneros dos anellovírus também foram representados: Betatorquevirus (Torque teno mini vírus, TTMV) e Gammatorquevírus (Torque teno midi vírus, TTMDV). O vírus HPgV-1 foi o mais abundante na maioria dos pools (com exceção do pool 1 e 5). A contagem de sequências de todos os vírus identificados pode ser visto na tabela 4.

Tabela 4. Número de sequências identificadas para cada agente viral no grupo de pacientes com anemia falciforme.

Grupo Anemia Falciforme							
Vírus	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	Pool 7
<i>Anelloviridae</i>	185	8,444	1,969	1,713	1,202	285	259
Torque Teno vírus	166	7,576	441	1,591	1,021	260	238
Torque Teno midi vírus	16	774	1,240	67	113	14	11
Torque Teno mini vírus	3	23	-	-	12	8	8
HPgV-1	1,765	27,738	1,707	26,441	-	3,170	3,152

4.5 Análise qualitativa de vírus de interesse clínico

No grupo de pacientes com beta-talassemia foi possível detectar vírus que não são considerados comensais. Uma quantidade significativa de sequências (1.390) no pool 6 que foram classificadas como pertencentes ao vírus da hepatite C (HCV, família *Flaviviridae*). A

detecção de HCV foi esperada, uma vez que três das amostras, que compuseram o pool positivo, pertenciam a pacientes infectados com HCV (Tabela 5). No entanto, apenas um paciente demonstrou carga viral para HCV caracterizada por 56.793 IU/ml. A análise filogenética do maior contig obtido, 3.786 pb, indicou que o vírus da Hepatite C pertence ao genótipo 1A. A pipeline de análise bioinformática também foi capaz de detectar 8 sequências pertencentes ao vírus da hepatite B (HBV, família *Hepadnaviridae*) no pool 4. A análise filogenética da sequência de 244 pb do HBV demonstrou que o vírus pertence ao genótipo A (Figura 12).

Tabela 5. Número de sequências identificadas para cada agente viral no grupo de pacientes com anemia falciforme.

Grupo Beta-Talassemia						
Vírus	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6
<i>Anelloviridae</i>	1,524	8,594	25,554	2,440	2,271	1,492
Torque Teno vírus	1,454	2,727	24,187	1,872	1,706	1,416
Torque Teno mini vírus	52	1,232	827	408	313	23
Torque Teno midi vírus	5	251	82	42	132	17
HPgV-1	-	130,290	5	26	-	-
Hepatitis B	-	-	-	8	-	-
Hepatitis C	-	-	-	-	-	1,390

4.6 Análise qualitativa e quantitativa dos pacientes com hemofilia

O sequenciamento Illumina (NovaSeq 6000) do grupo de pacientes com hemofilia obteve um rendimento de 46.650.422 sequências representativas de 1 único pool de 5 amostras de plasma. Do total de sequências geradas, 44.272.173 foram retidas para classificação e 2.378.247 foram excluídas durante o processo de filtragem. 32.550.213 sequências foram taxonomicamente classificadas, onde a maior parte (73,52%) foi atribuída a espécie *Homo sapiens* e apenas 0.01% a vírus. As famílias virais mais representativas foram *Anelloviridae* e *Flaviviridae*. A espécie viral mais abundante foi representada pelo Torque Teno vírus (*Anelloviridae*). Dentro da família *Flaviviridae* o vírus da Hepatite C teve uma quantidade expressiva de *reads* atribuídas taxonomicamente (448), sendo possível a recuperação de um contig de 1175 pb de comprimento. O flavivírus HPgV-1 também foi identificado nesse grupo, porém com uma representatividade inferior (41 *reads*) (Tabela 6).

Tabela 6. Número de sequências identificadas para cada agente viral no grupo de pacientes com hemofilia.

Grupo de Hemofilia	
Vírus	Leituras
<i>Anelloviridae</i>	2506
<i>Alphatorquevirus</i>	2454
<i>Gamatorquevirus</i>	14
<i>Betatorquevirus</i>	14
HPgV-1	41
HCV	448

4.6 Artefatos do sequenciamento

Nos dois grupos de pacientes com hemoglobinopatias (Anemia falciforme e Beta-talassemia) e nos doadores de sangue também detectamos algumas sequências com baixa porcentagem de similaridade nucleotídica com o banco de dados de NCBI. Essas sequências poderiam ser identificadas como vírus emergentes ou serem apenas artefatos do sequenciamento. O processo de classificação taxonômico identificou sequências potencialmente pertencentes as famílias *Marseilleviridae*, *Phycodnaviridae* e *Mimiviridae*, da ordem *Megavirales*. Um vírus particular, Shamonda vírus (gênero Orthobunyavirus, família *Peribunyaviridae*), teve um número considerável de sequências em todos os pools. Contudo, nenhum desses vírus foi confirmado quando analisados individualmente pelo Blastn, demonstrando uma baixa cobertura variando entre 20 e 30%. A análise através do Blastx não encontrou nenhuma similaridade significativa para a maioria das sequências, enquanto uma outra parte também foi relacionada com *homo sapiens* ou bactérias como *Ralstonia solanacearum*. Nós também identificamos outras sequências virais classificadas como Dengue, Guanarito mamarenavírus, vírus da febre hemorrágica Congo-Crimeia e o vírus Epstein-Barr, porém, estes agentes virais foram confirmados como falso-positivos através de análise manual.

Tabela 7. Informações genômicas dos vírus encontrados e contaminações do sequenciamento.

Vírus	Tamanho do contig (pb)	Posição no Genoma (pb)		Genoma de Referência	Gene
		Início	Fim		
<i>Hepatite B</i>	244	1.610	1.853	NC_003977	Proteína X
<i>Hepatite C</i>	3.786	3.441	7.158	NC_038882	NS3, NS4A e NS4B
<i>HPgV-1</i>	1.360	2.228	3.554	NC_001710	p7-NS2
Contaminações					
<i>Parvovírus B19</i>	~200	-	-	KC013305	VP2
<i>HIV</i>	~200	-	-	MK801288	RRE
<i>Indiana Vesiculovírus (VSV)</i>	~200	-	-	NC_038236	VSV-G
<i>Vírus da Leucemia Murina (MLV)</i>	~200	-	-	NC_001362	Env e gag
<i>Citomegalovírus (CMV)</i>	~200	-	-	MT044485.1	CMV-MIE-enhancer

4.7 Análise filogenética

Sequências *consensus* dos vírus mais abundantes foram analisadas filogeneticamente. Um total de 5 árvores filogenéticas foram reconstruídas: duas referentes ao HPgV-1, contendo genomas completos de HPgV-1, um referente ao Torque Teno vírus, contendo genomas parcialmente completos de TTV, uma árvore de HBV e uma de HCV.

Figura 11. Reconstrução de árvores filogenéticas através do método da Máxima Verossimilhança de Anelovírus (esquerda) e HPgV-1 (direita) no grupo de pacientes com anemia falciforme. identificados pela análise de metagenômica viral em pacientes com anemia falciforme. O modelo de substituição nucleotídica utilizado foi o *General Time Reversible* (GTR+I+G) selecionado através do *Bayesian Information Criteria* (BIC) implementado pelo programa IQtree v. 1.6.2. Para suporte estatístico foram utilizados 10.000 *bootstrap replicates* e somente valores de *bootstrap* superiores a 75% foram utilizados.

1 - Análise filogenética de HPgV-1: Para as duas árvores de HPgV-1 foram utilizados 45 genomas completos, representando todos os 7 genótipos de HPgV-1. No grupo de pacientes com anemia falciforme e talassemia todos os 4 genomas analisados de HPgV-1 foram classificados como genótipo 2 (subgenótipo 2A). Dentro desse cluster, o genoma obtido do grupo de pacientes com talassemia foi agrupado com a sequência MH053115, um isolado proveniente de um estudo com um grupo de doadores da França. Os outros três genomas obtidos do grupo de pacientes com anemia falciforme foram agrupados próximos as sequências AX338086 e AF309966 provenientes dos Estados Unidos e Alemanha, respectivamente. Os valores de *bootstrap* obtidos nesses clados foram de 100%. (Figura 10 e 12)

2 - Análise do TTV: as sequências de TTV provenientes de pacientes com anemia falciforme foram alinhadas com um bando de dados construído a partir de genomas completos depositados no NCBI. A árvore dos anelovírus consistiu em 29 genomas completos representando o gênero *Alphatorquevirus* (TTV 1-29), 9 genomas representando o gênero

Betatorquevirus (TTMV 1-9) e 9 genomas representando o gênero *Gamatorquevirus* (TTMDV 1-9). As três sequências *consensus* de TTV filogeneticamente analisadas foram classificadas como TTV 6, 10 e 12, o que corresponde aos tipos previamente classificados no viroma de pacientes com anemia falciforme em nossa análise. (Figura 10)

3 - Análise de HBV: Dentro do grupo de pacientes com talassemia a árvore de HBV (composta de 32 genomas completos) resultou na classificação da única sequência de HBV no genótipo A. Nossa sequência foi agrupada com as sequências M57663 e KC875259, isoladas de pacientes da Índia e das Filipinas, respectivamente. O menor valor de *bootstrap* representado nessa ramificação foi de 96%. (Figura 12).

4 - Análise de HCV: A análise filogenética da sequência de HCV recuperada do grupo de pacientes com beta-talassemia revelou a presença do genótipo 1 (subgenótipo 1A) nesse grupo. Nossa sequência foi agrupada com a sequência AF009606 isolada nos Estados Unidos, com um *bootstrap* de 100%. O banco de dados da árvore filogenética do HCV foi construído com 36 genomas completos. (Figura 11).

Figura 12. Árvore filogenética do HCV no grupo de pacientes com beta-talassemia. Reconstrução de árvores filogenéticas através do método de Máxima Verossimilhança do maior contig de HCV mostram que os vírus detectados foram atribuídos aos genótipos mais prevalentes circulando pelo Brasil. O modelo de substituição nucleotídica utilizado foi o *General Time Reversible* (GTR+I+G) selecionado através do *Bayesian Information Criteria* (BIC) implementado pelo programa IQtree v. 1.6.2. Para suporte estatístico foram utilizados 10.000 bootstrap replicates e somente valores de bootstrap superiores a 75% foram considerados.

Figura 13. Árvore filogenética do HPgV-1 (esquerda) e HBV (direita) no grupo de pacientes com beta-talassemia. Reconstrução de árvores filogenéticas através do método de Máxima Verossimilhança do maior contig de HCV mostram que os vírus detectados foram atribuídos aos genótipos mais prevalentes circulando pelo Brasil. O modelo de substituição nucleotídica utilizado foi o *General Time Reversible* (GTR+I+G) selecionado através do *Bayesian Information Criteria* (BIC) implementado pelo programa IQtree v. 1.6.2. Para suporte estatístico foram utilizados 10.000 *bootstrap replicates* e somente valores de *bootstrap* superiores a 75% foram utilizados.

4.8 Prevalência de RNA de HPgV-1 entre pacientes de anemia falciforme

O RNA de HPgV-1 foi detectado por PCR em 8 das 45 amostras (17,8%). A idade média dos pacientes positivos foi de 27 anos (intervalo: 8-57 anos, desvio padrão=18,24). No que se refere a distribuição de gênero, seis (75%) dos pacientes positivos para RNA de HPgV-1 eram do sexo feminino enquanto dois pacientes (25%) eram do sexo masculino. Três dos pacientes positivos demonstraram coinfeções: um HCV (anti-HCV IgG), um doença de Chagas (anti-*Trypanosoma cruzi* IgG) e um HBV (anti-HBs Ag). Considerando a distribuição étnica, cinco (71%) dos sete pacientes positivos se auto declararam afrodescendentes e dois (28%) se declararam brancos.

Todas as sete amostras positivas foram sequenciadas e analisadas filogeneticamente. O modelo de substituição nucleotídica utilizado foi o HKY+F+G4. O filograma demonstrou que seis sequências foram classificadas como genótipo 2 (1 subgenótipo 2A e 5 subgenótipos 2B) e uma sequência foi classificada como genótipo 1. O suporte estatístico por *bootstrap* dos

grupos formados foi acima de 75% e foi observado uma sequência que poderia ser classificada como *outlier* no genótipo 2A, o que significa que existe uma divergência molecular dentro desse genótipo. As amostras que foram classificadas como subgenótipo 2B formaram um grupo monofilético com a cepa MK684252 com origem da Amazônia brasileira. Isso demonstra que as cepas de HPgV-1 brasileiras tem uma grande similaridade e provavelmente são endêmicas no país. A cepa de HPgV-1 que foi classificada como subgenótipo 2A se agrupou com a cepa U63715, que foi isolada nos Estados Unidos de um paciente com origem do leste africano. A cepa de HPgV-1 classificada no genótipo 1, agrupou com um alto valor estatístico com a cepa AB013500 que foi isolada de indivíduos de Gana, África, o que está de acordo com a origem africana do genótipo 1.

4.9 Prevalência de RNA de HPgV-1 entre pacientes com Beta-talassemia

A prevalência de RNA de HPgV-1 obtida entre os pacientes com beta-talassemia foi de 14,3% (6 das 42 amostras analisadas). A idade média dos pacientes que testaram positivo para o HPgV-1 foi de 22,7 anos de idade (intervalo: 12-37 anos). Três das amostras positivas pertenciam a crianças com idades de 12, 13 e 14 anos. A distribuição por gênero se deu da seguinte forma: 33,3% (n = 2/6) do gênero feminino e 66,7% (n = 4/6) do gênero masculino.

4.10 Prevalência de RNA de HCV entre os pacientes com hemofilia

A prevalência de RNA de HCV entre os pacientes com hemofilia tratados com fatores plasmáticos foi de 60% (3 das 5 amostras analisadas). A análise por PCR em tempo real nas amostras individuais revelou pelo Ct que em todas as ocasiões se tratava de infecção com carga viral relativamente alta (Ct médio 24,0 e Cts individuais 22,8; 23,7 e 25,6). A idade média correspondeu a 35,8 anos de idade (intervalo: 28-42 anos) e todos os pacientes foram do sexo masculino.

Figura 14. Reconstrução da árvore filogenética do HCV através do método da Máxima Verossimilhança do grupo de pacientes com Hemofilia identificados pela análise de metagenômica viral. O modelo de substituição nucleotídica utilizado foi o TIM2+F+R5 selecionado através do *Bayesian Information Criteria* (BIC) implementado pelo programa IQtree v. 1.6.2. Para suporte estatístico foram utilizados 10.000 *bootstrap replicates*.

5. DISCUSSÃO

Nesse estudo realizado com pacientes brasileiros com hemoglobinopatias e hemofilia, foi avaliado o viroma em indivíduos altamente dependentes de procedimentos parenterais. Dessa forma, utilizamos SNG e técnicas de bioinformática e programação com o objetivo de determinar a diversidade viral desses grupos de pacientes e identificar possíveis novos agentes virais que podem ser transmitidos por transfusão de sangue e podem ser considerados como potenciais ameaças à segurança transfusional. As razões para nosso estudo incluem: (i) localização da região analisada em uma área endêmica para agentes arbovirais (com destaque principalmente para Dengue) que podem ser potencialmente transmitidos via transfusão sanguínea e desencadear quadros clínicos graves em pacientes que sofrem com doenças hematológicas (SLAVOV SN et al., 2017; DE SOUZA PEREIRA BB et al., 2017); (ii) a caracterização do viroma em pacientes nas regiões tropicais pode revelar a presença de possíveis infecções virais emergentes e não suspeitas nos pacientes estudados e desta maneira o nosso estudo pode contribuir para o monitoramento em pacientes com distúrbios hematológicos. Portanto, pela avaliação do viroma nos pacientes com múltiplas transfusões nós podemos verificar a abundância viral nesses pacientes que têm um potencial de fornecer informações importantes para os serviços de transfusão ao redor do mundo.

Em todas as amostras analisadas foi possível encontrar vírus que estabelecem viremia crônica e são transmitidos parenteralmente. Entre eles, os mais abundantes são os pertencentes à família *Anelloviridae*, que foi representada pelos gêneros Alphatorquetenovírus, Betatorquetenovírus e Gammatorquetenovírus. Esse é o grupo mais abundantes de vírus encontrado em amostras de sangue analisadas através de abordagens metagenômicas em outros grupos de pesquisa (SAUVAGE V et al., 2016; BAL A et al., 2018; PAN S et al., 2018; BRITO F et al., 2018). Os resultados obtidos estão de acordo com a diversidade genômica e mostra uma prevalência extremamente elevada na população humana alcançando quase 100% (AL-QAHTANI AA et al., 2016; BIAGINI P et al., 2006; BERNARDIN F et al., 2010). Apesar dessa alta prevalência de anelovírus na população em geral, a forma como eles interagem com o hospedeiro ainda não é completamente elucidada.

Outro vírus largamente distribuído detectado pela nossa pipeline foi o HPgV-1, que teve um número elevado de sequências em alguns dos pools testados. A análise filogenética dos maiores contigs revelou a presença do genótipo 2 do HPgV-1 (subgenótipo 2A), o tipo que mais prevalente no Brasil (ALCALDE R et al., 2010; RAMOS FILHO R et al., 2004;

NISHIYA AS et al., 2003). Similar aos anellovírus, HPgV-1 é altamente prevalente nos diferentes grupos de pacientes, especialmente nos pacientes com alto risco de exposição parenteral devido a suas rotas de transmissão (SANTOS ML et al., 2017; DA MOTTA LD et al., 2016, SEIFRIED C et al., 2004, ASTER V et al., 2005). Dessa forma, nosso estudo está de acordo com outros estudos que examinaram o viroma de plasma e demonstraram que o HPgV-1 e os anellovírus representam o tipo de vírus comensais mais comumente encontrado em plasma humano (ZHANG W et al., 2016).

No grupo de pacientes com beta-talassemia foi possível identificar vírus com transmissão parenteral como o HCV e HBV. Porém, a detecção de HCV nos pools testados não foi inesperada, devido à alta prevalência de anti-HCV IgG no grupo de pacientes (23.3%). A análise filogenética do maior contig do HCV (3.786 pb) revelou que ele pertence ao genótipo 1A. Esse achado está de acordo com a literatura, apontando que o genótipo 1A é o genótipo mais prevalente de HCV no Brasil (BRUGGMANN P et al., 2014). Nós também detectamos sequências, embora em menor quantidade, que foram classificadas como HBV. O único *contig* de HBV de 244 pb foi classificado como pertencente ao genótipo A, que é um dos genótipos mais comuns de HBV no Brasil (LAMPE E et al, 2017). Esses resultados demonstram que a presença de HCV e HBV em pacientes multiplamente transfundidos está de acordo com a distribuição genotípica desses dois vírus no Brasil e que o uso de SNG e da nossa pipeline foi capaz de determinar a presença ativa de viremia de infecções virais crônicas tanto de vírus de DNA como de RNA.

Nesse estudo não foi confirmada a identificação de sequências de nenhum novo vírus que pudesse ser atribuído como ameaça transfusional. Em geral, a identificação de vírus emergentes é dependente dos alinhamentos com o banco de dados de vírus conhecidos, portanto, vírus que são distantemente relacionados com outros agentes virais já identificados podem não ter sido detectados pela pipeline utilizada nesse trabalho. Contudo, nós identificamos algumas sequências com baixa similaridade nucleotídica (20 a 30%), pertencentes às famílias *Marseilleviridae*, *Phycodnaviridae* e *Mimiviridae* que representam vírus gigantes que poderiam estar envolvidos em casos de transmissão sanguínea, mas que são geralmente apontados como contaminações, uma vez que estão presentes virtualmente na maioria dos ambientes da terra. (HALARY S et al., 2016). O mesmo problema se aplica aos *contigs* relativamente grandes que apresentam baixa similaridade proteica. Um exemplo disso é a similaridade de 39% encontrada para o vírus Iguapé, conhecido por já ter circulado no estado de São Paulo (COIMBRA TL et al., 1993). Devido ao número limitado de *contigs* (apenas um), não é possível investigar mais características desse agente viral como genoma completo.

Nos dados obtidos do grupo de pacientes com anemia falciforme, nós podemos observar um perfil do viroma semelhante ao grupo de talassêmicos, com presença majoritária de vírus comensais da família *Anneloviridae* e *Flaviviridae*. Nossos resultados estão de acordo com um estudo metagenômico realizado em pacientes franceses com anemia falciforme onde foram identificados exclusivamente Anelovírus e HPgV-1. (SAUVAGE V et al., 2016). Curiosamente, ao contrário da proporção viral dos pacientes com beta-talassemia, o vírus HPgV-1 demonstrou uma contagem de sequências superior aos vírus da família *Anneloviridae* em praticamente todos os pools, exceto pool 3 e 5. Considerando o vírus comensal HPgV-1, observamos uma prevalência um pouco mais elevada em pacientes de anemia falciforme. utilizando a confirmação via PCR a prevalência do HPgV-1 na anemia falciforme foi de 15.6%, enquanto no grupo de beta-talassemia foi de 14.3%. Como pacientes com beta-talassemia recebem mais transfusões a prevalência mais alta em pacientes com anemia falciforme pode significar que este vírus apresenta outras vias de transmissão além da parenteral. Apesar de ser um vírus comensal, nós não sabemos o efeito clínico desse agente viral em pacientes com anemia falciforme. Curiosamente, diversos estudos reportaram um aumento da sobrevida e uma melhora da qualidade de vida de pacientes com HIV coinfectados com HPgV-1 (de MIRANDA BKB et al., 2017; SCHWARZE-ZANDER C et al., 2006; CANDUCCI F, et al., 2003). Nesse sentido, nós fizemos triagem para HPgV-1 em doadores de sangue onde nós observamos uma prevalência de 15.6%. Isso confirma a prevalência mais elevada em pacientes com alto risco e que a via parenteral é um eficiente meio de transmissão viral.

A presença de contaminações revela um problema comumente encontrado nos estudos metagenômicos. Existem diversos casos de vírus que inicialmente foram associados a doenças em estudos de metagenômica, (LOMBARDI VC et al., 2009, XU B et al., 2013) mas que posteriormente foram reclassificados como contaminantes (SMUTS H et al., 2014, NACCACHE SN et al, 2013). A observação da cobertura genômica e a distribuição dos *reads* sobre o genoma do vírus identificado, são importantes critérios a serem considerados na análise dos dados. Uma alta cobertura genômica e/ou sequências distribuídas em regiões diferentes do genoma de referência são indicativos de que as sequências virais em questão são provenientes de um vírus, ao invés de um artefato (ASPLUND M, 2019). O número de leituras identificadas também pode ser um primeiro indicador de confiabilidade, devido ao fato de sequências contaminantes serem inversamente correlacionadas com o número total de leituras detectadas (ZINTER MS et al., 2019), ou seja, quanto menor o número de leituras detectadas maior o risco do vírus em questão ser um falso positivo.

Um desafio enfrentado na análise dos dados provenientes do sequenciamento do grupo de amostras de pacientes com anemia falciforme se deu devido a presença de contaminações relacionadas a vetores de clonagem utilizados na nossa prática laboratorial e, possivelmente, de reagentes. Um pequeno número de sequências (entre 2 à 30 sequências) foram classificadas como HIV, Indiana Vesiculovírus (VSV) e o Vírus da Leucemia Murina (MLV). Essas sequências formaram contigs de aproximadamente 200 pb e foram classificadas em regiões genômicas específicas, geralmente associadas a genes virais utilizados em vetores de clonagem. As análises mais aprofundadas com o BLAST revelaram que as sequências de HIV correspondiam ao *REV Response Element* com referência à um vetor de clonagem depositado no Genbank (MK801288). As sequências de CMV corresponderam a regiões regulatórias classificados como promotoras (CMV-MIE-enhancer) que são comuns em vetores de clonagem. De forma similar, as sequências de VSV foram associadas à glicoproteína (VSV-G) do genoma do VSV, que é utilizada como parte de vetores retrovirais (GIROGLOU T et al., 2004). Essas sequências possuíam baixa cobertura e se concentravam em apenas uma região gênica, o que é considerado uma forte indicação de presença de contaminantes. Sato *et al* identificaram sequências relacionadas ao MLV em kits comerciais de RT-PCR que correspondiam as regiões retrovirais *env* e *gag* também identificadas em nosso estudo. Esses resultados, aliados a abundância viral de MLV relativamente baixa em nossas amostras, sugerem fortemente que essas sequências são provenientes de contaminações por reagentes (i.e., SuperScript III, Invitrogen).

Por outro lado, as contaminações de Indiana Vesiculovírus (VSV) e Citomegalovírus provavelmente ocorreram durante as extrações das amostras, que foram feitas no mesmo ambiente onde manipulações com vetores de clonagem também ocorreram, além disso, esses vírus não são comumente identificados em pacientes com anemia falciforme que recebem múltiplas transfusões. Devido à baixa quantidade de sequências (2-30), a presença desses *reads* em todos os pools, a análise através do BLAST e um alinhamento guiado foi suficiente para verificar que se tratava de contaminações uma vez que os principais *hits* foram atribuídos a vetores de clonagem presentes no banco de dados do NCBI. Ao mesmo tempo, o material genético desses vírus foi utilizado como componentes de vetores de clonagem utilizados em nossa instituição, como o pMD2.G (VSV-G, CMV). Todos os pacientes foram soronegativos para HIV, o que corrobora com o fato das sequências de HIV identificadas se originarem de contaminações. Adicionalmente, a detecção desse tipo de contaminação demonstra uma

sensibilidade elevada de nossa pipeline bem como dos procedimentos laboratoriais na detecção de sequências pouco abundantes. Esses resultados demonstram que mesmo sendo uma excelente ferramenta para descrição de vírus emergentes o SNG apresenta algumas limitações diagnósticas como a frequente contaminação que pode interferir na interpretação dos dados e portanto é imprescindível para comparar o histórico clínico do paciente com os dados obtidos do SNG.

Para se facilitar a identificação de contaminantes na análise bioinformática, a utilização de controles negativos durante o sequenciamento é uma prática recomendada, mas que, devido aos custos do sequenciamento, ainda é uma prática adotada em poucos laboratórios, inclusive no nosso.

Em ambos os grupos de pacientes, a análise filogenética dos genomas completos do HPgV-1 revelou a presença do genótipo 2 (subgenótipo 2A), um genótipo altamente prevalente no Brasil (ALCALDE R et al., 2010; RAMOS FILHO R et al., 2004; NISHIYA AS et al., 2003) e o principal genótipo transmitido via transfusão de sangue (Figura 10 e 12).

A análise filogenética do maior contig do HCV (3.786 pb) (Figura 11) revelou que ele pertence ao genótipo 1A. Esse achado está de acordo com a literatura, apontando que o genótipo 1A é genótipo mais prevalente de HCV no Brasil e na América Latina (BRUGGMANN P et al., 2014). Foram também detectadas sequências, embora em menor quantidade, que foram classificadas como HBV. O único contig de HBV de 244 pb foi classificado como pertencente ao genótipo A (Figura 12), que é um dos mais comuns genótipos de HBV no Brasil (LAMPE E et al., 2017). Esses resultados demonstram que a presença de HCV e HBV em pacientes multiplamente transfundidos está de acordo com a distribuição genotípica desses dois vírus no Brasil e que o uso de SNG e da nossa pipeline foi capaz de determinar a viremia estabelecida de infecções virais crônicas tanto de vírus de DNA como de RNA.

Atualmente existem poucos estudos sobre a prevalência de HPgV-1 em pacientes com anemia falciforme que recebem múltiplas transfusões. Como os nossos resultados de metagenômica revelaram um alto número de sequências de HPgV-1 nos grupos com hemoglobinopatias, especialmente no grupo de anemia falciforme, assumimos que seria importante avaliar a prevalência desse agente viral neste grupo de pacientes. A prevalência de HPgV-1 encontrada foi de 15,6%. Apesar da escassez de informação sobre a prevalência de HPgV-1 no grupo de anemia falciforme, a prevalência que nós determinamos pode ser considerada relativamente alta. O único estudo que examina a prevalência de HPgV-1 RNA entre pacientes com anemia falciforme foi feito no Gabão e mostrou uma prevalência de 9,7%

entre os 31 pacientes testados (TUVERI R, et al., 2000). Diferentes fatores podem explicar essa menor prevalência dos indivíduos de Gabão quando comparada com o nosso estudo. Uma possível explicação poderia ser a região geográfica de origem dos pacientes. Regiões geográficas diferentes possuem variedade na circulação de genótipos e níveis diferentes de disseminação e infectividade das cepas. Contudo, uma outra explicação poderia ser a diferença de idade dos grupos estudados. A média de idades dos pacientes no estudo de Gabão foi de 4.9 anos (todos os pacientes são crianças) enquanto a do no nosso grupo foi de 26.2 anos. À medida que a idade aumenta, a possibilidade de exposição a um vírus como o HPgV-1 também deve aumentar consideravelmente, o que tende a resultar em uma maior prevalência em grupos mais velhos. Mais estudos que avaliam a prevalência de HPgV-1 em diferentes países são necessários para entender melhor a circulação e disseminação desse vírus em pacientes de alto risco de aquisição de doenças parenterais como o grupo de anemia falciforme.

De forma semelhante ao estudo com pacientes com anemia falciforme, existem poucos estudos sobre a prevalência do HPgV-1 em pacientes com beta-talassemia ao redor do mundo. No nosso estudo na região de Ribeirão Preto, foi encontrado uma prevalência do RNA de HPgV-1 significativamente maior que a prevalência em pacientes indianos com talassemia na região de Mumbai (3.37%) (DESAI MM et al., 2004). Ao mesmo tempo, essa prevalência obtida em nossa região foi similar a alguns resultados obtidos em pacientes na Itália, 12.5% (GAMBERINI MR et al., 2004) e 14.5% (PRATI D et al., 1998), e foi igual a um estudo realizado em 2003 na região do estado de São Paulo (14.2%) (WATANABE MAE et al., 2003). Contudo, alguns outros estudos na Itália (35%) (CARNELI V et al., 1998) e Índia (39.7%) (PANIGRAHI AK et al., 1998) determinaram uma prevalência de RNA de HPgV-1 superior aos resultados previamente citados, bem como na Tailândia (32.6%) (POOVORAWAN Y et al., 1998) e Taiwan (51.4%)(YU ML et al., 2000). A concordância entre a prevalência determinada em nosso estudo e o resultado de Watanabe *et al* em 2003, ambas determinadas na região de São Paulo, sugere que o padrão de circulação e as taxas de infecção do HPgV-1 entre pacientes com beta-talassemia são relativamente estáveis, independentemente do ano testado ou do método de detecção utilizado. No entanto, esse estudo de 2003 não especificou características filogenéticas do HPgV-1, como o genótipo predominantemente circulante em São Paulo. As diferenças de prevalência entre os estudos citados podem ser explicados pela região geográfica estudada, as regiões genômicas utilizadas nos testes e a sensibilidade dos *primers* construídos para detectar genótipos específicos do HPgV-1. Por outro lado, a prevalência relativamente alta do RNA do HPgV-1 nesse grupo de pacientes pode ser

explicado principalmente pelo alto número de transfusões ao qual esses pacientes são submetidos e a rota de transmissão parenteral do HPgV-1.

De forma similar aos grupos de hemoglobinopatias estudados, nossa análise em amostras de pacientes com hemofilia identificou uma grande quantidade de vírus comensais representantes da família *Anelloviridae*, como o Torque Teno vírus. Além disso, também foi possível a identificação do vírus da Hepatite C, onde subsequentes testes de ácido nucléico (NAT) revelaram a prevalência de HCV em 60% das amostras, o que corresponde a 3 dos 5 pacientes que constituíram nosso pool de amostras. Dados do histórico de exames sorológicos de cada paciente confirmaram a infecção por HCV através da presença de anticorpos anti-HCV. A alta prevalência de HCV nos pacientes com hemofilia tratados com fatores plasmáticos, contudo, não é inesperada, considerando que a maioria desses pacientes iniciaram o tratamento em um período em que a inativação viral de hemoderivados ainda não era amplamente realizada (BAPTISTA GONZÁLEZ HÁ et al., 2002, BOLTON-MAGGS PH et al., 2003).

6. CONCLUSÃO

O presente estudo buscou caracterizar o perfil viral de pacientes com beta-talassemia, anemia falciforme e hemofilia. Nós detectamos vírus comensais conhecidos e parenteralmente transmitidos pertencentes a diferentes famílias. Isso demonstra que apesar da significativa abundância viral encontrada nesses pacientes, a detecção de vírus emergentes (ou até mesmo persistentes e com viremia crônica) é uma tarefa difícil, mesmo considerando o tratamento com múltiplas transfusões. Além disso, o Estado de São Paulo é considerado uma região onde ocorre a circulação de arbovírus emergentes com grande relevância como Zika e Dengue que também demonstram transmissão transfusional. Esses tipos de agentes virais exóticos podem apresentar quadros clínicos diferentes em pacientes com hemoglobinopatias e hemofilia como foi observado durante a epidemia de Dengue 2 onde alguns pacientes com anemia falciforme diagnosticados com esse agente viral demonstraram quadro clínico grave com crise vaso-oclusiva exacerbada e risco de choque. Portanto, a vigilância do viroma em pacientes com múltiplas transfusões e hemofilia torna-se de suma importância, não somente para identificação de agentes emergentes ou não suspeitos, mas também para o direcionamento das políticas públicas para o monitoramento dos mesmos e melhoria da vida desses grupos de pacientes largamente negligenciados. Apesar de não termos identificado agentes virais emergentes até o momento, nosso estudo demonstra que a abordagem de metagenômica representa uma estratégia interessante na avaliação da diversidade viral em pacientes submetidos a múltiplas transfusões e, portanto, reforça sua relevância para o monitoramento de possíveis agentes virais que possam ameaçar a segurança transfusional.

REFERÊNCIAS

1. ABDEL-GHAFFAR TY, Sira MM, El Naghi S. Hepatitis C genotype 4: The past, present, and future. *World J Hepatol.* 2015;7(28):2792-2810. doi:10.4254/wjh.v7.i28.
2. ABRAHAM P, Hallett R, Radhakrishnan S, Raghuraman S, Sridharan G, Teo CG. Hepatitis G virus genotypes in a tertiary care centre in southern India. *J Clin Virol.* 2002 Aug;25(2):187-92.
3. Alcalde R, Nishiya A, Casseb J, Inocêncio L, Fonseca LA, Duarte AJ. Prevalence and distribution of the GBV-C/HGV among HIV-1-infected patients under anti-retroviral therapy. *Virus Res.* 2010 Aug;151(2):148-52. doi: 10.1016/j.virusres.2010.04.008. Epub 2010 Apr 24. PMID: 20420864.
4. ALI S, van Pelt JF, Verslype C, Nevens F, Fevery J, Yap SH. TT virus infection in acute and chronic liver diseases and in patients regularly receiving blood products in Belgium. *Acta Gastroenterol Belg.* 2004;67(2):161-5.
5. AL-QAHTANI AA, Alabsi ES, AbuOdeh R, Thalib L, El Zowalaty ME, Nasrallah GK. Prevalence of anelloviruses (TTV, TTMDV, and TTMV) in healthy blood donors and in patients infected with HBV or HCV in Qatar. *Virol J.* 2016 Dec 28;13(1):208. doi: 10.1186/s12985-016-0664-6. Erratum in: *Virol J.* 2017 Jul 27;14 (1):143. PMID: 28031027; PMCID: PMC5198501.
6. ALTSCHUL SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–410.
7. ANGELETTI S, Lo Presti A, Giovanetti M, et al. Phylogenesis and homology modeling in Zika virus epidemic: food for thought. *Pathog Glob Health.* 2016;110:269-274.
8. ASPLUND M, Kjartansdóttir KR, Møllerup S, et al. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin Microbiol Infect.* 2019;25(10):1277-1285.
9. ASTER V, König J, Stanková M, Rozsypal H, Procházka B. Prevalence of GBV-C/HGV (HGV) in HIV-infected patients and potential influence of co-infection on the course of the disease. *Klin Mikrobiol Infekc Lek.* 2005 Dec;11(6):199-1203. Czech, English. PMID: 16382413.
10. BAL A, Sarkozy C, Josset L, Cheynet V, Oriol G, Becker J, Vilchez G, Sesques P, Mallet F, Pachot A, Morfin F, Lina B, Salles G, Reynier F, Trouillet-Assant S, Brengel-Pesce K. Metagenomic Next-Generation Sequencing Reveals Individual Composition and Dynamics of Anelloviruses during Autologous Stem Cell Transplant Recipient Management. *Viruses.* 2018 Nov 14;10(11):633. doi: 10.3390/v10110633. PMID: 30441786; PMCID: PMC6266913.
11. BANKEVICH A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012 May;19(5):455-77.
12. BAPTISTA GONZÁLEZ HA. Infección por hepatitis C en pacientes con hemofilia [Hepatitis C virus infection in patients with hemophilia]. *Rev Gastroenterol Mex.* 2002 Oct;67 Suppl 2:S75-9. Spanish. PMID: 12712860.
13. BAPTISTA GONZÁLEZ HA. Infección por hepatitis C en pacientes con hemofilia [Hepatitis C virus infection in patients with hemophilia]. *Rev Gastroenterol Mex.* 2002 Oct;67 Suppl 2:S75-9. Spanish. PMID: 12712860.

14. BARJAS-CASTRO ML, Angerami RN, Cunha MS, Suzuki A, Nogueira JS, Rocco IM, et al. Probable transfusion-transmitted Zika virus in Brazil. *Transfusion* 2016; 56:1684–8.
15. Bernardin F, Operskalski E, Busch M, Delwart E. Transfusion transmission of highly prevalent commensal human viruses. *Transfusion*. 2010 Nov;50(11):2474-83. doi: 10.1111/j.1537-2995.2010.02699.x. PMID: 20497515.
16. BESSER J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect*. 2018;24(4):335-341
17. BIAGINI P, de Micco P, de Lamballerie X. Identification of a third member of the Anellovirus genus ("small anellovirus") in French blood donors. *Arch Virol*. 2006 Feb;151(2):405-8. doi: 10.1007/s00705-005-0660-4. Epub 2005 Nov 21. PMID: 16328140.
18. BLATYTA PF, Kelly S, Sabino E, et al. Prevalence of serologic markers of transfusion and sexually transmitted infections and their correlation with clinical features in a large cohort of Brazilian patients with sickle cell disease. *Transfusion*. 2020;60(2):343–350.
19. BOLGER AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–2120.
20. BOLTON-MAGGS PH, Pasi KJ. Haemophilias A and B. *Lancet*. 2003 May 24;361(9371):1801-9. doi: 10.1016/S0140-6736(03)13405-8. PMID: 12781551.
21. BOLTON-MAGGS PH, Pasi KJ. Haemophilias A and B. *Lancet*. 2003 May 24;361(9371):1801-9. doi: 10.1016/S0140-6736(03)13405-8. PMID: 12781551.
22. BOUADJENEK MR, Verspoor K, Zobel J. Literature consistency of bioinformatics sequence databases is effective for assessing record quality. *Database (Oxford)*. 2017;2017(1):bax021. doi:10.1093/database/bax021
23. BRITO F, Cordey S, Delwart E, Deng X, Tirefort D, Lemoine-Chaduc C, Zdobnov E, Lecompte T, Kaiser L, Waldvogel-Abramowski S, Preynat-Seauve O. Metagenomics analysis of the virome of 300 concentrates from a Swiss platelet bank. *Vox Sang*. 2018 Jul 18;10.1111/vox.12695. doi: 10.1111/vox.12695. Epub ahead of print. PMID: 30022500; PMCID: PMC6338525.
24. BROCCHERI L. Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol*. 2001 Feb;59(1):27-40.
25. BRUGGMANN P, Berg T, Øvrehus AL, Moreno C, Brandão Mello CE, Roudot-Thoraval F, Marinho RT, Sherman M, Ryder SD, Sperl J, Akarca U, Balık I, Bihl. Historical epidemiology of hepatitis C virus (HCV) in selected countries. *J Viral Hepat*. 2014 May;21 Suppl 1:5-33
26. CAMCI C, Guney C, Balkan A, Buyukberber N, Buyukberber S, Kadayifci A, Kubar A. The prevalence of TT virus in cancer patients. *New Microbiol*. 2002 Oct;25(4):463-8.
27. CAO A, Galanello, R. Beta-thalassemia. *Genet Med* 12, 61–76 (2010).
28. CHATTOPADHYAY S, Rao S, Das BC, Singh NP, Kar P. Prevalence of transfusion-transmitted virus infection in patients on maintenance hemodialysis from New Delhi, India. *Hemodial Int*. 2005;9(4):362–366.

29. CHENG VCC, Sridhar S, Wong SC, et al. Japanese Encephalitis Virus Transmitted Via Blood Transfusion, Hong Kong, China. *Emerg Infect Dis.* 2018; 24(1).
30. CHONG Y, Lee JY, Thakur N, Kang CS, Lee EJ. Strong association of Torque teno virus/Torque teno-like minivirus to Kikuchi-Fujimoto lymphadenitis (histiocytic necrotizing lymphadenitis) on quantitative analysis. *Clin Rheumatol.* 2020;39(3):925-931.
31. COMPEAU PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol.* 2011;29(11):987-991. Published 2011 Nov 8. doi:10.1038/nbt.2023
32. DE MIRANDA BKB, de Sá KSG, da Silva ANR, Feitosa RNM, Cayres-Vallinoto IMV, Ishak R, Vallinoto ACR. GBV-C/HIV-1 coinfection is associated with low HIV-1 viral load and high CD4+ T lymphocyte count. *Arch Virol.* 2017 Nov;162(11):3431-3438.
33. DE SANCTIS V, Kattamis C, Canatan D, et al. β -Thalassemia Distribution in the Old World: an Ancient Disease Seen from a Historical Standpoint. *Mediterr J Hematol Infect Dis.* 2017;9(1):e2017018. Published 2017 Feb 20.
34. DE SOUZA PEREIRA BB, Darrigo Junior LG, de Mello Costa TC, Felix AC, Simoes BP, Stracieri AB, da Silva PM, Mauad M, Machado CM. Prolonged viremia in dengue virus infection in hematopoietic stem cell transplant recipients and patients with hematological malignancies. *Transpl Infect Dis.* 2017 Aug;19(4).
35. DELWART EL. Viral metagenomics. *Rev Med Virol.* 2007 Mar-Apr;17(2):115-31. doi: 10.1002/rmv.532. PMID: 17295196; PMCID: PMC7169062.
36. DOS SANTOS BEZERRA R, Santos EV, Maraninchi Silveira R, Silva Pinto AC, Covas DT, Kashima S, Slavov SN. Molecular prevalence and genotypes of human pegivirus-1 (HPgV-1) and SENV-like viruses among multiply transfused patients with beta-thalassemia. *Transfus Apher Sci.* 2019 Dec 13:102697.
37. ERNST D, Greer M, Akmatova R, Pischke S, Wedemeyer H, Heiken H, Tillmann HL, Schmidt RE, Stoll M. Impact of GB virus C viraemia on clinical outcome in HIV-1-infected patients: a 20-year follow-up study. *HIV Med.* 2014 Apr;15(4):245-50.
38. ESTCOURT LJ, Hopewell S, Trivella M, Hambleton IR, Cho G. Regular long-term red blood cell transfusions for managing chronic chest complications in sickle cell disease. *Cochrane Database Syst Rev.* 2019;2019(10)
39. FAHSBENDER E, Charlys da-Costa A, Elise Gill D, Augusto de Padua Milagres F, Brustulin R, Julio Costa Monteiro F, Octavio da Silva Rego M, Soares D'Athaide Ribeiro E, Cerdeira Sabino E, Delwart E. Plasma virome of 781 Brazilians with unexplained symptoms of arbovirus infection include a novel parvovirus and densovirus. *PLoS One.* 2020 Mar 5;15(3):e0229993
40. FARIA NR, Quick J, Claro IM, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature.* 2017;546(7658):406-410. doi: 10.1038/nature22401.
41. FLECKNOE D, Charles Wakefield B, Simmons A. Plagues & wars: the 'Spanish Flu' pandemic as a lesson from history. *Med Confl Surviv.* 2018 Jun;34(2):61-68. doi: 10.1080/13623699.2018.1472892. Epub 2018 May 15. PMID: 29764189.

42. FOULONGNE V, Sauvage V, Hebert C, Dereure O, Cheval J, Gouilh MA, Pariente K, Segondy M, Burguière A, Manuguerra JC, Caro V, Eloit M. Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS One*. 2012;7(6):e38499. doi: 10.1371/journal.pone.0038499. Epub 2012 Jun 19. PMID: 22723863; PMCID: PMC3378559.
43. FRANCO JM, Borges CCL, Ansaloni MA, Mauro RDL, Souza YCDC, Braga JAP. Chronic transfusion therapy effectiveness as primary stroke prophylaxis in sickle cell disease patients. *Hematol Transfus Cell Ther*. 2019;S2531-1379(19)30160-9.
44. FURUTA RA, Sakamoto H, Kuroishi A, Yasiui K, Matsukura H, Hirayama F. Metagenomic profiling of the viromes of plasma collected from blood donors with elevated serum alanine aminotransferase levels. *Transfusion* 2015;55:1889–99
45. GAO J, Couzens L, Burke DF, et al. Antigenic drift of the influenza A (H1N1) pdm09 Virus neuraminidase results in reduced effectiveness of A/California/7/2009 (H1N1pdm09)-specific. *mBio*. 2019;10:e00307-19.
46. GEOGHEGAN JL, Holmes EC. Predicting virus emergence amid evolutionary noise. *Open Biol*. 2017 Oct;7(10):170189. doi: 10.1098/rsob.170189. PMID: 29070612; PMCID: PMC5666085.
47. GERGELY P JR, Perl A, Poór G. Possible pathogenic nature of the recently discovered TT virus: does it play a role in autoimmune rheumatic diseases?. *Autoimmun Rev*. 2006;6(1):5-9.
48. GIOVANETTI M, Grifoni A, Lo Presti A, et al. Amino acid mutations in Ebola virus glycoprotein of the 2014 epidemic. *J Med Virol*. 2015;87(6):893-898.
49. GOODNOUGH LT, Panigrahi AK. Blood Transfusion Therapy. *Med Clin North Am*. 2017 Mar;101(2):431-447. doi: 10.1016/j.mcna.2016.09.012. Epub 2016 Dec 24.
50. GORBALENYA AE, Lauber C. Phylogeny of Viruses. Reference Module in Biomedical Sciences. 2017;B978-0-12-801238-3.95723-4. doi:10.1016/B978-0-12-801238-3.95723-4
51. HANDELSMAN J, Rondon MR, Brady SF et al (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 5:R245–R249
52. HANDELSMAN J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 2004 Dec;68(4):669-85. doi: 10.1128/MMBR.68.4.669-685.2004. PMID: 15590779; PMCID: PMC539003.
53. HOPKINS, Donald R. (1980). Ramses V: earliest known victim? *World Health*;1980 May p.22-26
54. HOWARD J. Sickle cell disease: when and how to transfuse. *Hematology Am Soc Hematol Educ Program*. 2016 Dec 2;2016(1):625-631.
55. IKUTA Y, Oba K, Nai E, et al. Aseptic meningitis caused by torque teno virus in an infant: a case report. *J Med Case Rep*. 2019;13(1):302. Published 2019 Sep 24.
56. ILUNGA KALENGA O, Moeti M, Sparrow A, Nguyen VK, Lucey D, Ghebreyesus TA. The Ongoing Ebola Epidemic in the Democratic Republic of Congo, 2018-2019. *N Engl J Med*. 2019;381(4):373-383. doi:10.1056/NEJMs1904253
57. JOGEDA EL, Huik K, Pauskar M, et al. Prevalence and genotypes of GBV-C and its associations with HIV infection among persons who inject drugs in Eastern Europe. *J Med Virol*. 2017;89(4):632-638.

58. KARIM F, Nasir N, Moiz B. Transfusion transmitted dengue: One donor infects two patients. *Transfus Apher Sci.* 2017;56(2):151–153.
59. KARIM F, Nasir N, Moiz B. Transfusion transmitted dengue: One donor infects two patients. *Transfus Apher Sci.* 2017;56(2):151–153.
60. KATOH K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002;30(14):3059–3066
61. KCHOUK M, Gibrat JF, Elloumi M (2017) Generations of Sequencing Technologies: From First to Next Generation. *Biol Med (Aligarh)* 9:395.
62. KUMAR KR, Cowley MJ, Davis RL. Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost.* 2019 Oct;45(7):661-673. doi: 10.1055/s-0039-1688446. Epub 2019 May 16. PMID: 31096307.
63. LAMPE E, Mello FCA, do Espírito-Santo MP, Oliveira CMC, Bertolini DA, Gonçalves NSL, Moreira RC, Fernandes CAS, Nascimento HCL, Grotto RMT, Pardini MIMC; The Brazilian Hepatitis B Research Group. Nationwide overview of the distribution of hepatitis B virus genotypes in Brazil: a
64. LEFRÈRE JJ, Roudot-Thoraval F, Lefrère F, et al. Natural history of the TT virus infection through follow-up of TTV DNA-positive multiple-transfused patients. *Blood.* 2000;95(1):347-51.
65. LIU HF, Goderniaux E, Burtonboy G, Goubau P. Molecular analysis of GB virus C/hepatitis G virus in HIV-1-positive intravenous drug users in Belgium. *J Hum Virol.* 1999;2(2):115-20.
66. LYSHOLM F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, Lindberg AM, Persson B, Allander T, Andersson B. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS One.* 2012;7(2):e30875. doi: 10.1371/journal.pone.0030875. Epub 2012 Feb 15. PMID: 22355331; PMCID: PMC3280267.
67. MA YY, Guo Y, Zhao X, Wang Z, Lv MM, Yan QP, Zhang JG. Human parvovirus PARV4 in plasma pools of Chinese origin. *Vox Sang.* 2012 Oct;103(3):183-5. doi: 10.1111/j.1423-0410.2012.01605.x. Epub 2012 Mar 27. PMID: 22458565.
68. MAKHLOUF MM, Elwakil SG, Ibrahim NS. Molecular and serological assessment of parvovirus B-19 infection in Egyptian children with sickle cell disease. *J Microbiol Immunol Infect.* 2017 Oct;50(5):565-569. doi: 10.1016/j.jmii.2015.10.016. Epub 2015 Dec 1. PMID: 26698686.
69. MANSOUR AK, Aly RM, Abdelrazek SY, Elghannam DM, Abdelaziz SM, Shahine DA, Elmenshawy NM, Darwish AM. Prevalence of HBV and HCV infection among multi-transfused Egyptian thalassemic patients. *Hematol Oncol Stem Cell Ther.* 2012;5(1):54-9. doi: 10.5144/1658-3876.2012.54. PMID: 22446611.
70. MAROUF R. Blood transfusion in sickle cell disease. *Hemoglobin.* 2011;35(5-6):495-502.
71. MARZ M, Beerenwinkel N, Drosten C, Fricke M, Frishman D, Hofacker IL, Hoffmann D, Middendorf M, Rattei T, Stadler PF, Töpfer A. Challenges in RNA virus bioinformatics. *Bioinformatics.* 2014 Jul 1;30(13):1793-9.
72. MIKHEYEV AS, Tin MM. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour.* 2014 Nov;14(6):1097-102. doi: 10.1111/1755-0998.12324. Epub 2014 Sep 24. PMID: 25187008.

73. MINOT S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* 2011 Oct;21(10):1616-25. doi: 10.1101/gr.122705.111. Epub 2011 Aug 31. PMID: 21880779; PMCID: PMC3202279.
74. MOKILI JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol.* 2012 Feb;2(1):63-77.
75. MOTTA IJ, Spencer BR, Cordeiro da Silva SG, Arruda MB, Dobbin JA, Gonzaga YB, Arcuri IP, Tavares RC, Atta EH, Fernandes RF, Costa DA, Ribeiro LJ, Limonte F, Higa LM, Voloch CM, Brindeiro RM, Tanuri A, Ferreira OC Jr. Evidence for Transmission of Zika Virus by Platelet Transfusion. *N Engl J Med.* 2016 Sep 15;375(11):1101-3. doi: 10.1056/NEJMc1607262. Epub 2016 Aug 17. PMID: 27532622.
76. NAIK P, Dave VP, Joseph J. Detection of Torque Teno Virus (TTV) and TTV-Like Minivirus in patients with presumed infectious endophthalmitis in India. *PLoS One.* 2020;15(1):e0227121.
77. NGUYEN LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;32(1):268–274.
78. NISHIGUCHI S, Enomoto M, Shiomi S, et al. GB virus C and TT virus infections in Japanese patients with autoimmune hepatitis. *J Med Virol.* 2002;66(2):258-262. 9.
79. 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.* 2003 Jul-Aug;45(4):213-6. doi: 10.1590/s0036-46652003000400007. Epub 2003 Sep 17. PMID: 14502349.
80. PAN S, Yu T, Wang Y, Lu R, Wang H, Xie Y, Feng X. Identification of a Torque Teno Mini Virus (TTMV) in Hodgkin's Lymphoma Patients. *Front Microbiol.* 2018 Jul 26;9:1680. doi: 10.3389/fmicb.2018.01680. PMID: 30093892; PMCID: PMC6070622.
81. PÁR A, Takács M, Brojnás J, et al. Co-infections with hepatitis G and TT virus in patients with chronic hepatitis C in Hungary. *Acta Microbiol Immunol Hung.* 2004;51(4):437-447. doi:10.1556/AMicr.51.2004.4.4
82. PATTERSON KB, Runge T. Smallpox and the Native American. *Am J Med Sci.* 2002 Apr;323(4):216-22. doi: 10.1097/00000441-200204000-00009. PMID: 12003378.
83. PETERSEN LM, Martin IW, Moschetti WE, Kershaw CM, Tsongalis GJ. Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. *J Clin Microbiol.* 2019 Dec 23;58(1):e01315-19. doi: 10.1128/JCM.01315-19. PMID: 31619531; PMCID: PMC6935936.
84. PHAN TG, Kapusinszky B, Wang C, Rose RK, Lipton HL, Delwart EL. The fecal viral flora of wild rodents. *PLoS Pathog.* 2011 Sep;7(9):e1002218. doi: 10.1371/journal.ppat.1002218. Epub 2011 Sep 1. PMID: 21909269; PMCID: PMC3164639.
85. PISANI G, Cristiano K, Pupella S, Liumbruno GM. West Nile Virus in Europe and Safety of Blood Transfusion. *Transfus Med Hemother.* 2016 May;43(3):158-67. doi: 10.1159/000446219. Epub 2016 May 10. PMID: 27403087; PMCID: PMC4924479.

86. POLLOCK J., Glendinning L., Wisedchanwet T., Watson M. The Madness of Microbiome: Attempting To Find Consensus “Best Practice” for 16S Microbiome Studies. *Appl. Environ. Microbiol.* 2018;84 doi: 10.1128/AEM.02627-17.
87. PRATI D, Lin YH, De Mattei C, et al. A prospective study on TT virus infection in transfusion-dependent patients with beta-thalassemia. *Blood.* 1999;93(5):1502-1505.
88. PRATI D, Zanella A, Bosoni P, et al. The incidence and natural course of transfusion-associated GB virus C/hepatitis G virus infection in a cohort of thalassemic patients. The CooleyCare Cooperative Group. *Blood.* 1998;91(3):774-777
89. Ramos Filho R, Carneiro MA, Teles SA, Dias MA, Cardoso DD, Lampe E, Yoshida CF, Martins RM. GB virus C/hepatitis G virus infection in dialysis patients and kidney transplant recipients in Central Brazil. *Mem Inst Oswaldo Cruz.* 2004 Oct;99(6):639-43. doi: 10.1590/s0074-02762004000600019. Epub 2004 Nov 18. PMID: 15558178.
90. RESHETNYAK VI, Maev IV, Burmistrov AI, Chekmazov IA, Karlovich TI. Torque teno virus in liver diseases: On the way towards unity of view. *World J Gastroenterol.* 2020;26(15):1691-1707.
91. REYES A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature.* 2010 Jul 15;466(7304):334-8.
92. RHOADS A, Au KF. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics.* 2015 Oct;13(5):278-89. doi: 10.1016/j.gpb.2015.08.002. Epub 2015 Nov 2. PMID: 26542840; PMCID: PMC4678779.]
93. RIVEIRO-BARCIELA M, Saulea S, Quer J, Salvador F, Gregori J, Pirón M, Rodríguez-Frías F, Buti M. Red blood cell transfusion-transmitted acute hepatitis E in an immunocompetent subject in Europe: a case report. *Transfusion.* 2017 Feb;57(2):244-247. doi: 10.1111/trf.13876. Epub 2016 Oct 26. PMID: 27785789.
94. ROSE R, Constantinides B, Tapinos A, Robertson DL, Prosperi M. Challenges in the analysis of viral metagenomes. *Virus Evol.* 2016 Aug 3;2(2):vew022. doi: 10.1093/ve/vew022. PMID: 29492275; PMCID: PMC5822887.
95. SABINO EC, Loureiro P, Lopes ME, Capuani L, McClure C, Chowdhury D, Di-Lorenzo-Oliveira C, Oliveira LC, Linnen JM, Lee TH, Gonçalves T, Brambilla D, Kleinman S, Busch MP, Custer B; International Component of the NHLBI Recipient Epidemiology and Donor Evaluation Study-III. Transfusion-Transmitted Dengue and Associated Clinical Symptoms During the 2012 Epidemic in Brazil. *J Infect Dis.* 2016 Mar 1;213(5):694-702. doi: 10.1093/infdis/jiv326. Epub 2015 Jun 8. PMID: 26908780; PMCID: PMC4747611.
96. SALEMI M, Fitch WM, Ciccozzi M, Ruiz-Alvarez MJ, Rezza G, Lewis MJ. Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. *J Virol.* 2004; 78:1602-1603.
97. SAUVAGE V, Gomez J, Boizeau L, Laperche S. The potential of viral metagenomics in blood transfusion safety. *Transfus Clin Biol.* 2017;24(3):218-222. doi: 10.1016/j.traccli.2017.06.018
98. SAUVAGE V, Laperche S, Cheval J, et al. Viral metagenomics applied to blood donors and recipients at high risk for blood-borne infections. *Blood Transfus.* 2016;14(5):400–407.

99. SAUVAGE V, Laperche S, Cheval J, Muth E, Dubois M, Boizeau L, Hébert C, Lionnet F, Lefrère JJ, Eloit M. Viral metagenomics applied to blood donors and recipients at high risk for blood-borne infections. *Blood Transfus.* 2016 Sep;14(5):400-7.
100. SAVAS MC, Guney C, Kadayifci A, Balkan A, Koruk M, Kubar A, Uygun A. High prevalence of transfusion-transmitted virus infection in patients with chronic liver diseases in an endemic area of hepatitis B and C virus. *Med Princ Pract.* 2003 Jul-Sep;12(3):176-9.
101. SCHMIDT HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics.* 2002;18(3):502–504.
102. SCHMIEDER R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One.* 2011;6(3):e17288.
103. SCHNEIDER H. 2003. Métodos de Análise filogenética: um guia prático. Ribeirão Preto: Holos Editora, 162p
104. Seifried C, Weber M, Bialleck H, Seifried E, Schrezenmeier H, Roth WK. High prevalence of GBV-C/HGV among relatives of GBV-C/HGV-positive blood donors in blood recipients and in patients with aplastic anemia. *Transfusion.* 2004 Feb;44(2):268-74. doi: 10.1111/j.1537-2995.2004.00665.x. PMID: 14962319.
105. SHAN T, Li L, Simmonds P, Wang C, Moeser A, Delwart E. The fecal virome of pigs on a high-density farm. *J Virol.* 2011 Nov;85(22):11697-708. doi: 10.1128/JVI.05217-11. Epub 2011 Sep 7. PMID: 21900163; PMCID: PMC3209269.
106. SIQUEIRA JD, Dominguez-Bello MG, Contreras M, Lander O, Caballero-Arias H, Xutao D, Noya-Alarcon O, Delwart E. Complex virome in feces from Amerindian children in isolated Amazonian villages. *Nat Commun.* 2018 Oct 15;9(1):4270. doi: 10.1038/s41467-018-06502-9. Erratum in: *Nat Commun.* 2018 Nov 14;9(1):4852. PMID: 30323210; PMCID: PMC6189175.
107. SLAVOV SN, Haddad SK, Silva-Pinto AC, et al. Molecular and phylogenetic analyses of human Parvovirus B19 isolated from Brazilian patients with sickle cell disease and β -thalassemia major and healthy blood donors. *J Med Virol.* 2012;84(10):1652-1665.
108. SLAVOV SN, Hespanhol MR, Ferreira AR, Rodrigues ES, Covas DT, Kashima S. Silent dengue virus circulation among asymptomatic blood donors from a hyperendemic Brazilian region. *Transfus Med.* 2018;28(6):465-467.
109. SLAVOV SN, Kashima S, Silva-Pinto AC, Amarilla AA, Aquino VH, Covas DT. Molecular and clinical evaluation of the acute human parvovirus B19 infection: comparison of two cases in children with sickle cell disease and discussion of the literature. *Braz J Infect Dis.* 2013;17(1):97-101.
110. 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;9:e0003631.
111. TICEHURST JR, Pisanic N, Forman MS, Ordak C, Heaney CD, Ong E, Linnen JM, Ness PM, Guo N, Shan H, Nelson KE. Probable transmission of hepatitis E virus (HEV) via transfusion in the United States. *Transfusion.* 2019 Mar;59(3):1024-1034. doi: 10.1111/trf.15140. Epub 2019 Jan 31. PMID: 30702157; PMCID: PMC6501795.

112. VENTER JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001 Feb 16;291(5507):1304-51. Erratum in: *Science* 2001 Jun 5;292(5523):1838. PMID: 11181995.
113. VIRGIN HW. The virome in mammalian physiology and disease. *Cell*. 2014 Mar 27;157(1):142-50.
114. WAGNER, D. M., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., Sahl, J. W., Enk, J., Birdsell, D. N., Kuch, M., Lumibao, C., Poinar, D., Pearson, T., Fourment, M., Golding, B., Riehm, J. M., Earn, D. J. D., DeWitte, S., Rouillard, J. M., Grupe, G., ... Poinar, H. (2014). *Yersinia pestis* and the Plague of Justinian 541-543 AD: A genomic analysis. *The Lancet Infectious Diseases*, 14(4), 319-326.
115. WEATHERALL DJ, Miller LH, Baruch DI, Marsh K, Doumbo OK, Casals-Pascual C, Roberts DJ. Malaria and the red cell. *Hematology Am Soc Hematol Educ Program*. 2002:35-57.
116. WOOD DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014;15:R46.
117. ZEHENDER G, Pariani E, Piralla A, et al. Reconstruction of the evolutionary dynamics of the A(H1N1)pdm09 influenza virus in Italy during the pandemic and post-pandemic phases. *PLOS One*. 2012;7:e47517.
118. ZHANG W, Li L, Deng X, Blümel J, Nübling CM, Hunfeld A, Baylis SA, Delwart E. Viral nucleic acids in human plasma pools. *Transfusion*. 2016 Sep;56(9):2248-55.
119. ZHANG YZ, Shi M, Holmes EC. Using Metagenomics to Characterize an Expanding Virosphere. *Cell*. 2018 Mar 8;172(6):1168-1172