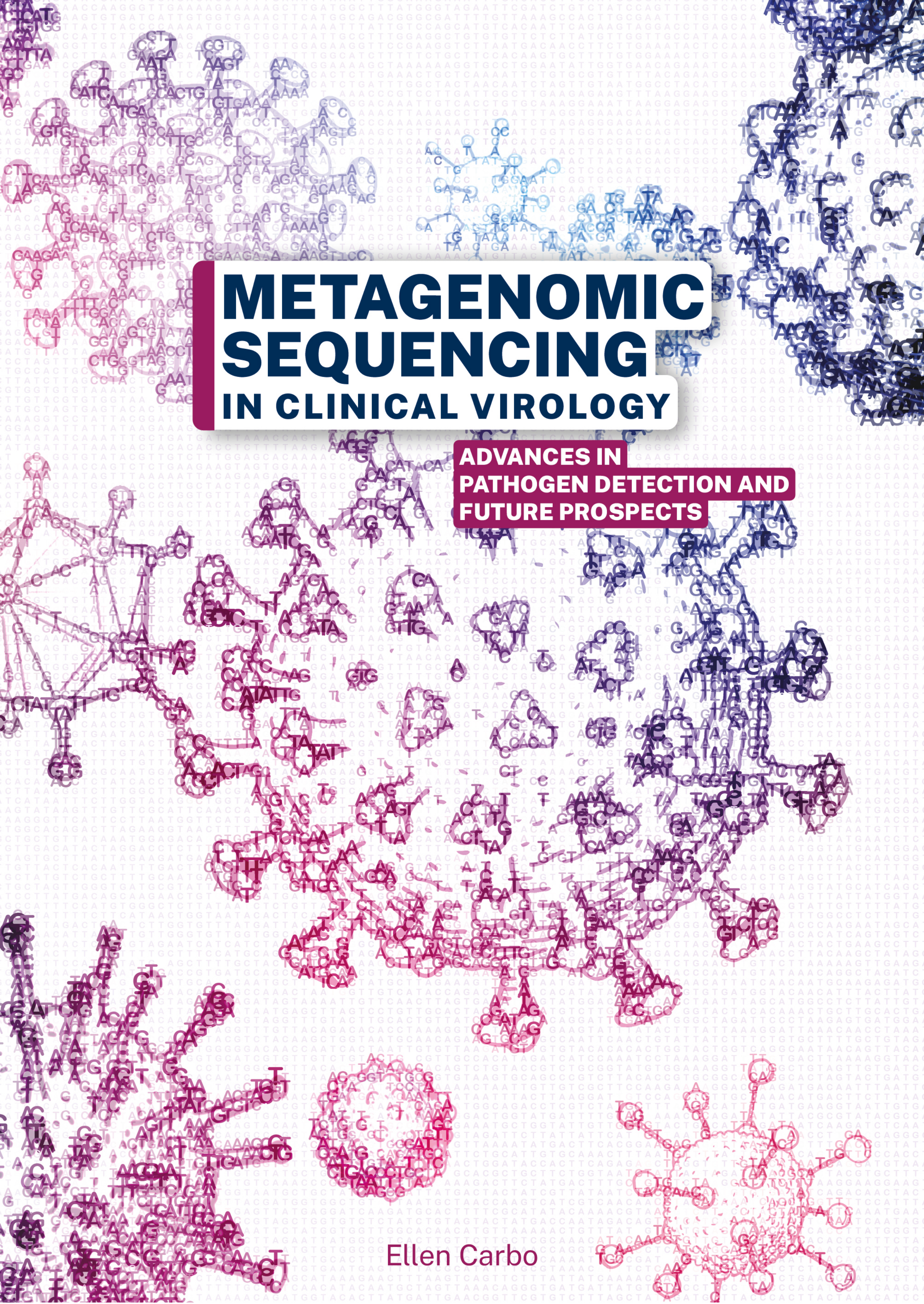
Metagenomic sequencing in clinical virology: advances in pathogen detection and future prospects

Carbo, E.C.



METAGENOMIC SEQUENCING IN CLINICAL VIROLOGY

ADVANCES IN PATHOGEN DETECTION AND FUTURE PROSPECTS

Ellen Carbo

Metagenomic sequencing in clinical virology: advances in pathogen detection and future prospects

PhD dissertation by:

Ellen Carbo

Colophon

ISBN ISBN 978-90-824257-4-1

Thesis **Metagenomic sequencing in clinical virology: advances in pathogen detection and future prospects**

Design&Layout Ralph Boer

Printing Ipskamp printing

Funding Printing of this thesis was financially supported by the Medical Microbiology department of the Leiden University Medical Center. Printing of this thesis was financially supported by the Netherlands Society of Medical Microbiology (NVMM) and the Royal Netherlands Society for Microbiology (KNVM).

This thesis is dedicated to:

Nino van Rooijen
1957–2015

Metagenomic sequencing in clinical virology:
advances in pathogen detection and future prospects

PROEFSCHRIFT

Ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op
woensdag 17 mei 2023
klokke 16:15 uur

door

Ellen Catharine Carbo
geboren te Utrecht
in 1982

Promotor

Prof. dr. A.C.M. Kroes

Co-promotoren

Dr. J.J.C. de Vries
Dr. I.A. Sidorov

Leden Promotiecommissie

Prof. dr. N. Fischer (University Medical Center Hamburg-Eppendorf)
Prof. dr. M.D. de Jong (Amsterdam University Medical Center, Amsterdam)
Prof. dr. L.G. Visser
Prof. dr. E.J. Snijder
Dr. E.C.J. Claas

Scientific research is one of the most

exciting and rewarding of occupations.

It is like a voyage of discovery

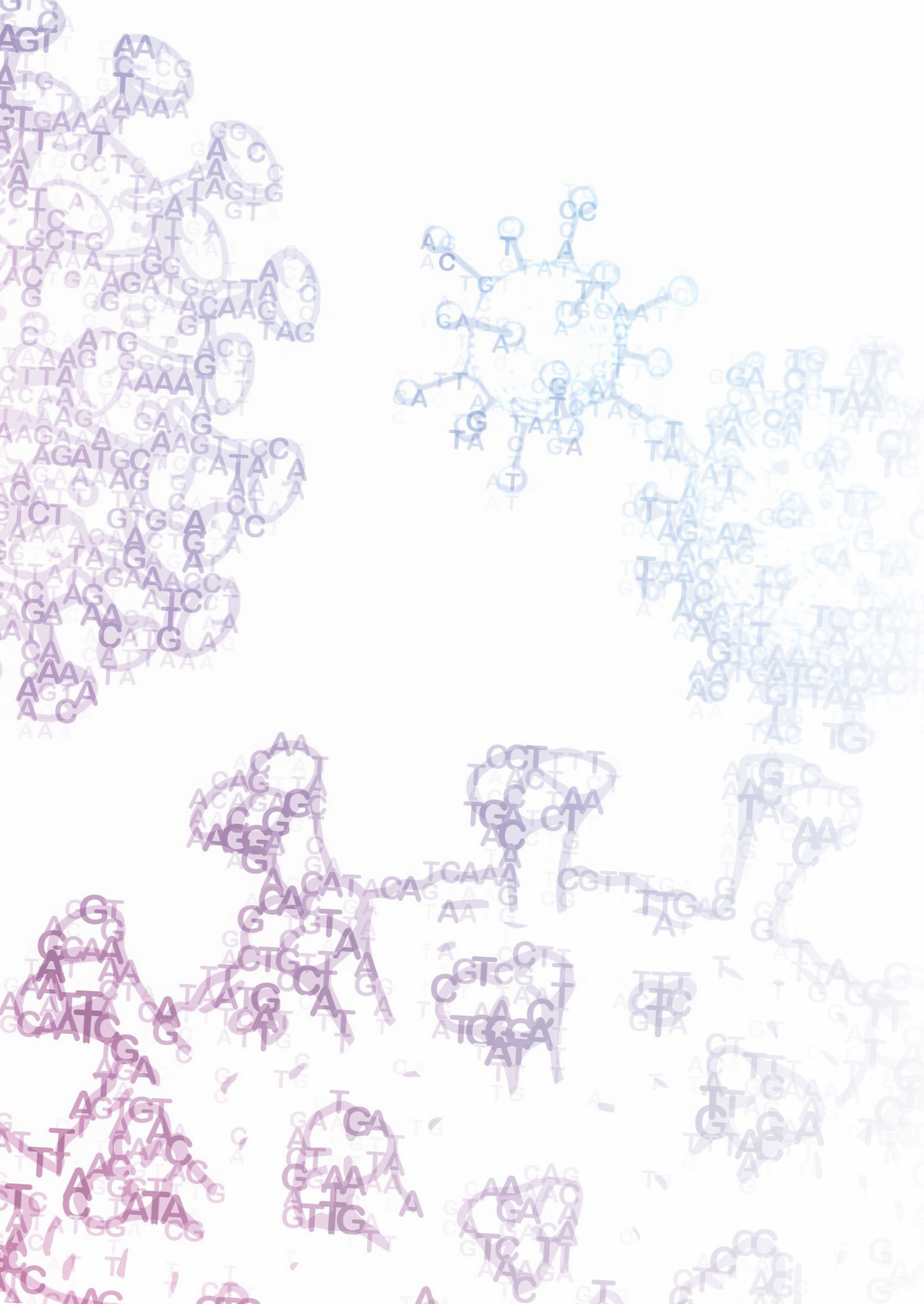
into unknown lands, seeking not for new territory

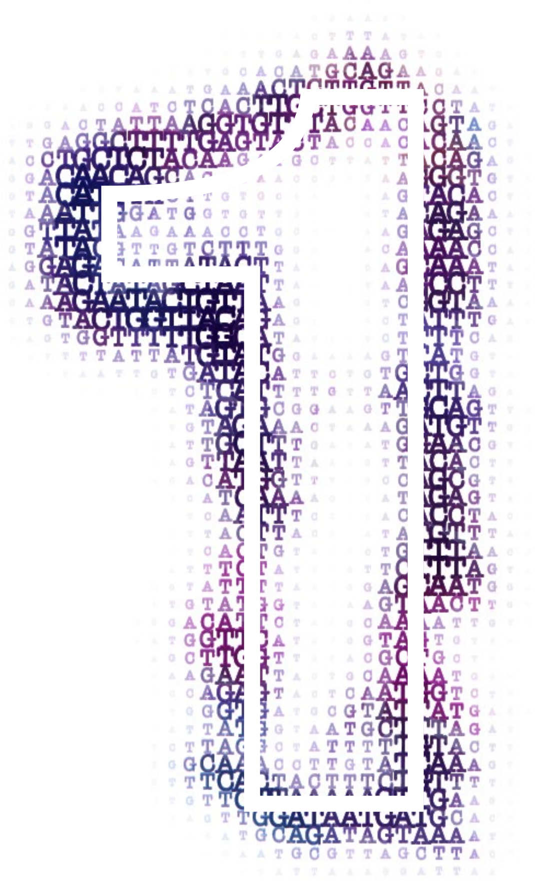
but for new knowledge. It should appeal to

those with a good sense of adventure.

Frederick Sanger, speech at the Nobel Banquet, 1980

Table of Contents





Chapter 1 **Introduction**

Molecular viral diagnostics is based on sequential tests identifying up to a handful of viruses at a time using polymerase chain reaction (PCR). However, with over 1,000 virus species known to infect humans [1], it is impossible to test all these viruses at once in a patient sample using this PCR method. And with a total presence of 10^{31} virus particles on earth [2] and novel or zoonotic viruses potentially infecting humans [3], there is a great expectation for a one-test-catches-all method. Viral metagenomic next-generation sequencing is such a method, though some early studies have shown a lower sensitivity compared to PCR [4,5]. The work performed in this thesis investigates the application and possibilities of viral metagenomic sequencing, further enhancing the accuracy of this test to make it suitable for application in a clinical setting.

This current chapter will first illustrate the relevance of infectious diseases as a societal burden and provide a summary of the history of infectious diseases and sequencing. Subsequently, several applications of sequencing will be explained with a focus on their utilization in viral metagenomic sequencing.

Background

Infections and pandemics

Infectious diseases have troubled humankind since recorded history. Tuberculosis (TB) is perhaps the longest known [6], and the causative pathogen may have already existed 15,000-20,000 years ago. Additionally, several types of infections have been described by both Hippocrates and the ancient Egyptians [7], and one of the first plagues described was the Athenian plague of 430 BC [8]. More recent pandemics of the last century with a large effect on society were the influenza pandemic between 1918-1920 resulting in approximate 50 million fatalities [9], and the outbreak of HIV/AIDs in the ‘80s and up until present day leading to over 38 million deaths [10]. Currently, infectious diseases and pandemics are still a threat, with SARS-CoV-2 as the causative agent of the latest large known outbreak.

All eukaryote cells harbour small pieces of bacterial DNA in their mitochondria, demonstrating that infections as encounters between different cells have a very long history [11]. The human genome is composed of 8% viruses due to integration of human endogenous retroviruses in our DNA [12]. Additionally, the human microbiome consists of about 10-100 trillion microbial cells that are permanent residents of the human gastrointestinal tract [13]. These symbiotic interactions with microorganisms are often ignored until the moment we get afflicted by an infection that will give us specific symptoms. Infectious diseases are part of our existence; we exist because they exist, and they exist since we exist, as many of them use the Homo sapiens species as a host. In addition, pathogen infection and replication are dependent on population density, with pathogens becoming or remaining endemic in a highly dense population [14,15]. This was already known in ancient history with the Harappan civilization, in the present day Pakistan, building a brick sewer system before 2000 BC, that probably was intended for proper sanitation [16]. With the current increasing urbanization of the world’s population and ever more travel movements we are more at risk for infectious outbreaks [17,18].

There are over 1,000 virus species known to infect humans [1], and there are currently 10^{31} viral particles on Earth [2], of which a small number might be identified as infectious for humans in the future, and of which some might be zoonotic viral infections that can potentially infect humans [3]. Estimates indicate that up to 60% of human infectious diseases are from zoonotic origin [19], exemplified by the recent outbreak of SARS-CoV-2 [20]. Not only infectious diseases can directly affect an individual's health, but they can also have broader consequences. What is often overlooked is that infectious diseases have also shaped economic, political and social aspects of our society. Black Death (Yersinia pestis in different forms) is an example of an infectious disease that impacted society greatly. It was the largest catastrophe to have ever happened to mankind, resulting in the death of one third of Europe's population around the 1300s [21,22]. Large outbreaks did not only have deleterious effects but also resulted in improvements in health care, and pushed the need for epidemiologic insight in prevention, immunity and antimicrobial treatment [8]. The SARS-CoV-2 outbreak also necessitated an urgent update of our surveillance of the infection in all details, and therefore viral DNA sequencing needs [23-25].

Infectious disease burden

Worldwide, the leading cause of death is thought to be ischemic heart disease with a crude death rate (CDR) of 115.3 per 100,000 individuals in 2019 [26] (Figure 1). However, when taking infectious and parasitic diseases into account and all lower respiratory infections, the combined death rate is 100 per 100,000 individuals, illustrating the impact of infectious diseases on human health. Underestimations of infectious diseases have been common due to absent or inconsistent surveillance, identification, and registration by healthcare organizations [27]. This problem can be larger in underdeveloped countries where infectious diseases are more frequent than ischemic heart diseases. The latest disease burden statistics presented by the WHO date from 2019, and are excluding an estimate of over six million deaths worldwide due to the outbreak of SARS-CoV-2 [28]. SARS-CoV-2 epidemiological data published in 2021 showed a crude death rate of 81.7 per 100,000 individuals in the Netherlands and 180.9 per 100,000 individuals in Belgium, further increasing the infectious disease burden [29]. Infectious disease are also part of the cause for diarrhoeal diseases and can cause cancer, with estimates of approximately one in six cancers having an infectious origin [30]. With cancer having a combined total CDR of 120.6 [26], another 20.1 deaths per 100,000 would be linked to infectious disease for a total of approximately 120 deaths per 100,000 – surpassing ischemic heart disease, and illustrating the impact of infectious diseases have on human health and the importance of studying this topic.

Microorganisms and metagenomics

In 1677, Dutch merchant Antonie van Leeuwenhoek wrote in his letters to the Royal Society about 'concerning little animals' he observed in several materials using one of his own custom-built microscopes [31], which must have been the first visualisation of individual bacteria. After this first known recognition of microorganisms, it took two more centuries before Robert Koch developed new methods to grow microorganisms. Koch isolated and pinpointed a bacterium to cause tuberculosis: Mycobacterium tuberculosis [32,33] and formulated criteria for establishing the causality of a microbe (Koch's postulates) [34]. Shortly afterwards, in 1898, Martinus Beijerink, who received his doctorate at the University of Leiden, discovered the tobacco mosaic pathogen. Martinus Beijerink called it a 'virus' and he is now considered as one of the founders of virology [35].

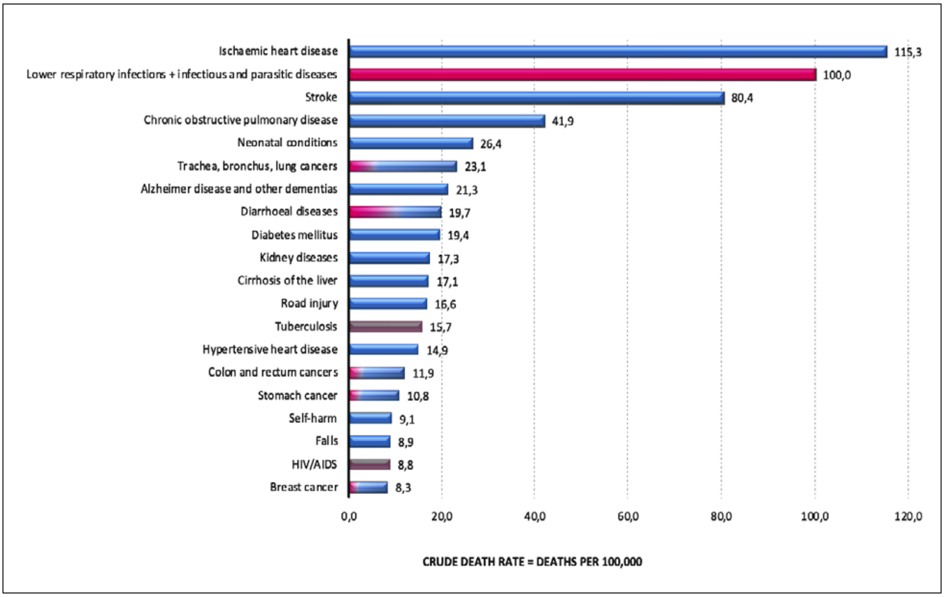
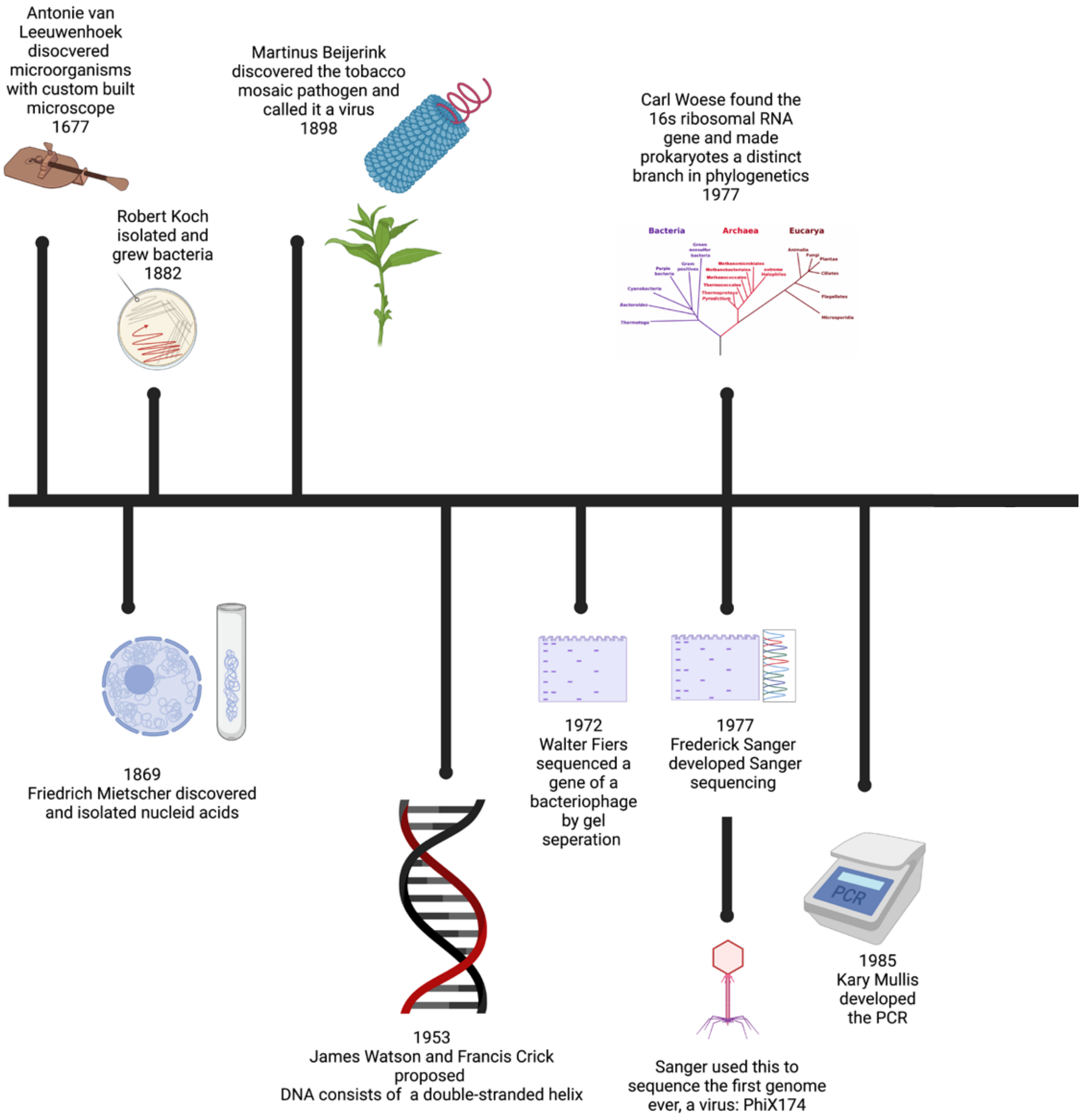


Figure 1. Disease burden worldwide: Top 20 causes of death worldwide.

Data adapted from World Health Organization, last presented for the year 2019 [26], therefore death cause data excludes a majority of COVID-19 deaths. Crude death rate (CDR) associated with infectious diseases are shown in red. CDR of lower respiratory infections and infectious and parasitic diseases is the total number including the CDR of Tuberculosis and HIV/AIDS, therefore the CDR of these individual diseases are masked.

Another century later Carl Woese distinguished the different biological kingdoms using the common 16S ribosomal RNA gene present in prokaryotes. Because this region is quite well preserved amongst species, it is a good basis for phylogenetics [36]. From this moment the prokaryotes were distinct in the tree of life and later 16S amplicon sequencing was marked as the start of microbial sequencing. This was achieved first by means of Sanger sequencing, simply by looking at short limited sequences at once and later using next-generation sequencing looking at a high number of sequences at once. In 1996 (Stein et al. 1996) sequenced


Antonie van Leeuwenhoek disocvered microorganisms with custom built microscope 1677
Robert Koch isolated and grew bacteria 1882
Martinus Beijerink discovered the tobacco mosaic pathogen and called it a virus 1898
Carl Woese found the 16s ribosomal RNA gene and made prokaryotes a distinct branch in phylogenetics 1977
Bacteria
Archaea
Eucarya
Green nonsulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Bacteroides
Thermotoga
Methanomicrobiales
Methanobacteriales
Methanococcales
Thermococcales
Thermoproteus
Pyrodictium
extreme Halophiles
Animalia
Fungi
Plantae
Ciliates
Flagellates
Microsporidia
1869 Friedrich Mietscher discovered and isolated nucleid acids
1953 James Watson and Francis Crick proposed DNA consists of a double-stranded helix
1972 Walter Fiers sequenced a gene of a bacteriophage by gel seperation
1977 Frederick Sanger developed Sanger sequencing
Sanger used this to sequence the first genome ever, a virus: PhiX174
PCR
1985 Kary Mullis developed the PCR

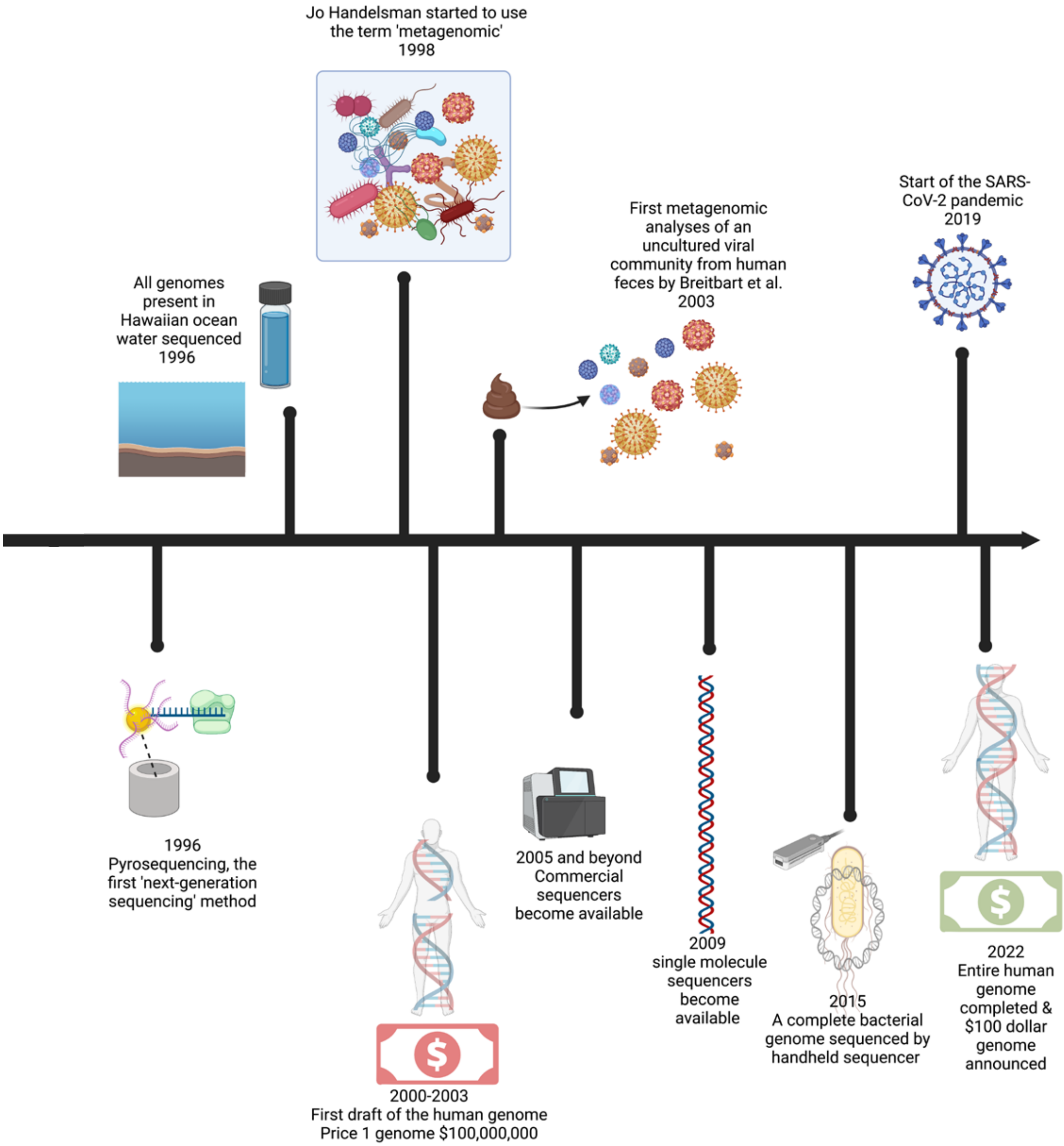

Jo Handelsman started to use the term 'metagenomic' 1998
All genomes present in Hawaiian ocean water sequenced 1996
First metagenomic analyses of an uncultured viral community from human feces by Breitbart et al. 2003
Start of the SARS-CoV-2 pandemic 2019
1996 Pyrosequencing, the first 'next-generation sequencing' method
2000-2003 First draft of the human genome Price 1 genome $100,000,000
2005 and beyond Commercial sequencers become available
2009 single molecule sequencers become available
2015 A complete bacterial genome sequenced by handheld sequencer
2022 Entire human genome completed & $100 dollar genome announced

Figure 2. Timeline with milestones in microbiology and sequencing.

Above: important events over time in microbiology, virology and metagenomics, and below: milestones in molecular genetics and sequencing. Adapted from the timeline of the study of Escobar-Zepeda et al. [40] Created using Biorender.com.

Hawaiian ocean water to look at all the genomes in a sample [37], pioneering the field towards metagenomics, though the name metagenomic sequencing was coined in 1998 when Jo Handelsman used the term metagenomics in one of her studies [38], meaning sequencing all metagenomes present in a sample (see Figure 2). The first metagenomic analysis of the uncultured viral community present in human feces was studied in Breitbart et al. in 2003 [39].

History of sequencing

Just before Robert Koch isolated and pinpointed a specific disease-causing micro-organism, Friedrich Mietscher was able to discover and isolate DNA from cells for the first time in 1869. He first observed there was another structure with different kind of characteristics [41]. James Watson and Francis Crick proposed that the DNA structure consisted of a right-handed helix composed of two anti-parallel DNA strands [42]. Rosalind Franklin in her fellow research on the molecular structure of RNA/DNA and her X-Ray diffraction work provided insight in this structure that enabled them to publish this work [43]. Many attempts were undertaken to sequence DNA to determine the order of nucleotides, and Robert Holley in 1965 sequenced the first transfer RNA extracted from yeast [44]. Shortly after, in 1968, Crick published the genetic code behind all the different amino acids [45].The first gene of which the complete nucleotide code was deciphered was that of a virus, as in 1972 Walter Fiers from Belgium sequenced the first gene of bacteriophage MS2. The first sequencing protocol consisted of digesting the virus RNA into small pieces and then separating them on a gel [46]. Frederick Sanger in 1977 developed Sanger sequencing based on a chain termination method that generates partially fragmented DNA, each fragment with a radiolabelled termination, enabling the sequence to be constructed [47] Sanger also used this method to sequence and determine the nucleotide order of the first genome ever, a virus, bacteriophage PhiX174 [48]. Up to this day, PhiX174 is a virus that haunts genetic research in every sector, since one of the large sequence companies, Illumina, advises to sequence this virus in every run for a quality control check; indirectly leading to contaminated assembled genomes being uploaded to public databases, including parts of this PhiX174 genome [49]. It was also in the group of Sanger where there was the first need for computer processing of DNA data [50], whereas computer processing of biological data, bioinformatics, was previously mainly still applied in the protein field with Margaret Dayhoff probably being the first bioinformatician [50].

In 1985 Kary Mullis developed the PCR, a method we are still currently using [51]. Merely one year later Leroy Hood and Michael Hunkapiller automated the Sanger sequencing method making sequencing possible in a quicker way [52]. In order to expand sequence capacity, pyrosequencing was introduced in 1996 [53] as the first high throughput or next-generation sequence method.

Next-generation sequencing (NGS)

The development of massively parallel sequencing contributed to the first drafts of the human genome in 2000 and 2003 [54-56]. Sanger sequencing was subsequently used to sequence Craig Venter's genome for 100 million dollars, while resequencing James Watson's genome was less than 1 million dollars using NGS [54,57]. Around that time several other companies also started to use a similar method for NGS: Solexa in 1998 (later acquired by Illumina); 454 in 2005 (acquired by Roche in 2007); SOLiD in 2007; and the IonTorrent system of Life Technologies in 2001 [58].

While NGS methods first aimed at sequencing short sequence fragments, the new era of NGS sequencers focuses on single molecule sequencing, first described in 2009 [59]. With this technique performed now by Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT), a complete genome can be sequenced in one single run. These longer sequence reads greatly improve assembling novel genomes [60]. In 2015 a complete bacterial genome was assembled using the ONT method [61]. And only in this year, 2022, the complete human genome has been finally completely sequenced by means of several different sequencing methods including Illumina, PacBio and ONT filling in the last blanks of the human genome that still currently existed [62]. Whereas sequencing Watson's genome first had a price of $1 million, in 2014 the 1000-dollar human genome became available [63]. Sequencing costs are still declining, and in 2020 the author of this thesis paid €300 for privately sequencing her whole genome and since this year, 2022, the $100 human genome is available [64].

The need for informatics in biology: bioinformatics

With the growing number of sequence reads that need interpretation, bioinformatics becomes of vital importance. Bioinformatics started with protein analysis [50], as Margaret Dayhoff, 'the mother and father of bioinformatics', used punched-card business machines to establish molecular energies of organic molecules as these calculations could not be handled on regular calculators [65]. Together with Robert S. Ledley she developed a way to use computational resources in order to establish protein structure. The software 'COMPROTEIN' running on a IBM7090 needed for this was written in FORTRAN on punch cards (see Figure 3), and it bears resemblance to current *de novo* sequence assembly methods [66]. She also developed the one-letter amino acid codes that we still use today [67].

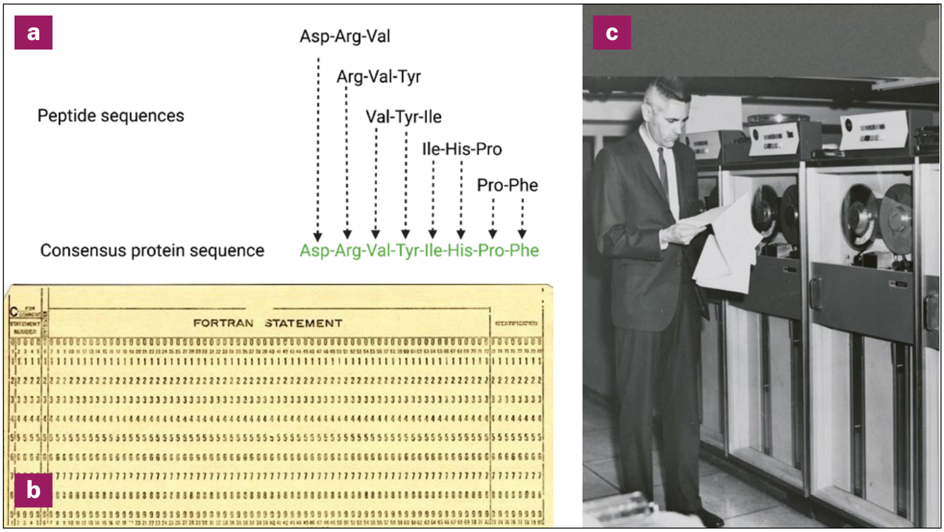


Figure 3. The first bioinformatics software.

(a) Overview of COMPROTEIN. COMPROTEIN used peptide sequences as input and gave a consensus protein sequence as output. Created using Biorender.com. **(b)** FORTRAN punch card. COMPROTEIN was written in the language FORTRAN, only one line of code could be punched per punch card. **(c)** An IBM 7090 mainframe that could run COMPROTEIN.

Two decades later came the need to use computational methods to aid the analysis of nucleotides for comparisons, calculations and matching patterns [68]. Roger Staden wrote the first software to analyse DNA data from Sanger sequencing. His software looked for overlap between gel reads, to join reads into contigs, and to annotate sequence files [69]. He also extended the DNA alphabet with codes for when base calling could quantitatively not be correct, and this is now recorded in the official nomenclature for unclear bases in sequences [70], Roger Staden's software can currently still be downloaded [71].

Next-generation sequencing in clinical settings

Nowadays, utilizing sequencing within a clinical setting is common practice within human genetics and pathology, but its use in microbiology still lags behind. In clinical genetics, the identification of the first disease causing mutations were explored mid-20th century. Linkage analysis was suggested to make a connection between a genome locus and a disease [72], for instance the Huntington disease gene in 1983 [73]. In 1986 the first gene CYBB was linked to a chronic granulomatous disease without exactly knowing what the function of the gene was [74]. In 2009, an autosomal dominant mutation was found to be causal for Freeman-Sheldon syndrome [75]. Shortly after, whole exome sequencing became common to find causes for autosomal recessive diseases [76]. Whole genome sequencing, having a larger horizontal genome coverage of the entire human genome, is since 2010 also used for finding disease causes [77].

In microbiology, the first attempt for metagenomic sequencing in Hawaiian ocean water was already undertaken in 1996 [37] and whole viral genomes were sequenced in 2004 [78]. Sequencing was not popular in a clinical setting, due the wide availability of fast traditional methods to identify pathogens. In clinical virology, viral culture, PCR, or serology testing were used to detect 'known' viral pathogens, and later, multiplex PCR reactions were used to detect several viruses at once. Viral metagenomics, investigating all viral nucleotides from an often uncultured sample, made it possible to identify novel and unexpected but previously identified viruses [79,80]. Viral metagenomics, testing in an unbiased and agnostic way, has been suggested as a powerful tool for virus discovery in a clinical healthcare setting [81]. Viral identification and discovery are in great need, with a constant threat of zoonotic virus transfer and an estimate of at least 320,000 mammalian viruses that still need to be discovered [82]. In the beginning of the recent pandemic, it was a viral metagenomic technique that identified the disease-causing viral pathogen directly from patients' material, and established the genome at once [23,24].

Clinical viral metagenomics

Sequencing all genomes in a sample at once

Metagenomics enables detection of all the genetic material of organisms present ina sample, making it a pathogen-agnostic approach for detecting common and rare or novel pathogens that are not included in conventional testing. Beforehand, a clinician does not need to have a hypothesis of what pathogen is expected, unlike traditional PCR testing. Another benefit is that this technique enables investigation of multiple species at once (Figure 4). The first case reports on the identification of viral pathogens in patients by means of metagenomics focused mainly on encephalitis patients [83–95]. Immunocompromised patients are of specific interest as they are at increased risk of infections by unexpected and novel viruses and bacteria without having regular symptoms [83,96]. In 2019, two prospective clinical utility studies were performed and published where viral metagenomics was used in parallel with conventional diagnostics [97,98]. These findings demonstrated that viral enrichment was beneficial for virus detection improving the potential for diagnostics [97,99-101].

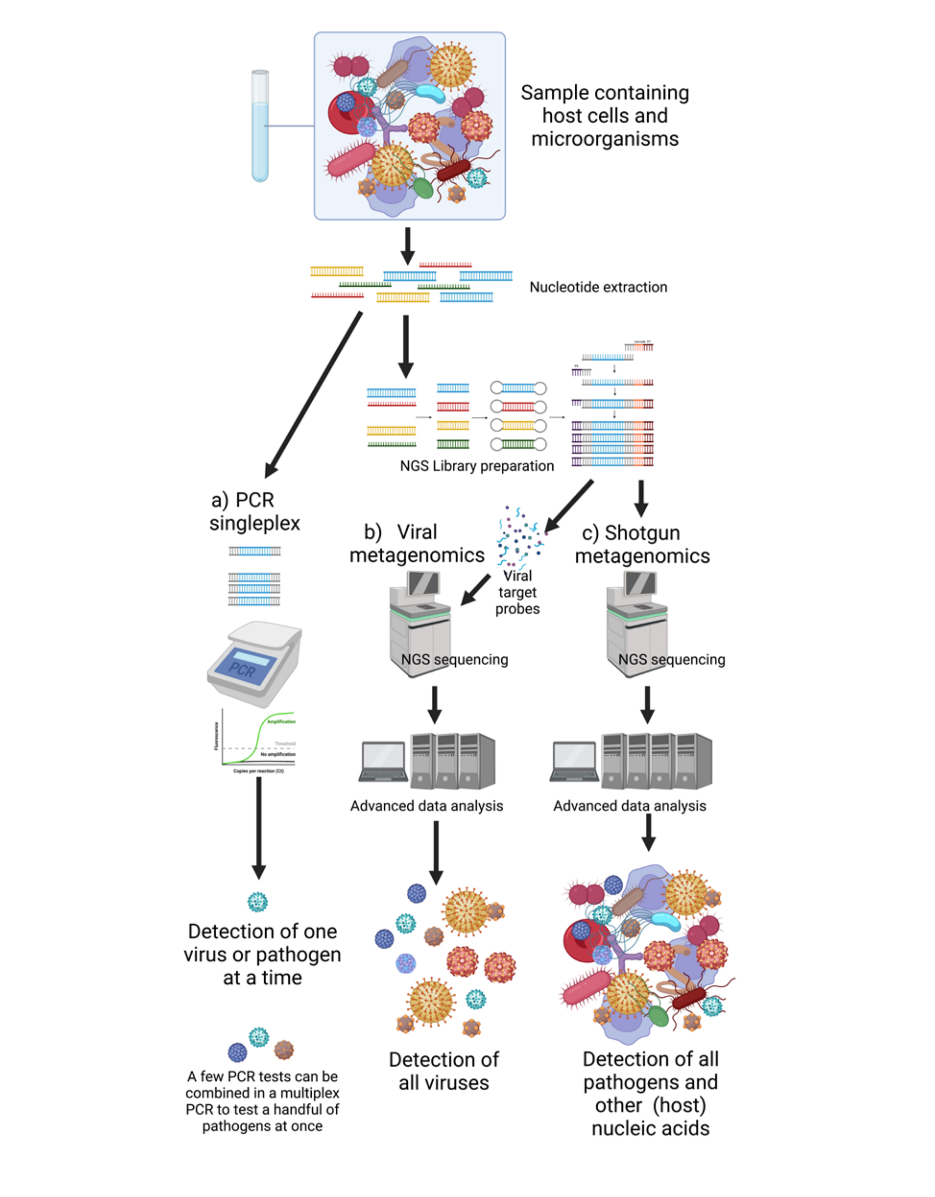


Figure 4. Overview of PCR and (targeted) metagenomic sequencing methods.

An overview of three methods for viral pathogen detection in a sample containing host cells and microorganisms. a) Conventional singleplex PCR testing of just one virus target at a time, though in a multiplex test several PCR tests can be combined to test multiple viruses at once b) Targeted viral metagenomics, using viral probes to capture only certain viral sequences after library preparation, c) Sequencing all genetic material in a sample using shotgun metagenomics, so all pathogens/species/genetic host material is available after data analysis. Created using Biorender.com.

While the genomes of viruses are sequenced, information about the type/subtype is often additionally available as well as information about resistance mutations.Another benefit of metagenomics is that a virus will not be missed due to mismatching primer pair sequences in cases of virus mutations in the primer regions. Recently, metagenomics has shown to be useful for the discovery and classification of the SARS-Cov-2 virus directly from patient material [24,25].

Increase in diagnostic yield

The additional diagnostic yield of metagenomic sequencing reported in literature is 28.73% (CI 19.80-37.63), with diagnostic yield defined as finding a potentially causative pathogen by means of metagenomic sequencing after initial/diagnostic standard tests were negative for a pathogen. The numbers are based on a systematic review and meta-analysis (Figure 5) performed by the author of this thesis and the metagenomics group. Forty-seven studies were included for systematic review, of which 27 studies were selected that targeted potential causative pathogens by means of metagenomic sequencing after the initial or diagnostic testing was negative in patients with a wide range of clinical syndromes [97,98,102-126].

Viral metagenomics laboratory protocol

For viral metagenomic sequencing, nucleotides are first extracted from a sample of RNA, DNA or both. Then in most cases RNA is converted into DNA by means of cDNA synthesis. In the beginning of the library preparation, genetic material is fragmented, after which the sequences are end-repaired and ligated with adapters needed for sequencing (Figure 6). The nucleotide sequences are ligated with barcodes to differentiate samples after sequencing. Usually, library preparation protocols involve amplification of the prepared libraries [83,97,98,128-132]. When no special filtering or additional target probes are used, this is called shotgun metagenomics (Figure 4). If required, steps can be undertaken to filter out more human cells, or ribosomal RNA before the library prep either via centrifugation or additional prep kits [133-137]. After the library preparation it is also possible to use a targeted probe kit designed to capture specifically viral sequences [100,103,133,138].

The highly sensitive PCR procedure is the gold standard to establish whether a virus is present in a given sample, and currently it is a challenge to establish a similar sensitivity using metagenomics. Due to highly abundant host material and microbiome, viral pathogen sequences are like the proverbial needle in a haystack and sensitivity remains to be improved. Additionally, it is common to detect contaminating microbial genetic material from reagents specifically used in test kits, the 'kitome' [139,140]. The current protocols are expensive and time-consuming [141] and, due to the complexity, not every lab can perform this method of sequencing [140,142].

Research was needed to further improve the viral metagenomic test sensitivity and reduce the amount of background (host) sequences and contaminating sequences. Due to the limited number of studies presenting data on the diagnostic yield of the test, additional patient cohorts with specific clinical syndromes needed to be tested to investigate the yield in different populations. In addition, the potential for virus discovery straight from clinical samples, while utilizing viral metagenomics, had to be tested for accuracy and applicability.

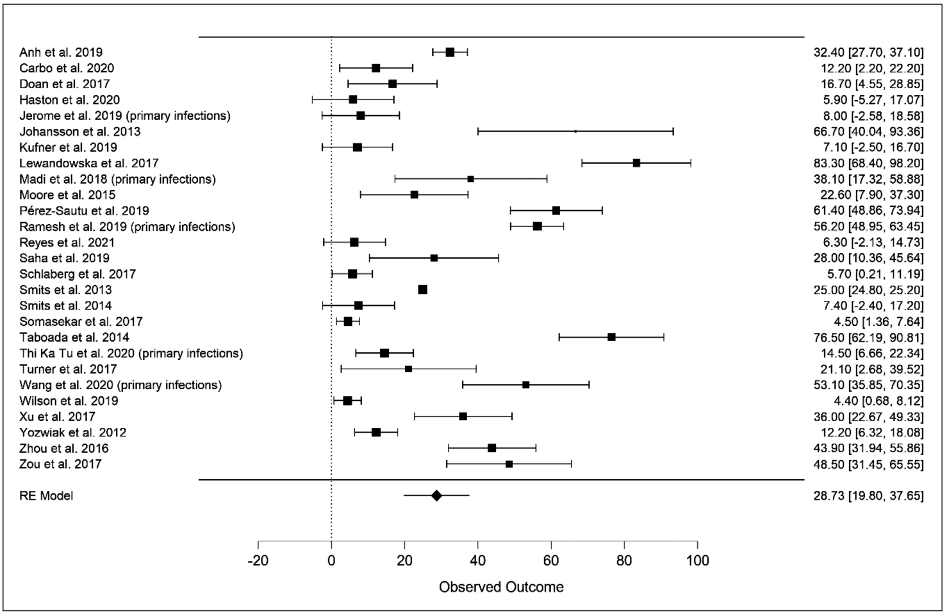


Figure 5. Meta-analysis of studies using metagenomics as a diagnostic tool to detect infectious diseases, in patients with a wide range of clinical syndromes.

Derived from author's unpublished data of a systematic review on viral metagenomics and diagnostic yield. A search in PubMed was conducted and reference lists were searched in December 2020, of which 644 studies were obtained. Of the 644 studies, microbiome/virome studies, technical validation studies, reviews, opinion papers, studies in other languages than English, case reports, and studies with sample number <7 were excluded. Forty-seven remained for systematic review, 27 identified potentially causative pathogen by means of metagenomic sequencing after the initial or diagnostic testing was negative. A random effects meta-analysis was performed given the heterogeneity of the 27 papers using JASP [127].

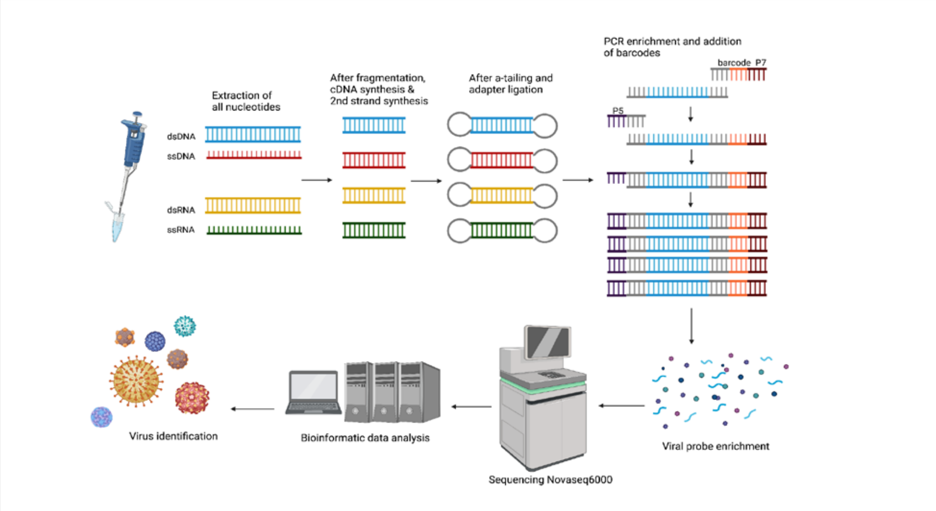


Figure 6. Steps in our current metagenomics NGS protocol applicable in a clinical setting.
Library preparation protocol as performed in this thesis starts after extracting all nucleotides of a clinical sample with enzymatic fragmentation, cDNA synthesis and 2nd strand synthesis. Next, A-tailing of the sequences is carried out and adapters are ligated to the blunt ended A-tailed overhang. A PCR enrichment step is carried out and sample specific barcodes are added to the sample.

To select viruses and minimize sequence reads of other (host) species, a viral probe enrichment can be carried out, though samples could also be sequenced for a shotgun approach without probe enrichment. Sequencing is carried out on an Illumina sequencer followed by bioinformatic data analysis using a taxonomic classifier for virus identification. Created using Biorender.com.

Viral metagenomics data analysis

The large amount of data that needs to be handled after generation of sequence reads is often a bottleneck, since specifically trained personnel, hardware (both sequencing platform and computing methods) and software is required. Data analysis of up to millions of sequence reads usually starts with removing bad quality sequence reads and sequence adapters. These adapters are DNA fragments that aid sequencing but do not provide information about the original content of the sample. It is optional to first filter out host material by means of mapping to the human genome or to first assemble reads into genome parts called contigs. Thereafter, read classification takes place by a taxonomic classifier to lay out the virome present in a sample. If one is additionally interested to zoom in at the genome of specific viruses that were detected, immediate mapping against the reference genomes can be performed, followed by variant calling analysis if more specific characterization is needed. If the user is interested in completely novel viruses, a *de novo* assembly step can be performed after the first step.

Bioinformatic sequence analyses tools are mainly orientated on human genetics and most taxonomic classification tools have originally been designed for bacteria. Therefore, analysis and optimization of sensitivity and specificity specifically for viral metagenomic testing is needed. Furthermore, the impact of filtering out host reads on accuracy needs to be investigated for viral metagenomics, and the possibilities for quantification and typing by means of NGS.

Thesis aim

The research in this thesis has several aims. Firstly, establishing the diagnostic yield of viral metagenomics in specific patient populations: patients suspected of encephalitis and travellers returning with febrile illness. Secondly, the identification, typing and quantification of viruses by means of viral metagenomics as a diagnostic tool is evaluated. Another aim is to improve sensitivity and specificity of the wet and dry (bioinformatic) lab components of viral metagenomics, in order to achieve a better performance of the method in clinical practice.

Lastly, we investigated the best methods and approaches of performing genetic analysis of just one viral genome.

Outline of this thesis

This thesis is focusing on diagnostic yield, clinical findings and enhancing technical opportunities in viral metagenomics. **Chapter 7 and 8** are devoted to **whole genome sequencing** of one specific viral genome by means of metagenomics, and a comparison of sequencing methods of SARS-CoV-2.

Chapter 2 is focused on the estimated **diagnostic yield**, the number of extra viruses that can be found using metagenomics after traditional test remain negative in cases of meningoencephalitis. In this invited review, the technical and bioinformatic advances of viral metagenomics and the remaining challenges are explained.

To further enhance sensitivity, shotgun metagenomic sequencing and sequencing with viral capture probes was compared in a cohort of **encephalitis patients** with a known virus in **chapter 3**. In this chapter, an additional cohort of adult and paediatric hematological patients without etiologic agent detected by conventional assays was assessed using metagenomic sequencing.

Chapter 4 describes a metagenomics protocol with viral capture probes that was applied on a cohort of **international travellers** with febrile illness. We focus on confirming and typing of the original positive test results, and on detection of viruses that remained undetected using traditional assays.

Almost a billion sequence reads generated for 88 respiratory samples were used to assess the performance of various bioinformatic **taxonomic classification** tools based on the original qPCR results in **chapter 5**.

In **chapter 6** a metagenomics protocol is applied to **quantify** the number of viruses in transplant patients over the course of the disease. Thus, in addition to establishing the type of virus, the number of viral particles was assessed.

In order to assess the performance of a metagenomic protocol for **virus discovery** directly in a patient sample, viral metagenomic sequencing is performed on clinical samples containing SARS-CoV, MERS-CoV and SARS-CoV-2 with viral databases from the time of original discovery. In **chapter 7,** we explain the process, and the steps taken for virus discovery in a clinical setting.

In chapter 8 it is described how **SARS-CoV-2** samples were handled by one metagenomic sequencing and four amplicon-based **WGS protocols** of three different sequence platforms to assess the performance for analyses of this one specific genome.

Chapter 9 contains the **general discussion** on the methodological breakthroughs, remaining challenges in viral metagenomics and viral sequencing. In addition, the future opportunities for metagenomic NGS in the future viral or molecular diagnostic laboratories are discussed.

References

[1] G. Lasso et al., ‘A Structure-Informed Atlas of Human-Virus Interactions’, Cell, vol. 178, no. 6, pp. 1526-1541.e16, Sep. 2019, doi: 10.1016/j.cell.2019.08.005.

[2] M. Breitbart and F. Rohwer, ‘Here a virus, there a virus, everywhere the same virus?’, Trends in Microbiology, vol. 13, no. 6, pp. 278–284, Jun. 2005, doi: 10.1016/j.tim.2005.04.003.

[3] L. H. Taylor, S. M. Latham, and M. E. J. woolhouse, ‘Risk factors for human disease emergence’, Phil. Trans. R. Soc. Lond. B, vol. 356, no. 1411, pp. 983–989, Jul. 2001, doi: 10.1098/rstb.2001.0888.

[4] Z. Diao, D. Han, R. Zhang, and J. Li, ‘Metagenomics next-generation sequencing tests take the stage in the diagnosis of lower respiratory tract infections’, Journal of Advanced Research, vol. 38, pp. 201–212, May 2022, doi: 10.1016/j.jare.2021.09.012.

[5] K. N. Govender, T. L. Street, N. D. Sanderson, and D. W. Eyre, ‘Metagenomic Sequencing as a Pathogen-Agnostic Clinical Diagnostic Tool for Infectious Diseases: a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies’, J Clin Microbiol, vol. 59, no. 9, pp. e02916-20, Aug. 2021, doi: 10.1128/JCM.02916-20.

[6] I. Barberis, N. L. Bragazzi, L. Galluzzo, and M. Martini, ‘The history of tuberculosis: from the first historical records to the isolation of Koch’s bacillus’, J Prev Med Hyg, vol. 58, no. 1, pp. E9–E12, Mar. 2017.

[7] M. Mohammadpour, M. Abrishami, A. Masoumi, and H. Hashemi, ‘Trachoma: Past, present and future’, Journal of Current Ophthalmology, vol. 28, no. 4, pp. 165–169, Dec. 2016, doi: 10.1016/j.joco.2016.08.011.

[8] D. Huremović, ‘Brief History of Pandemics (Pandemics Throughout History)’, in Psychiatry of Pandemics, D. Huremović, Ed. Cham: Springer International Publishing, 2019, pp. 7–35. doi: 10.1007/978-3-030-15346-5_2.

[9] J. K. Taubenberger and D. M. Morens, ‘1918 Influenza: the Mother of All Pandemics’, Emerg. Infect. Dis., vol. 12, no. 1, pp. 15–22, Jan. 2006, doi: 10.3201/eid1209.05-0979.

[10] ‘UNAIDS statistics’, Accessed: Aug. 03, 2022. [Online]. Available: https://www.unaids.org/en/resources/fact-sheet

[11] C. G. Kurland and S. G. E. Andersson, ‘Origin and Evolution of the Mitochondrial Proteome’, Microbiol Mol Biol Rev, vol. 64, no. 4, pp. 786–820, Dec. 2000, doi: 10.1128/MMBR.64.4.786-820.2000.

[12] M. Horie et al., ‘Endogenous non-retroviral RNA virus elements in mammalian genomes’, Nature, vol. 463, no. 7277, pp. 84–87, Jan. 2010, doi: 10.1038/nature08695.

[13] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, ‘The Human Microbiome Project’, Nature, vol. 449, no. 7164, pp. 804–810, Oct. 2007, doi: 10.1038/nature06244.

[14] R. M. May and R. M. Anderson, ‘Endemic infections in growing populations’, Mathematical Biosciences, vol. 77, no. 1–2, pp. 141–156, Dec. 1985, doi: 10.1016/0025-5564(85)90093-8.

[15] R. M. Anderson and R. M. May, Eds., Population Biology of Infectious Diseases: Report of the Dahlem Workshop on Population Biology of Infectious Disease Agents Berlin 1982, March 14 – 19. Berlin, Heidelberg: Springer Berlin Heidelberg, 1982. doi: 10.1007/978-3-642-68635-1.

[16] 'Harappan civilization building a brick sewer system', Accessed: Aug. 03, 2022. [Online]. Available: https://www.jstor.org/stable/26426051

[17] J. F. Lindahl and D. Grace, 'The consequences of human actions on risks for infectious diseases: a review', Infection Ecology & Epidemiology, vol. 5, no. 1, p. 30048, Jan. 2015, doi: 10.3402/iee.v5.30048.

[18] C. Connolly, R. Keil, and S. H. Ali, 'Extended urbanisation and the spatialities of infectious disease: Demographic change, infrastructure and governance', Urban Studies, vol. 58, no. 2, pp. 245–263, Feb. 2021, doi: 10.1177/0042098020910873.

[19] 'Zoonotic origin', Accessed: Aug. 03, 2022. [Online]. Available: https://iris.wpro.who.int/bitstream/handle/10665.1/7819/9789290615040_eng.pdf

[20] K. Dhama et al., 'SARS-CoV-2 jumping the species barrier: Zoonotic lessons from SARS, MERS and recent advances to combat this pandemic virus', Travel Medicine and Infectious Disease, vol. 37, p. 101830, Sep. 2020, doi: 10.1016/j.tmaid.2020.101830.

[21] O. Benedictow, 'The black death: the greatest catastrophe ever.', Hist Today, p. 55(3):42–9, 2005.

[22] W. Scheidel, The great leveler: violence and the history of inequality from the Stone Age to the twenty-first century. Princeton, New Jersey: Princeton University Press, 2017.

[23] N. Zhu et al., 'A Novel Coronavirus from Patients with Pneumonia in China, 2019', N Engl J Med, vol. 382, no. 8, pp. 727–733, Feb. 2020, doi: 10.1056/NEJMoa2001017.

[24] C. Huang et al., 'Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China', The Lancet, vol. 395, no. 10223, pp. 497–506, Feb. 2020, doi: 10.1016/S0140-6736(20)30183-5.

[25] P. Zhou et al., 'Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin', Microbiology, preprint, Jan. 2020. doi: 10.1101/2020.01.22.914952.

[26] 'CDR', Accessed: Aug. 03, 2022. [Online]. Available: ttps://www.who.int/data/maternal-newborn-child-adolescent-ageing/indicator-explorer-new/mca/crude-death-rate-(deaths-per-1000-population)

[27] C. L. Gibbons et al., 'Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods', BMC Public Health, vol. 14, no. 1, p. 147, Dec. 2014, doi: 10.1186/1471-2458-14-147.

[28] 'https://covid19.who.int', Accessed: Aug. 03, 2022. [Online]. Available: https://covid19.who.int

[29] S. Soneji, H. Beltrán-Sánchez, J. W. Yang, and C. Mann, 'Population-level mortality burden from novel coronavirus (COVID-19) in Europe and North America', Genus, vol. 77, no. 1, p. 7, Dec. 2021, doi: 10.1186/s41118-021-00115-9.

[30] D. M. Parkin, 'The global health burden of infection-associated cancers in the year 2002', Int. J. Cancer, vol. 118, no. 12, pp. 3030–3044, Jun. 2006, doi: 10.1002/ijc.21731.

[31] A. Leewenhoeck, ‘Observations, Communicated to the Publisher by Mr. Antony van Leewenhoeck, in a Dutch Letter of the 9th of Octob. 1676. Here English’d: concerning Little Animals by Him Observed in Rain-Well-Sea. and Snow Water; as Also in Water Wherein Pepper Had Lain Infused’, Philosophical Transactions, vol. Vol. 12 (1677-1678), pp. 821–831.

[32] S. R. Lakhani, ‘Early clinical pathologists: Robert Koch (1843-1910).’, Journal of Clinical Pathology, vol. 46, no. 7, pp. 596–598, Jul. 1993, doi: 10.1136/jcp.46.7.596.

[33] S. M. Blevins and M. S. Bronze, ‘Robert Koch and the “golden age” of bacteriology’, International Journal of Infectious Diseases, vol. 14, no. 9, pp. e744–e751, Sep. 2010, doi: 10.1016/j.ijid.2009.12.003.

[34] R. Koch, Xth International Congress of Medicine, vol. Berlin, 1890.

[35] A. Kammen, ‘Beijerinck’s contribution to the virus concept — an introduction’, in 100 Years of Virology, C. H. Calisher and M. C. Horzinek, Eds. Vienna: Springer Vienna, 1999, pp. 1–8. doi: 10.1007/978-3-7091-6425-9_1.

[36] C. R. Woese and G. E. Fox, ‘Phylogenetic structure of the prokaryotic domain: The primary kingdoms’, Proc. Natl. Acad. Sci. U.S.A., vol. 74, no. 11, pp. 5088–5090, Nov. 1977, doi: 10.1073/pnas.74.11.5088.

[37] J. L. Stein, T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong, ‘Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon’, J Bacteriol, vol. 178, no. 3, pp. 591–599, Feb. 1996, doi: 10.1128/jb.178.3.591-599.1996.

[38] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, ‘Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products’, Chemistry & Biology, vol. 5, no. 10, pp. R245–R249, Oct. 1998, doi: 10.1016/S1074-5521(98)90108-9.

[39] M. Breitbart et al., ‘Metagenomic Analyses of an Uncultured Viral Community from Human Feces’, J Bacteriol, vol. 185, no. 20, pp. 6220–6223, Oct. 2003, doi: 10.1128/JB.185.20.6220-6223.2003.

[40] A. Escobar-Zepeda et al., ‘Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics’, Sci Rep, vol. 8, no. 1, p. 12034, Dec. 2018, doi: 10.1038/s41598-018-30515-5.

[41] R. Dahm, ‘Friedrich Miescher and the discovery of DNA’, Developmental Biology, vol. 278, no. 2, pp. 274–288, Feb. 2005, doi: 10.1016/j.ydbio.2004.11.028.

[42] J. D. Watson and F. H. C. Crick, ‘Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid’, Nature, vol. 171, no. 4356, pp. 737–738, Apr. 1953, doi: 10.1038/171737a0.

[43] B. Maddox, ‘The double helix and the “wronged heroine”’, Nature, vol. 421, no. 6921, pp. 407–408, Jan. 2003, doi: 10.1038/nature01399.

[44] R. W. Holley, G. A. Everett, J. T. Madison, and A. Zamir, ‘NUCLEOTIDE SEQUENCES IN THE YEAST ALANINE TRANSFER RIBONUCLEIC ACID’, J Biol Chem, vol. 240, pp. 2122–2128, May 1965.

[45] F. H. C. Crick, ‘The origin of the genetic code’, Journal of Molecular Biology, vol. 38, no. 3, pp. 367–379, Dec. 1968, doi: 10.1016/0022-2836(68)90392-6.

[46] W. M. Jou, G. Haegeman, M. Ysebaert, and W. Fiers, ‘Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein’, Nature, vol. 237, no. 5350, pp. 82–88, May 1972, doi: 10.1038/237082a0.

[47] F. Sanger, S. Nicklen, and A. R. Coulson, ‘DNA sequencing with chain-terminating inhibitors’, Proc. Natl. Acad. Sci. U.S.A., vol. 74, no. 12, pp. 5463–5467, Dec. 1977, doi: 10.1073/pnas.74.12.5463.

[48] F. Sanger et al., ‘Nucleotide sequence of bacteriophage □X174 DNA’, Nature, vol. 265, no. 5596, pp. 687–695, Feb. 1977, doi: 10.1038/265687a0.

[49] S. Mukherjee, M. Huntemann, N. Ivanova, N. C. Kyrpides, and A. Pati, ‘Large-scale contamination of microbial isolate genomes by Illumina PhiX control’, Stand in Genomic Sci, vol. 10, no. 1, p. 18, Dec. 2015, doi: 10.1186/1944-3277-10-18.

[50] ‘Computer processing of DNA sequence data’, Journal of Molecular Biology, vol. 116, no. 1, pp. 29–30, Oct. 1977, doi: 10.1016/0022-2836(77)90116-4.

[51] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich, ‘Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction’, Cold Spring Harbor Symposia on Quantitative Biology, vol. 51, no. 0, pp. 263–273, Jan. 1986, doi: 10.1101/SQB.1986.051.01.032.

[52] L. M. Smith et al., ‘Fluorescence detection in automated DNA sequence analysis’, Nature, vol. 321, no. 6071, pp. 674–679, Jun. 1986, doi: 10.1038/321674a0.

[53] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén, ‘Real-Time DNA Sequencing Using Detection of Pyrophosphate Release’, Analytical Biochemistry, vol. 242, no. 1, pp. 84–89, Nov. 1996, doi: 10.1006/abio.1996.0432.

[54] J. C. Venter et al., ‘The Sequence of the Human Genome’, Science, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001, doi: 10.1126/science.1058040.

[55] J. Schmutz et al., ‘Quality assessment of the human genome sequence’, Nature, vol. 429, no. 6990, pp. 365–368, May 2004, doi: 10.1038/nature02390.

[56] International Human Genome Sequencing Consortium et al., ‘Initial sequencing and analysis of the human genome’, Nature, vol. 409, no. 6822, pp. 860–921, Feb. 2001, doi: 10.1038/35057062.

[57] D. A. Wheeler et al., ‘The complete genome of an individual by massively parallel DNA sequencing’, Nature, vol. 452, no. 7189, pp. 872–876, Apr. 2008, doi: 10.1038/nature06884.

[58] J. M. Heather and B. Chain, ‘The sequence of sequencers: The history of sequencing DNA’, Genomics, vol. 107, no. 1, pp. 1–8, Jan. 2016, doi: 10.1016/j.ygeno.2015.11.003.

[59] E. Check Hayden, ‘Genome sequencing: the third generation’, Nature, p. news.2009.86, Feb. 2009, doi: 10.1038/news.2009.86.

[60] J. Quick, A. R. Quinlan, and N. J. Loman, ‘A reference bacterial genome dataset generated on the MinIONTM portable single-molecule nanopore sequencer’, GigaSci, vol. 3, no. 1, p. 22, Dec. 2014, doi: 10.1186/2047-217X-3-22.

[61] N. J. Loman, J. Quick, and J. T. Simpson, ‘A complete bacterial genome assembled de novo using only nanopore sequencing data’, Nat Methods, vol. 12, no. 8, pp. 733–735, Aug. 2015, doi: 10.1038/nmeth.3444.

[62] S. Nurk et al., ‘The complete sequence of a human genome’, Science, vol. 376, no. 6588, pp. 44–53, Apr. 2022, doi: 10.1126/science.abj6987.

[63] E. Check Hayden, ‘Technology: The $1,000 genome’, Nature, vol. 507, no. 7492, pp. 294–295, Mar. 2014, doi: 10.1038/507294a.

[64] G. Almogy et al., ‘Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform’, Genomics, preprint, May 2022. doi: 10.1101/2022.05.29.493900.

[65] G. Moody, Digital code of life: how bioinformatics is revolutionizing science, medicine, and business. Hoboken, N.J: Wiley, 2004.

[66] M. O. Dayhoff and R. S. Ledley, ‘Comprotein: a computer program to aid primary protein structure determination’, in Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS ’62 (Fall), Philadelphia, Pennsylvania, 1962, pp. 262–274. doi: 10.1145/1461518.1461546.

[67] M. O. Dayhoff and National Biomedical Research Foundation., Atlas of protein sequence and structure. Silver Spring. Md: National Biomedical Research Foundation, 1969.

[68] J. Gauthier, A. T. Vincent, S. J. Charette, and N. Derome, ‘A brief history of bioinformatics’, Briefings in Bioinformatics, vol. 20, no. 6, pp. 1981–1996, Nov. 2019, doi: 10.1093/bib/bby063.

[69] R. Staden, ‘A strategy of DNA sequencing employing computer programs’, Nucl Acids Res, vol. 6, no. 7, pp. 2601–2610, 1979, doi: 10.1093/nar/6.7.2601.

[70] A. Cornish-Bowden, ‘Nomenclature for incompletely specified bases in nucleic acid sequences: rcommendations 1984’, Nucl Acids Res, vol. 13, no. 9, pp. 3021–3030, 1985, doi: 10.1093/nar/13.9.3021.

[71] ‘Staden’, Accessed: Aug. 03, 2022. [Online]. Available: http://staden.sourceforge.net

[72] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis, ‘Construction of a genetic linkage map in man using restriction fragment length polymorphisms’, Am J Hum Genet, vol. 32, no. 3, pp. 314–331, May 1980.

[73] J. F. Gusella et al., ‘A polymorphic DNA marker genetically linked to Huntington’s disease’, Nature, vol. 306, no. 5940, pp. 234–238, Nov. 1983, doi: 10.1038/306234a0.

[74] R. L. Baehner et al., ‘DNA linkage analysis of X chromosome-linked chronic granulomatous disease.’, Proc. Natl. Acad. Sci. U.S.A., vol. 83, no. 10, pp. 3398–3401, May 1986, doi: 10.1073/pnas.83.10.3398.

[75] S. B. Ng et al., ‘Targeted capture and massively parallel sequencing of 12 human exomes’, Nature, vol. 461, no. 7261, pp. 272–276, Sep. 2009, doi: 10.1038/nature08250.

[76] M. Choi et al., ‘Genetic diagnosis by whole exome capture and massively parallel DNA sequencing’, Proc. Natl. Acad. Sci. U.S.A., vol. 106, no. 45, pp. 19096–19101, Nov. 2009, doi: 10.1073/pnas.0910672106.

[77] J. R. Lupski et al., ‘Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy’, N Engl J Med, vol. 362, no. 13, pp. 1181–1191, Apr. 2010, doi: 10.1056/NEJMoa0908094.

[78] J. C. Venter et al., ‘Environmental Genome Shotgun Sequencing of the Sargasso Sea’, Science, vol. 304, no. 5667, pp. 66–74, Apr. 2004, doi: 10.1126/science.1093857.

[79] R. A. Edwards and F. Rohwer, ‘Viral metagenomics’, Nat Rev Microbiol, vol. 3, no. 6, pp. 504–510, Jun. 2005, doi: 10.1038/nrmicro1163.

[80] E. L. Delwart, ‘Viral metagenomics’, Rev. Med. Virol., vol. 17, no. 2, pp. 115–131, Mar. 2007, doi: 10.1002/rmv.532.

[81] S. Svraka, K. Rosario, E. Duizer, H. van der Avoort, M. Breitbart, and M. Koopmans, ‘Metagenomic sequencing for virus identification in a public-health setting’, Journal of General Virology, vol. 91, no. 11, pp. 2846–2856, Nov. 2010, doi: 10.1099/vir.0.024612-0.

[82] S. J. Anthony et al., 'A Strategy To Estimate Unknown Viral Diversity in Mammals', mBio, vol. 4, no. 5, pp. e00598-13, Nov. 2013, doi: 10.1128/mBio.00598-13.

[83] J. R. Brown, T. Bharucha, and J. Breuer, 'Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases', Journal of Infection, vol. 76, no. 3, pp. 225–240, Mar. 2018, doi: 10.1016/j.jinf.2017.12.014.

[84] C. Y. Chiu et al., 'Diagnosis of Fatal Human Case of St. Louis Encephalitis Virus Infection by Metagenomic Sequencing, California, 2016', Emerg. Infect. Dis., vol. 23, no. 10, pp. 1964–1968, Oct. 2017, doi: 10.3201/eid2310.161986.

[85] A. W. D. Edridge et al., 'Novel Orthobunyavirus Identified in the Cerebrospinal Fluid of a Ugandan Child With Severe Encephalopathy', Clinical Infectious Diseases, vol. 68, no. 1, pp. 139–142, Jan. 2019, doi: 10.1093/cid/ciy486.

[86] H. Fridholm et al., 'Human pegivirus detected in a patient with severe encephalitis using a metagenomic pan-virus array', Journal of Clinical Virology, vol. 77, pp. 5–8, Apr. 2016, doi: 10.1016/j.jcv.2016.01.013.

[87] B. Hoffmann et al., 'A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis', N Engl J Med, vol. 373, no. 2, pp. 154–162, Jul. 2015, doi: 10.1056/NEJMoa1415627.

[88] D. Lipowski et al., 'A Cluster of Fatal Tick-borne Encephalitis Virus Infection in Organ Transplant Setting', The Journal of Infectious Diseases, vol. 215, no. 6, pp. 896–901, Mar. 2017, doi: 10.1093/infdis/jix040.

[89] N. T. H. Mai et al., 'Central Nervous System Infection Diagnosis by Next-Generation Sequencing: A Glimpse Into the Future?', Open Forum Infectious Diseases, vol. 4, no. 2, p. ofx046, Apr. 2017, doi: 10.1093/ofid/ofx046.

[90] J. A. Murkey et al., 'Hepatitis E Virus–Associated Meningoencephalitis in a Lung Transplant Recipient Diagnosed by Clinical Metagenomic Sequencing', Open Forum Infectious Diseases, vol. 4, no. 3, p. ofx121, Jul. 2017, doi: 10.1093/ofid/ofx121.

[91] S. N. Naccache et al., 'Diagnosis of Neuroinvasive Astrovirus Infection in an Immunocompromised Adult With Encephalitis by Unbiased Next-Generation Sequencing', Clinical Infectious Diseases, vol. 60, no. 6, pp. 919–923, Mar. 2015, doi: 10.1093/cid/ciu912.

[92] K. Perlejewski et al., 'Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: Unexpected detection of human herpesvirus 1 while searching for RNA pathogens', Journal of Virological Methods, vol. 226, pp. 1–6, Dec. 2015, doi: 10.1016/j.jviromet.2015.09.010.

[93] A. Piantadosi et al., 'Rapid Detection of Powassan Virus in a Patient With Encephalitis by Metagenomic Sequencing', Clinical Infectious Diseases, vol. 66, no. 5, pp. 789–792, Feb. 2018, doi: 10.1093/cid/cix792.

[94] F. Tschumi et al., 'Meningitis and epididymitis caused by Toscana virus infection imported to Switzerland diagnosed by metagenomic sequencing: a case report', BMC Infect Dis, vol. 19, no. 1, p. 591, Dec. 2019, doi: 10.1186/s12879-019-4231-9.

[95] M. R. Wilson et al., 'A novel cause of chronic viral meningoencephalitis: Cache Valley virus: Orthobunyavirus Encephalitis', Ann Neurol., vol. 82, no. 1, pp. 105–114, Jul. 2017, doi: 10.1002/ana.24982.

[96] D. Saylor, K. Thakur, and A. Venkatesan, ‘Acute encephalitis in the immunocompromised individual’, Current Opinion in Infectious Diseases, vol. 28, no. 4, pp. 330–336, Aug. 2015, doi: 10.1097/QCO.0000000000000175.

[97] M. R. Wilson et al., ‘Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis’, N Engl J Med, vol. 380, no. 24, pp. 2327–2340, Jun. 2019, doi: 10.1056/NEJMoa1803396.

[98] Kufner et al., ‘Two Years of Viral Metagenomics in a Tertiary Diagnostics Unit: Evaluation of the First 105 Cases’, Genes, vol. 10, no. 9, p. 661, Aug. 2019, doi: 10.3390/genes10090661.

[99] B. M. O’Flaherty et al., ‘Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing’, Genome Res., vol. 28, no. 6, pp. 869–877, Jun. 2018, doi: 10.1101/gr.226316.117.

[100] T. N. Wylie, K. M. Wylie, B. N. Herter, and G. A. Storch, ‘Enhanced virome sequencing using targeted sequence capture’, Genome Res., vol. 25, no. 12, pp. 1910–1920, Dec. 2015, doi: 10.1101/gr.191049.115.

[101] J. S. Kalpoe, A. C. M. Kroes, S. Verkerk, E. C. J. Claas, R. M. Y. Barge, and M. F. C. Beersma, ‘Clinical relevance of quantitative varicella-zoster virus (VZV) DNA detection in plasma after stem cell transplantation’, Bone Marrow Transplant, vol. 38, no. 1, pp. 41–46, Jul. 2006, doi: 10.1038/sj.bmt.1705397.

[102] L.-A. Le, T.-P. Nguyen-Hoang, V.-P. Huynh, T.-H. Nguyen, T.-V. Nguyen, and T.-D. Ho-Huynh, ‘Microbiome dataset analysis from a shrimp pond in Ninh Thuan, Vietnam using shotgun metagenomics’, Data in Brief, vol. 31, p. 105731, Aug. 2020, doi: 10.1016/j.dib.2020.105731.

[103] E. C. Carbo et al., ‘Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics’, Journal of Clinical Virology, vol. 130, p. 104566, Sep. 2020, doi: 10.1016/j.jcv.2020.104566.

[104] T. Doan et al., ‘Metagenomic DNA Sequencing for the Diagnosis of Intraocular Infections’, Ophthalmology, vol. 124, no. 8, pp. 1247–1248, Aug. 2017, doi: 10.1016/j.ophtha.2017.03.045.

[105] J. C. Haston et al., ‘Prospective Cohort Study of Next-Generation Sequencing as a Diagnostic Modality for Unexplained Encephalitis in Children’, Journal of the Pediatric Infectious Diseases Society, vol. 9, no. 3, pp. 326–333, Jul. 2020, doi: 10.1093/jpids/piz032.

[106] H. Jerome et al., ‘Metagenomic next-generation sequencing aids the diagnosis of viral infections in febrile returning travellers’, Journal of Infection, vol. 79, no. 4, pp. 383–388, Oct. 2019, doi: 10.1016/j.jinf.2019.08.003.

[107] H. Johansson, D. Bzhalava, J. Ekström, E. Hultin, J. Dillner, and O. Forslund, ‘Metagenomic sequencing of “HPV-negative” condylomas detects novel putative HPV types’, Virology, vol. 440, no. 1, pp. 1–7, May 2013, doi: 10.1016/j.virol.2013.01.023.

[108] D. W. Lewandowska et al., ‘Unbiased metagenomic sequencing complements specific routine diagnostic methods and increases chances to detect rare viral strains’, Diagnostic Microbiology and Infectious Disease, vol. 83, no. 2, pp. 133–138, Oct. 2015, doi: 10.1016/j.diagmicrobio.2015.06.017.

[109] N. Madi, W. Al-Nakib, A. S. Mustafa, and N. Habibi, ‘Metagenomic analysis of viral diversity in respiratory samples from patients with respiratory tract infections in Kuwait’, J Med Virol, vol. 90, no. 3, pp. 412–420, Mar. 2018, doi: 10.1002/jmv.24984.

[110] N. E. Moore et al., ‘Metagenomic Analysis of Viruses in Feces from Unsolved Outbreaks of Gastroenteritis in Humans’, J Clin Microbiol, vol. 53, no. 1, pp. 15–21, Jan. 2015, doi: 10.1128/JCM.02029-14.

[111] U. Pérez-Sautu et al., ‘Target-independent high-throughput sequencing methods provide evidence that already known human viral pathogens play a main role in respiratory infections with unexplained etiology’, Emerging Microbes & Infections, vol. 8, no. 1, pp. 1054–1065, Jan. 2019, doi: 10.1080/22221751.2019.1640587.

[112] A. Ramesh et al., ‘Metagenomic next-generation sequencing of samples from pediatric febrile illness in Tororo, Uganda’, PLoS ONE, vol. 14, no. 6, p. e0218318, Jun. 2019, doi: 10.1371/journal.pone.0218318.

[113] A. Reyes et al., ‘Viral metagenomic sequencing in a cohort of international travellers returning with febrile illness’, Journal of Clinical Virology, vol. 143, p. 104940, Oct. 2021, doi: 10.1016/j.jcv.2021.104940.

[114] S. Saha et al., ‘Unbiased Metagenomic Sequencing for Pediatric Meningitis in Bangladesh Reveals Neuroinvasive Chikungunya Virus Outbreak and Other Unrealized Pathogens’, mBio, vol. 10, no. 6, pp. e02877-19, /mbio/10/6/mBio.02877-19.atom, Dec. 2019, doi: 10.1128/mBio.02877-19.

[115] R. Schlaberg et al., ‘Viral Pathogen Detection by Metagenomics and Pan-Viral Group Polymerase Chain Reaction in Children With Pneumonia Lacking Identifiable Etiology’, The Journal of Infectious Diseases, vol. 215, no. 9, pp. 1407–1415, May 2017, doi: 10.1093/infdis/jix148.

[116] S. L. Smits et al., ‘Novel Cyclovirus in Human Cerebrospinal Fluid, Malawi, 2010–2011’, Emerg. Infect. Dis., vol. 19, no. 9, Sep. 2013, doi: 10.3201/eid1909.130404.

[117] S. L. Smits et al., ‘New Viruses in Idiopathic Human Diarrhea Cases, the Netherlands’, Emerg. Infect. Dis., vol. 20, no. 7, Jul. 2014, doi: 10.3201/eid2007.140190.

[118] B. Taboada et al., ‘Is There Still Room for Novel Viral Pathogens in Pediatric Respiratory Tract Infections?’, PLoS ONE, vol. 9, no. 11, p. e113570, Nov. 2014, doi: 10.1371/journal.pone.0113570.

[119] N. Thi Kha Tu et al., ‘The Virome of Acute Respiratory Diseases in Individuals at Risk of Zoonotic Infections’, Viruses, vol. 12, no. 9, p. 960, Aug. 2020, doi: 10.3390/v12090960.

[120] P. Turner et al., ‘The aetiologies of central nervous system infections in hospitalised Cambodian children’, BMC Infect Dis, vol. 17, no. 1, p. 806, Dec. 2017, doi: 10.1186/s12879-017-2915-6.

[121] H. Wang et al., ‘Clinical diagnostic application of metagenomic next-generation sequencing in children with severe nonresponding pneumonia’, PLoS ONE, vol. 15, no. 6, p. e0232610, Jun. 2020, doi: 10.1371/journal.pone.0232610.

[122] L. Xu et al., ‘Characterization of the nasopharyngeal viral microbiome from children with community-acquired pneumonia but negative for Luminex xTAG respiratory viral panel assay detection’, J Med Virol, vol. 89, no. 12, pp. 2098–2107, Dec. 2017, doi: 10.1002/jmv.24895.

[123] N. L. Yozwiak, P. Skewes-Cox, M. D. Stenglein, A. Balmaseda, E. Harris, and J. L. DeRisi, ‘Virus Identification in Unknown Tropical Febrile Illness Cases Using Deep Sequencing’, PLoS Negl Trop Dis, vol. 6, no. 2, p. e1485, Feb. 2012, doi: 10.1371/journal.pntd.0001485.

[124] Y. Zhou et al., ‘Metagenomics Study of Viral Pathogens in Undiagnosed Respiratory Specimens and Identification of Human Enteroviruses at a Thailand Hospital’, The American Journal of Tropical Medicine and Hygiene, vol. 95, no. 3, pp. 663–669, Sep. 2016, doi: 10.4269/ajtmh.16-0062.

[125] X. Zou et al., ‘Simultaneous virus identification and characterization of severe unexplained pneumonia cases using a metagenomics sequencing technique’, Sci. China Life Sci., vol. 60, no. 3, pp. 279–286, Mar. 2017, doi: 10.1007/s11427-016-0244-8.

[126] S. I. Sardi et al., ‘Coinfections of Zika and Chikungunya Viruses in Bahia, Brazil, Identified by Metagenomic Next-Generation Sequencing’, J Clin Microbiol, vol. 54, no. 9, pp. 2348–2353, Sep. 2016, doi: 10.1128/JCM.00877-16.

[127] ‘JASP’, Accessed: Mar. 23, 2021. [Online]. Available: https://jasp-stats.org

[128] C. Rodriguez et al., ‘Fatal Encephalitis Caused by Cristoli Virus, an Emerging Orthobunyavirus, France’, Emerg Infect Dis, vol. 26, no. 6, pp. 1287–1290, Jun. 2020, doi: 10.3201/eid2606.191431.

[129] C. Rodriguez et al., ‘Viral genomic, metagenomic and human transcriptomic characterization and prediction of the clinical forms of COVID-19’, PLoS Pathog, vol. 17, no. 3, p. e1009416, Mar. 2021, doi: 10.1371/journal.ppat.1009416.

[130] J. R. Brown et al., ‘Astrovirus VA1/HMO-C: An Increasingly Recognized Neurotropic Pathogen in Immunocompromised Patients’, Clinical Infectious Diseases, vol. 60, no. 6, pp. 881–888, Mar. 2015, doi: 10.1093/cid/ciu940.

[131] M. Alawi et al., ‘DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples’, Sci Rep, vol. 9, no. 1, p. 16841, Nov. 2019, doi: 10.1038/s41598-019-52881-4.

[132] M. Christopeit et al., ‘Suspected encephalitis with Candida tropicalis and Fusarium detected by unbiased RNA sequencing’, Ann Hematol, vol. 95, no. 11, pp. 1919–1921, Nov. 2016, doi: 10.1007/s00277-016-2770-3.

[133] F. X. López-Labrador et al., ‘Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure’, Journal of Clinical Virology, vol. 134, p. 104691, Jan. 2021, doi: 10.1016/j.jcv.2020.104691.

[134] C. P. Oechslin et al., ‘Limited Correlation of Shotgun Metagenomics Following Host Depletion and Routine Diagnostics for Viruses and Bacteria in Low Concentrated Surrogate and Clinical Samples’, Front. Cell. Infect. Microbiol., vol. 8, p. 375, Oct. 2018, doi: 10.3389/fcimb.2018.00375.

[135] A. L. van Rijn et al., ‘The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease’, PLoS ONE, vol. 14, no. 10, p. e0223952, Oct. 2019, doi: 10.1371/journal.pone.0223952.

[136] S. van Boheemen et al., ‘Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients’, The Journal of Molecular Diagnostics, vol. 22, no. 2, pp. 196–207, Feb. 2020, doi: 10.1016/j.jmoldx.2019.10.007.

[137] J. R. Brown, T. Bharucha, and J. Breuer, ‘Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases’, Journal of Infection, vol. 76, no. 3, pp. 225–240, Mar. 2018, doi: 10.1016/j.jinf.2017.12.014.

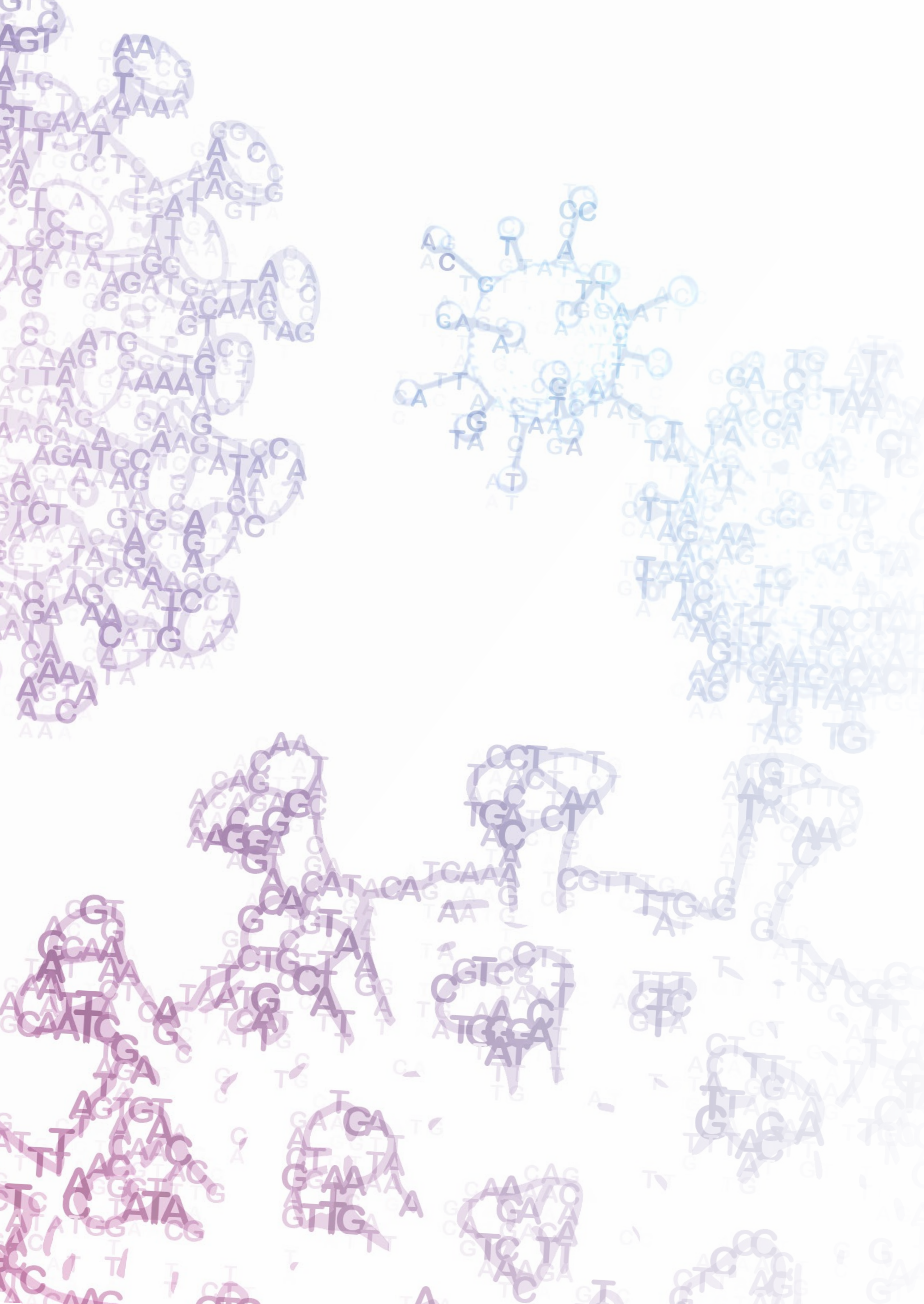
[138] T. Briese et al., ‘Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis’, mBio, vol. 6, no. 5, pp. e01491-15, Sep. 2015, doi: 10.1128/mBio.01491-15.

[139] M. Asplund et al., ‘Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries’, Clinical Microbiology and Infection, vol. 25, no. 10, pp. 1277–1285, Oct. 2019, doi: 10.1016/j.cmi.2019.04.028.

[140] J. C. Wooley and Y. Ye, ‘Metagenomics: Facts and Artifacts, and Computational Challenges’, J. Comput. Sci. Technol., vol. 25, no. 1, pp. 71–81, Jan. 2010, doi: 10.1007/s11390-010-9306-4.

[141] A. L. Greninger, ‘The challenge of diagnostic metagenomics’, Expert Review of Molecular Diagnostics, vol. 18, no. 7, pp. 605–615, Jul. 2018, doi: 10.1080/14737159.2018.1487292.

[142] T. Li et al., ‘Metagenomic Next-Generation Sequencing of the 2014 Ebola Virus Disease Outbreak in the Democratic Republic of the Congo’, J Clin Microbiol, vol. 57, no. 9, pp. e00827-19, Sep. 2019, doi: 10.1128/JCM.00827-19.



Chapter 2 Viral metagenomic sequencing in the diagnosis of meningo-encephalitis: a review of technical advances and diagnostic yield

Ellen C. Carbo[1], Ivar Blankenspoor[1], Jelle J. Goeman[2], Aloys C.M. Kroes[1],
Eric C.J. Claas[1], Jutte J.C. de Vries[1]

1. Clinical Microbiological Laboratory, department of Medical Microbiology, Leiden University Medical Center, Leiden, the Netherlands
2. Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherland

Published Invited article in
Expert Review of Molecular Diagnostics

Abstract

Introduction: Meningoencephalitis patients are often severely impaired and benefit from early etiological diagnosis though many cases remain without identified cause. Metagenomics as pathogen agnostic approach can result in additional etiological findings, however the exact diagnostic yield when used as a secondary test remains unknown.

Areas covered: This review aims to highlight recent advances with regard to wet and dry lab methodologies of metagenomic testing and technical milestones that have been achieved. A selection of procedures currently applied in accredited diagnostic laboratories is described in more detail to illustrate best practices. Furthermore, a meta-analysis was performed to assess the additional diagnostic yield utilizing metagenomic sequencing in meningoencephalitis patients. Finally, the remaining challenges for successful widespread implementation of metagenomic sequencing for the diagnosis of meningoencephalitis are addressed in a future perspective.

Expert opinion: The last decade has shown major advances in technical possibilities for using mNGS in diagnostic settings including cloud-based analysis. An additional advance may be the current established infrastructure of platforms for bioinformatic analysis of SARS-CoV-2, which may assist to pave the way for global use of clinical metagenomics.

Highlights

- **The additional diagnostic yield of metagenomic sequencing for pathogen detection when used as a secondary test after conventional testing is 5-20% and is dependent on the endemic pathogens in combination with the available diagnostic facilities.**
- **Best metagenomic practices for wet lab procedures include virus enrichment by means of depletion of ribosomal RNA and probes capturing vertebrate viruses**
- **Best practices for bioinformatic analysis of metagenomic data include algorithms to minimize false positive findings and to assist interpretation, for example using post-probability scores**
- **Future comparisons of metagenomic protocols with regard to sensitivity, specificity, feasibility in terms of laborious workflows, and turn-around time are needed. Recently, the ENNGS has initiated the sharing and comparison of viral metagenomic protocols in the project METASHARE.**

Introduction

Meningoencephalitis is a severe inflammation of the brain tissue and meninges, with an overall mortality of 30% and long-term residual sequelae in the majority of the patients that survive [1]. All age groups can be affected and immunocompromised patients are at higher risk of infection with unexpected and novel viral pathogens [2]. Disease outcome improves with a proper and timely diagnosis and correct identification of disease etiology [3]. Strikingly, more than 30% of cases remain without identified etiologic agent [4]. A wide range of causative agents can be involved, and besides host immune status, the etiology is also dependent on geographical location, as exemplified by tick-born encephalitis, Toscana virus encephalitis and Japanese encephalitis. The clinical severity of the disease in combination with frequently negative routine qPCR panel results and wide range of causative agents makes this type of patients attractive candidates for metagenomic next generation sequencing (mNGS), as mNGS can detect all pathogens, including rare and novel pathogens not included in conventional testing.

Over the past decade, an increasing number of studies has been published on metagenomic sequencing in cases of meningoencephalitis of unknown cause, using mainly cerebrospinal fluid and sporadically brain tissue (Figure 1). Most reports are on individual clinical cases with either novel viruses or known viruses not previously associated with a specific clinical syndrome. A growing but still modest number of prospective evaluations have been reported on the application of metagenomic sequencing in routine diagnostic settings. This review aims to summarize findings with regard to the diagnostic yield of viral metagenomics in meningoencephalitis patients with negative conventional test results, to highlight milestones and share technical details of a selection of viral metagenomic methods that have been implemented in routine diagnostic laboratories as examples of best practice. Finally, remaining technical challenges for implementation of viral mNGS are addressed.

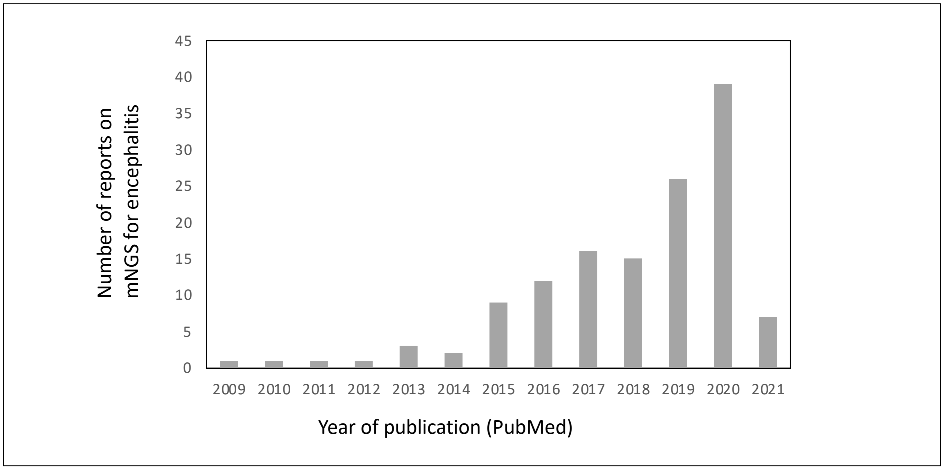


Figure 1. The increasing number of reports in literature (PubMed) on metagenomic sequencing in meningoencephalitis of unknown cause.

Date of access 2021 April 20, query [encephalitis AND metagenomic OR metagenomics]

Additional diagnostic yield of mNGS in cases of meningoencephalitis

Appropriate management of patients with meningoencephalitis is dependent on timely identification of the etiological agent. The distinction between infection and inflammatory causes is of importance since inflammatory meningoencephalitis is typically treated with anti-inflammatory drugs, which can have counter-effective results in patients with active virus replication. Viral mNGS provides a broad and untargeted approach to identify all pathogenic viruses from the differential diagnosis and beyond in one single test. Metagenomics for pathogen detection is currently used in a growing number of laboratories as secondary test for difficult to diagnose cases with negative conventional diagnostic test results. To analyze the added value of mNGS in the clinical setting the additional diagnostic yield in patients with meningoencephalitis as reported in literature was reviewed, and a meta-analysis was performed. The additional diagnostic yield was defined as the proportion of extra etiological agents identified by conducting mNGS as a secondary test compared to conventional testing. The included papers consisted of studies applying metagenomic next-generation sequencing (mNGS) in meningoencephalitis patient cohorts, suspected of an infectious aetiology and negative by conventional testing. Only studies using mNGS for pathogen detection were included using the search strategy, search terms, and exclusion criteria described in Supplementary Table 1.

Additional relevant studies were selected by screening the reference lists of the included studies. Two authors (IB, ECC) independently reviewed and extracted data from the included manuscripts [5]–[14]. From the cohort studies investigating mNGS for pathogen detection, the additional diagnostic yield was determined by analysis of the proportion (%) of meningoencephalitis patients with additional findings by mNGS. Next, diagnostic yield data was analysed using JASP statistical software [15] based on the R package Metafor [16], for a study with an estimator of 0 [10] the 95% confidence interval was calculated with Clopper-Pearson exact. A restricted maximum likelihood meta-analysis was performed to summarize the results of all the studies included, followed by subgroup analysis based on patient origin. The forest plot of the additional metagenomic yield is shown in Figure 2.

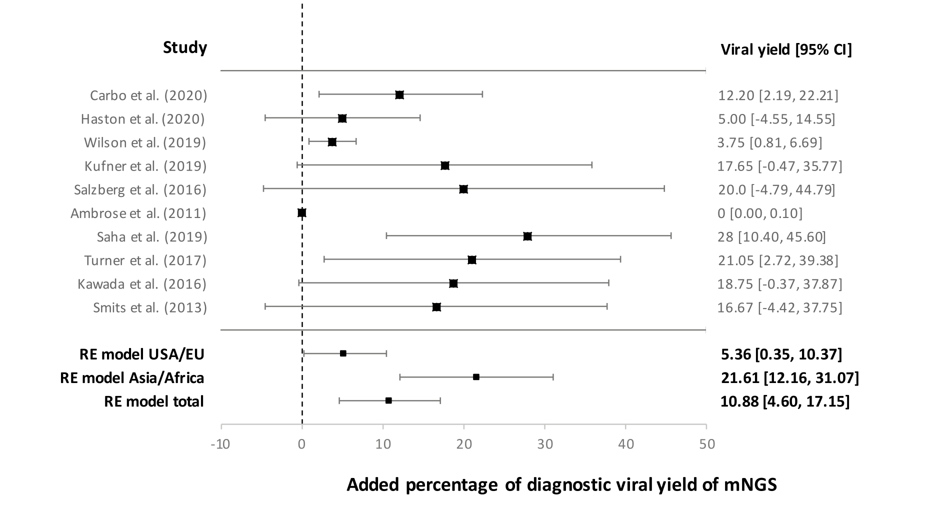


Figure 2. Forest plot of the diagnostic yields of the included cohort studies using mNGS for pathogen detection in cases of meningoencephalitis of unknown cause.

Additional viral yield is shown as percentage per study including the 95% confidence interval (CI). Viral yield is defined as the percentage of additional diagnoses that are being made due to the utilization of metagenomic NGS, compared to only using conventional tests. Total RE model and RE models specified on patient origin are included.

The studies included show significant heterogeneity in design, geographic location and causative agents. Therefore, a restricted maximum likelihood model was used, which lead to an overall additional viral diagnostic yield by mNGS of 10.88% (95% CI 4.6-17.15). The viral diagnostic yield in moderate climate zones (USA, EU) was 5.36%

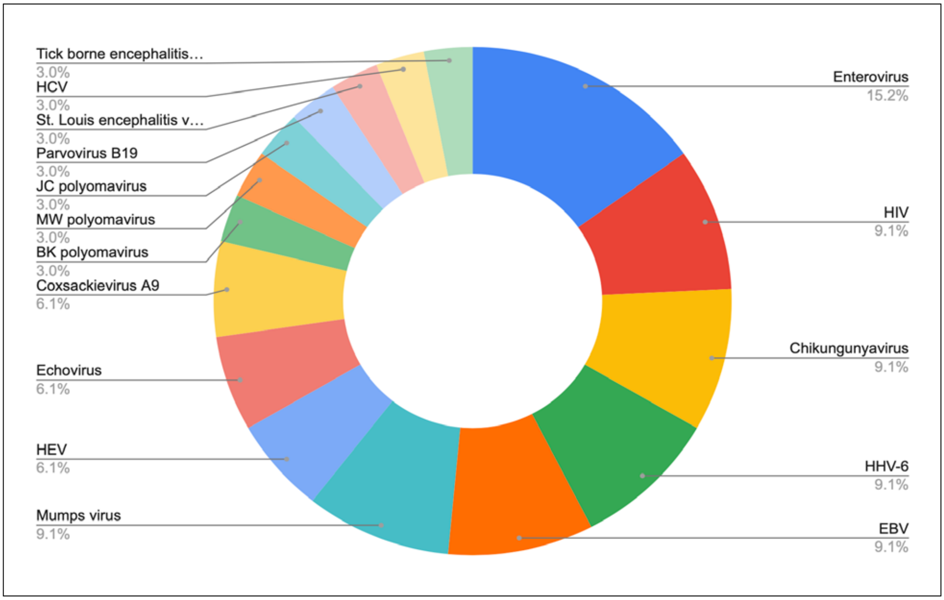


Figure 3. Pie chart of all pathogenic viruses detected by mNGS in cerebrospinal fluid in the reports included in the current meta-analysis.

(95% CI 0.35-10.37), generally lower than (sub)tropical climate zones: 21.61% (95% CI 12.16-31.07). Additional pathogenic virus yield in (sub)tropical climate zones included mosquito born viral disease viruses such as CHIKV [11]. An additional factor for a higher yield was the detection of pathogens part of a vaccination program present in western but absent in non-western countries, like mumps [11]. An overview of the additional viruses detected is depicted in Figure 3. These viruses can be categorized as known viruses that were not included in the conventional testing panel since they were rarely or previously not detected in meningoencephalitis. It must be noted that no novel viruses were published as part of these cohort studies. Detection of novel viruses has been mainly described in case reports, from which no data on proportionality could be deducted for meta-analysis. The above described rate of additional yield of 5.36% in moderate climate zones was based on studies that included all idiopathic meningoencephalitis patients based on clinical, biological and radiological data. However, when selecting cases with potentially high risk of viral infection the yield was higher: 12.2% additional mNGS yield in hematological adult and pediatric patients with meningoencephalitis [5]. A Swiss study with a 17,65% yield reported that in the majority of patients (>67%) an infectious disease specialist was consulted to select patients with higher suspicion for viral etiology [8]. It must be noted that the yield will increases when bacterial and other pathogens are taken into account.

With these additional diagnostic yield percentages and an annual incidence of over 500,000 meningoencephalitis cases worldwide [3,17], widespread implementation of mNGS diagnostics is expected to lead to a substantial increase in the number of identified etiologies and correctly diagnosed cases.

Technical advances in the wet lab: viral enrichment

A diversity of mNGS library preparation and sample pre-treatment methods is in use for diagnosing patients with meningoencephalitis. In contrast to viral meningitis, viral infection of the brain usually does not result in high virus concentration in the cerebrospinal fluid (CSF). Most diagnostic laboratories receive CSF for virus diagnostics, while the higher viral loads in brain biopsies are better suited for metagenomic sequencing assays. A recent benchmarking study [18] underscores that sensitivity remains a challenge in the presence of abundant background host sequences. Table 1 provides an overview of a selection of technical wet and dry lab methodologies currently implemented in diagnostic settings or extensively prospectively validated for clinical diagnostic use. Viral enrichment before extraction of nucleic acids, by centrifugation and filtration and in some protocols using DNase treatment can be beneficial but is not readily automatized and furthermore has not consistently been reported as effective [19,20,21,22]. Enrichment of RNA virus sequences is commonly performed either by removal of ribosomal RNA depletion or enrichment by poly A tail binding of mRNA. The mRNA of eukaryotic viruses is usually poly A tailed, in addition to the genome itself in some viruses (e.g. picornaviruses) [19,23,24]. Some viruses initiate translation in the absence of poly A tail by using functional analogues (e.g. hepatitis C viruses, rotaviruses) and viruses that are in a non-replicative phase may be missed when using this type of selection method [19]. After nucleic acid extraction, reverse transcription and library preparation is commonly performed using separate library preps for RNA and DNA viruses, though a one-tube protocol can be used as cost-effective alternative [21,22]. Enrichment after library preparation using capture probes specific for all known vertebrate viruses resulted in a significant improvement in sensitivity and 100-10,000 fold increase in virus read counts [5,19,25,26]. Despite these enrichment techniques, sensitivity remains an issue to be addressed in the validation phase when implementing viral metagenomic sequencing in routine diagnostic settings as shown in recent benchmarking studies [18,27]. Some methods have resulted in sensitivities comparable to PCR, but not all protocols have been proven equally efficient for detecting DNA viruses in various types of patient samples. A validation study should include the different sample types selected for application in combination with the selected wet and dry lab protocol [18].

Detection of DNA viruses in brain biopsies tends to be more sensitive due to higher abundance of virus material, but is not often performed as it is an invasive method [2]. Sequencing of tissue biopsies can be hindered by large amounts of host sequences as compared to analysis of cerebrospinal fluids [28]. DNA derived from Formalin-Fixed Paraffin-Embedded (FFPE) material can be impaired due to the required specimen processing workflow [29] leading to sequence artifacts [30]. It is advised to follow the evidence-based practices for e.g. formalin fixation time, storage condition and extraction methods [31]. Due to sequence artifacts the use of molecular tags or unique molecular identifiers should be considered. In this way, each molecule prior to library prepping is labeled and can be analyzed by additional bioinformatic tools [32,33].

Advances in bioinformatic analysis and cloud-based analysis

The performance of metagenomic methods is heavily dependent on accurate bio-informatic analysis, and both the classification algorithms as well as the databases are crucial determinants of the overall performance of available pipelines. A wide range of metagenomic pipelines and taxonomic classifiers have been developed, often for the purpose of biodiversity studies analyzing the composition of the microbiome including the virome in different samples and cohorts [28]. In contrast, when applying mNGS for patient diagnostics, potential false-negative and false-positive bioinformatic classification results can have significant consequences for patient care. Reports on specific bioinformatic tools for metagenomic analysis for virus diagnostics typically describe algorithms and validations of single pipelines developed and used by the authors themselves, stressing the need for high quality validation and comparison studies [28]. The development of guidelines and recommendations on mNGS bioinformatic analysis methods and reporting will assist the implementation of mNGS in diagnostic laboratories, ensuring the validity of results and thus optimizing patient management [34]. A recent benchmark of bioinformatic tools and pipelines conducted by the ESCV Network on NGS [18], where datasets from clinical samples including CSF and brain biopsies from patients with viral meningoencephalitis were analyzed, showed that virus infections with Ct-values of ≤ 28 were challenging for most tools and pipelines. The tools/pipelines with the combination of highest sensitivity and selectivity were metaMix [35], Centrifuge [36] and VirMet [37]. An extra correction for increasing the specificity can be made with additional tools for deduction of contamination or the 'kitome' [38,39] or align the sequence reads to a potential given specie, like in GenomeDetective[40], to see whether reads are evenly distributed to avoid artifacts.

Processing of mNGS data can be done via command line tools compiled by bioinformaticians, or by user-friendly interfaces containing tools and pipelines. Potent computer hardware can be situated locally in the format of high-performance computing (HPC) cluster, or remotely via cloud computing. Cloud-based platforms usually have web front-end interfaces which facilitate direct uploading of the raw files from sequencing instruments and direct downloading of the final output analyses from the server [28]. Galaxy [41] and BlueBee [42] are examples of web-based platforms with user-friendly interfaces for hosting in-house tools and pipelines. Recently, several web-based, user-friendly, and complete pipelines for viral metagenomic analyses have become available, including DNASTAR [43], Genome Detective [40], One Codex [44], Taxonomer [45], and IDbyDNA [46], the latter including library preparation and sequencing. The availability of these complete analyses as a service package, enables laboratories with no access to a HPC cluster or with limited bioinformatic knowledge, to analyze mNGS datasets, which can be considered a milestone. These service packages should be validated locally to assure accurate identification and classification of potential target viruses, and to analyze the limit of detection and variation. Common practice is in silico validation using a selection of viral RNA and DNA sequences, single stranded and double stranded, followed by a validation of the entire workflow using well-characterized patient samples. Precision, recall, and the F1 score as a combination of these, are the measures applicable when using patient samples since in practice it is impossible to subject every negative metagenomic finding to PCR. It is expected that implementation of these software packages will be beneficial for broader implementation of metagenomic sequencing, especially in the new in vitro diagnostic regulated (IVDR) era.

Remaining challenges for implementation

Several challenges remain and hamper the widespread implementation of viral mNGS for meningoencephalitis cases in routine diagnostic laboratories. These challenges can be found in both wet and dry lab procedures. There is no optimal and highly sensitive procedure for library preparation for viral metagenomic detection yet. Lack of standardization has impact on the ability of labs to select a procedure that is easily introduced into the routine diagnostic testing process with a time to result within a clinically relevant timeframe. Clearly, there is a need for future comparisons of wet lab protocols with regard to sensitivity, specificity, feasibility in terms of laborious workflows, and turnaround time. Recently, the ENNGS has initiated the sharing and comparison of viral metagenomic protocols in the project METASHARE.

With regard to the remaining dry lab challenges, bioinformatic analysis software and bioinformaticians have not typically been part of the infrastructure of the diagnostic microbiological lab in the past decades and cloud-bases analyses have only recently been introduced for metagenomics for pathogen detection. The validation procedure for bioinformatic analysis has yet not been standardized and the IVDR may stimulate manufacturers to implement and share further standardization of the process of validation of the pipelines and software updates. The IVDR may prove to be useful for mNGS (end-)users in this aspect: it requires that users will have access to information on the validation process of the pipeline and updated versions. Agreements will need to be in place to cover details on the storage and access of sequence data, results and logging. Sharing databases and pipelines for comparison will support laboratories during their mNGS protocol selection process. User-friendly access to databases and metagenomic pipelines provided with information on their sensitivity, specificity and clinical usage in a user-friendly way will be an impactful factor for the widespread implementation of viral metagenomics in diagnostic laboratories. Furthermore, in the coming years, some efforts are expected on the interpretation of the reports, possibly provided with post-probability scores in a user-friendly format, as the consultant benefits most from binary results that can guide a clinical course of action. It is anticipated that further development of interpretation algorithms may be beneficial here.

Expert commentary

A pro-con debate on viral metagenomics as a frontline approach was organized at the last Molecular Virology Workshop by the Pan American Society for Clinical Virology. It was an effective platform to contrast views on the challenges to the integration of viral metagenomics as a frontline diagnostic approach. Approximately half of the participants estimated that within the next 10 years clinical metagenomics would be implemented as frontline diagnostic approach, at least for a significant part of clinical cases. It remains to be seen whether this time-frame is sufficient to gather all the evidence for clinical utility in different patient populations and, importantly, to achieve cost effectiveness. Although sequence costs are rapidly decreasing, the manual workload and turn-around time are currently the main drawbacks and both have to be reduced to compete with rapid syndromic PCR panel testing with increasing numbers of target pathogens.

Whereas one decade ago the predominant question raised was whether metagenomic sequencing could be integrated in diagnostic laboratories for use in clinical care at all, now clinical metagenomics is being implemented in an increasing number of specialized diagnostic laboratories within the scope of their accreditation. The time-frame for widespread implementation is currently largely dependent on technical development: index hopping, 'kitome' sequences [47], low sensitivity and inaccurate quantification of target viruses are technical challenges that are expected to be resolved within the next decade. Algorithms are being developed to correct for factors interfering with pathogen detection and quantification such as background reads and contaminants [38,39]. Importantly, the momentum of the COVID-19 pandemic is in place and embodies the ultimate example of the value of metagenomic surveillance for detection of emerging and novel viruses. Additionally, the infrastructure for SARS-CoV-2 analysis and variant detection is being improved and extended. In a recent WHO meeting global accessibility to pipelines for SARS-CoV-2 variant analysis was discussed. Today's established infrastructure of platforms for bioinformatic analysis of SARS-CoV-2 can technically in the future also be used for harboring metagenomic pipelines and may pave the way for global use of clinical metagenomics.

The last decade has shown major advances in technical possibilities for using mNGS in diagnostic settings. A growing number of commercial parties is interested in providing cloud-based services for metagenomic bioinformatic analyses and seems to be preparing for IVD and FDA regulations. Hopefully, the next decade will be characterized by progress in technology and clinical implementation, perhaps resulting in one of the ultimate applications of the implementation of viral metagenomics: patient bed-side virus discovery.

Acknowledgements

We thank Christophe Rodriquez (CHU Henri Mondor, Créteil, France), Verena Kufner and Michael Huber (both from the Institute of Medical Virology, Zurich),), Julianne Brown and Judith Breuer (Great Ormond Street Hospital for Children NHS Foundation Trust, London, United Kingdom), Sofia Morfopoulou (University College London, London, United Kingdom), and Nicole Fisher (UKE Institute for Medical Microbiology, Virology and Hygiene, Germany) for providing information on their viral mNGS lab protocol.

Conflicts of interest

None

Funding

None

Table 1. Technical aspects of a selection of wet and dry lab methodologies currently in use in settings where viral metagenomic sequencing on cerebrospinal fluid has been implemented or prospectively validated for clinical diagnostics and reported.

Year, author	Library prep	Enrichment	Internal controls
2019 Wilson et al. [7]	Separate RNA and DNA libraries: Nextera XT DNA library prep kit (Illumina), sequenced pooled	DNase treatment of extract for RNA libraries, removal of CpG-methylated host DNA for DNA libraries (NEB Microbiome Kit)	T1 and MS2 Escherichia coli bacteriophages, for resp. DNA and RNA viruses
2020 2021 Rodriguez et al. [49,50,51]	Separate RNA and DNA library Nextera XT library and Stranded Total RNA pooled in the same run	No enrichment	No internal control. DNA and human RNA are used to verify quality of extraction.
2019, Kufner et al.[8]	Separate library prep for RNA and DNA viruses: NexteraXT DNA library preparation kit (Illumina) RNA and DNA sequenced separately	Pre-extraction: Low-speed centrifugation, filtration (0.45 um)	MS2 Escherichia coli bacteriophage for RNA viruses, T1 E. coli bacteriophage for DNA viruses in establishment
Brown et al. 2016, 2017 [18,53-55]	RNA: ROCHE Hyperprep kit, and the riboerase depletion kit for tissues. DNA: NEBNext ULTRA II FS DNA Library Prep Kit for Illumina	RNA-seq: ribodepletion (ROCHE Hyperprep RiboErase) DNA enrichment uses the NEBNext Microbiome Enrichment Kit	RNA-seq: Enterobacteria phage MS2 DNA-seq: not applicable (explanation: were using Lambda DNA up until February but there were issues and hence it was dropped)
2020, Carbo et al. [5]	Single library prep for RNA and DNA viruses: NEBNext Ultra II DNA library prep with adaptation	Post-library prep: Capture probes targeting known vertebrate viruses	Equine arteritis virus (EAV) for RNA and phocid herpesvirus-1 (PhHV-1) for DNA viruses
Alawi et al., [57] Christopeit et al., [58]	RNASeq: SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Takara Bio Europe)	No enrichment	No internal control

NEB; New England Biolabs. NA; not analyzed

Controls	Sequencer	Pipeline	Threshold determination for reporting pathogens	Limit of detection
No template control: elution buffer, positive control: RNA and DNA pathogen mixture (7 organisms)	Illumina HiSeq, 5-10 million reads	Modified SURPI+ in-house pipeline [48]	Receiver-operator curve (ROC) analysis based on clinical samples with established pathogens	DNA virus (CMV): 14 copies/ml RNA virus (HIV) 313 copies/ml
Environmental control: sterile water Positive control: ZymoBiomics (Zymo Research)	NextSeq 500 (2*150 bp) >20 millions (30/40millions mean)	MetaMIC in-house pipeline [49]	Above background: environmental control	<3 log copies/ml for DNA and RNA viruses. Bacteria, fungi, parasites quite similar to culture or PCR
No template: PBS (included in nucleic acid extraction)	Illumina MiSeq average 7 million reads/ sample (1*150 bp)	VirMet in-house pipeline [52]	>=3 reads distributed over the genome with high coverage score and not detected >100 times in NC or other samples (carry-over)	NA
RNA-seq: Total Brain RNA spiked in with feline calcivirus DNA-seq: Human genomic DNA spiked in with cowpox DNA	Illumina NextSeq500 (2*81 bp) 100 million reads	metaMix in-house pipeline [35]	≥3 regions, >10 reads [13,56] posterior probability & bayes factor	Similar to PCR in CSF for both RNA and DNA viruses. In tissue, for DNA viruses approx. 100-fold reduced
No template control: elution buffer	Illumina NovaSeq 6000, 10 million reads (2*150 bp)	Genome Detective commercial pipeline [40]	ROC analysis based on clinical samples with established pathogens, Coverage >=3 distant genome locations	RNA virus 10-60 copies/ml DNA virus 100-1000 copies/ml [28]
No template control	Illumina NextSeq 150PE 5-10 Mio reads	DAMIAN Pipeline (Alawi et al, [57]	Contig assembly approach; contigs > 400bp are reported See Alawi et al., [57]	Assembly is independent of host reads/background reads; minimum of 250 reads necessary for contig assembly Fischer et al., [59]

References

[1] A. Mailles et al., ‘Long-term Outcome of Patients Presenting With Acute Infectious Encephalitis of Various Causes in France’, Clin. Infect. Dis., vol. 54, no. 10, pp. 1455–1464, May 2012, doi: 10.1093/cid/cis226.

[2] J. R. Brown, T. Bharucha, and J. Breuer, ‘Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases’, J. Infect., vol. 76, no. 3, pp. 225–240, Mar. 2018, doi: 10.1016/j.jinf.2017.12.014.

[3] J. Granerod and N. S. Crowcroft, ‘The epidemiology of acute encephalitis’, Neuropsychol. Rehabil., vol. 17, no. 4–5, pp. 406–428, Aug. 2007, doi: 10.1080/09602010600989620.

[4] C. A. Glaser et al., ‘Beyond Viruses: Clinical Profiles and Etiologies Associated with Encephalitis’, Clin. Infect. Dis., vol. 43, no. 12, pp. 1565–1577, Dec. 2006, doi: 10.1086/509330.

[5] E. C. Carbo et al., ‘Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics’, J. Clin. Virol., p. 104566, Jul. 2020, doi: 10.1016/j.jcv.2020.104566.

[6] J. C. Haston et al., ‘Prospective Cohort Study of Next-Generation Sequencing as a Diagnostic Modality for Unexplained Encephalitis in Children’, J. Pediatr. Infect. Dis. Soc., vol. 9, no. 3, pp. 326–333, Jul. 2020, doi: 10.1093/jpids/piz032.

[7] M. R. Wilson et al., ‘Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis’, N. Engl. J. Med., vol. 380, no. 24, pp. 2327–2340, Jun. 2019, doi: 10.1056/NEJMoa1803396.

[8] Kufner et al., ‘Two Years of Viral Metagenomics in a Tertiary Diagnostics Unit: Evaluation of the First 105 Cases’, Genes, vol. 10, no. 9, p. 661, Aug. 2019, doi: 10.3390/genes10090661.
* Overview of results of using metagenomics in a diagnostic setting

[9] S. L. Salzberg et al., ‘Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system’, Neurol. -Neuroimmunol. Neuroinflammation, vol. 3, no. 4, p. e251, Aug. 2016, doi: 10.1212/NXI.0000000000000251.

[10] H. E. Ambrose et al., ‘Diagnostic Strategy Used To Establish Etiologies of Encephalitis in a Prospective Cohort of Patients in England’, J. Clin. Microbiol., vol. 49, no. 10, pp. 3576–3583, Oct. 2011, doi: 10.1128/JCM.00862-11.

[11] S. Saha et al., ‘Unbiased Metagenomic Sequencing for Pediatric Meningitis in Bangladesh Reveals Neuroinvasive Chikungunya Virus Outbreak and Other Unrealized Pathogens’, mBio, vol. 10, no. 6, pp. e02877-19, /mbio/10/6/mBio.02877-19.atom, Dec. 2019, doi: 10.1128/mBio.02877-19.

[12] P. Turner et al., ‘The aetiologies of central nervous system infections in hospitalised Cambodian children’, BMC Infect. Dis., vol. 17, no. 1, p. 806, Dec. 2017, doi: 10.1186/s12879-017-2915-6.

[13] J. Kawada et al., ‘Next-Generation Sequencing for the Identification of Viruses in Pediatric Acute Encephalitis and Encephalopathy’, Open Forum Infect. Dis., vol. 3, no. suppl_1, p. 1172, Dec. 2016, doi: 10.1093/ofid/ofw172.875.

[14] S. L. Smits et al., 'Novel Cyclovirus in Human Cerebrospinal Fluid, Malawi, 2010–2011', Emerg. Infect. Dis., vol. 19, no. 9, Sep. 2013, doi: 10.3201/eid1909.130404.

[15] 'JASP TEAM', 2020. [Online]. Available: https://jasp-stats.org/

[16] W. Viechtbauer, 'Conducting Meta-Analyses in R with the **metafor** Package', J. Stat. Softw., vol. 36, no. 3, 2010, doi: 10.18637/jss.v036.i03.

[17] J. Granerod et al., 'Causes of encephalitis and differences in their clinical presentations in England: a multicentre, population-based prospective study', Lancet Infect. Dis., vol. 10, no. 12, pp. 835–844, Dec. 2010, doi: 10.1016/S1473-3099(10)70222-X.

[18] J. J. C. de Vries et al., 'Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples', Infectious Diseases (except HIV/AIDS), preprint, May 2021. doi: 10.1101/2021.05.04.21256618.

[19] F. X. López-Labrador et al., 'Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure', J. Clin. Virol., vol. 134, p. 104691, Jan. 2021, doi: 10.1016/j.jcv.2020.104691.

[20] C. P. Oechslin et al., 'Limited Correlation of Shotgun Metagenomics Following Host Depletion and Routine Diagnostics for Viruses and Bacteria in Low Concentrated Surrogate and Clinical Samples', Front. Cell. Infect. Microbiol., vol. 8, p. 375, Oct. 2018, doi: 10.3389/fcimb.2018.00375.

[21] A. L. van Rijn et al., 'The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease', PLOS ONE, vol. 14, no. 10, p. e0223952, Oct. 2019, doi: 10.1371/journal.pone.0223952.

[22] S. van Boheemen et al., 'Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients', J. Mol. Diagn., vol. 22, no. 2, pp. 196–207, Feb. 2020, doi: 10.1016/j.jmoldx.2019.10.007.

[23] S. A. Ogram and J. B. Flanegan, 'Non-template functions of viral RNA in picornavirus replication', Curr. Opin. Virol., vol. 1, no. 5, pp. 339–346, Nov. 2011, doi: 10.1016/j.coviro.2011.09.005.

[24] L. L. M. Poon, D. C. Pritlove, E. Fodor, and G. G. Brownlee, 'Direct Evidence that the Poly(A) Tail of Influenza A Virus mRNA Is Synthesized by Reiterative Copying of a U Track in the Virion RNA Template', J. Virol., vol. 73, no. 4, pp. 3473–3476, Apr. 1999, doi: 10.1128/JVI.73.4.3473-3476.1999.

[25] T. N. Wylie, K. M. Wylie, B. N. Herter, and G. A. Storch, 'Enhanced virome sequencing using targeted sequence capture', Genome Res., vol. 25, no. 12, pp. 1910–1920, Dec. 2015, doi: 10.1101/gr.191049.115.

[26] T. Briese et al., 'Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis', mBio, vol. 6, no. 5, pp. e01491-15, Sep. 2015, doi: 10.1128/mBio.01491-15.

[27] D. Liu et al., 'Development and Multicenter Assessment of a Reference Panel for Clinical Shotgun Metagenomics for Pathogen Detection', In Review, preprint, Feb. 2021. doi: 10.21203/rs.3.rs-208796/v1.

* most comprehensive benchmark to date including both wet and dry lab component

[28] 'Labs NHS'. Accessed: May 07, 2021. [Online]. Available: http://www.labs.gosh.nhs.uk/laboratory-services/microbiology-virology-and-infection-control/metagenomics-pathogen-detection

[29] B. P. Bass, K. B. Engel, S. R. Greytak, and H. M. Moore, 'A Review of Preanalytical Factors Affecting Molecular, Protein, and Morphological Analysis of Formalin-Fixed, Paraffin-Embedded (FFPE) Tissue: How Well Do You Know Your FFPE Specimen?', Arch. Pathol. Lab. Med., vol. 138, no. 11, pp. 1520–1530, Nov. 2014, doi: 10.5858/arpa.2013-0691-RA.

[30] H. Do and A. Dobrovic, 'Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization', Clin. Chem., vol. 61, no. 1, pp. 64–71, Jan. 2015, doi: 10.1373/clinchem.2014.223040.

[31] S. R. Greytak et al., 'National Cancer Institute Biospecimen Evidence-Based Practices: Harmonizing Procedures for Nucleic Acid Extraction from Formalin-Fixed, Paraffin-Embedded Tissue', Biopreservation Biobanking, vol. 16, no. 4, pp. 247–250, Aug. 2018, doi: 10.1089/bio.2018.0046.

[32] Q. Peng, R. Vijaya Satya, M. Lewis, P. Randad, and Y. Wang, 'Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes', BMC Genomics, vol. 16, no. 1, p. 589, Dec. 2015, doi: 10.1186/s12864-015-1806-8.

[33] T. Smith, A. Heger, and I. Sudbery, 'UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy', Genome Res., vol. 27, no. 3, pp. 491–499, Mar. 2017, doi: 10.1101/gr.209601.116.

[34] J. J. C. de Vries et al., 'Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting', J. Clin. Virol., vol. 138, p. 104812, May 2021, doi: 10.1016/j.jcv.2021.104812.

[35] S. Morfopoulou and V. Plagnol, 'Bayesian mixture analysis for metagenomic community profiling', Bioinformatics, vol. 31, no. 18, pp. 2930–2938, Sep. 2015, doi: 10.1093/bioinformatics/btv317.

[36] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg, 'Centrifuge: rapid and sensitive classification of metagenomic sequences', Genome Res., vol. 26, no. 12, pp. 1721–1729, Dec. 2016, doi: 10.1101/gr.210641.116.

[37] D. W. Lewandowska et al., 'Unbiased metagenomic sequencing complements specific routine diagnostic methods and increases chances to detect rare viral strains', Diagn. Microbiol. Infect. Dis., vol. 83, no. 2, pp. 133–138, Oct. 2015, doi: 10.1016/j.diagmicrobio.2015.06.017.

[38] J. M. Martí, 'Recentrifuge: Robust comparative analysis and contamination removal for metagenomics', PLOS Comput. Biol., vol. 15, no. 4, p. e1006967, Apr. 2019, doi: 10.1371/journal.pcbi.1006967.

* Algorithm for correction of environmental sequences, potentially many more will follow as clearing out contamination will increase specificity

[39] V. C. Piro and B. Y. Renard, 'Contamination detection and microbiome exploration with GRIMER', Bioinformatics, preprint, Jun. 2021. doi: 10.1101/2021.06.22.449360.

[40] M. Vilsker et al., 'Genome Detective: an automated system for virus identification from high-throughput sequencing data', Bioinformatics, vol. 35, no. 5, pp. 871–873, Mar. 2019, doi: 10.1093/bioinformatics/bty695.

[41] B. Giardine, 'Galaxy: A platform for interactive large-scale genome analysis', Genome Res., vol. 15, no. 10, pp. 1451–1455, Sep. 2005, doi: 10.1101/gr.4086505.

[42] 'BlueBee'. Accessed: May 07, 2021. [Online]. Available: https://emea.illumina.com/company/about-us/mergers-acquisitions/bluebee.html

[43] T. G. Burland, 'DNASTAR's Lasergene Sequence Analysis Software', in Bioinformatics Methods and Protocols, vol. 132, New Jersey: Humana Press, 1999, pp. 71–91. doi: 10.1385/1-59259-192-2:71.

[44] S. S. Minot, N. Krumm, and N. B. Greenfield, 'One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification', Bioinformatics, preprint, Sep. 2015. doi: 10.1101/027607.

[45] S. Flygare et al., 'Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling', Genome Biol., vol. 17, no. 1, p. 111, Dec. 2016, doi: 10.1186/s13059-016-0969-1.

[46] 'IDbyDNA'. Accessed: May 11, 2021. [Online]. Available: https://www.idbydna.com

[47] M. Asplund et al., 'Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries', Clin. Microbiol. Infect., vol. 25, no. 10, pp. 1277–1285, Oct. 2019, doi: 10.1016/j.cmi.2019.04.028.

[48] S. N. Naccache et al., 'A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples', Genome Res., vol. 24, no. 7, pp. 1180–1192, Jul. 2014, doi: 10.1101/gr.171934.113.

[49] C. Rodriguez et al., 'Pathogen identification by shotgun metagenomics of patients with necrotizing soft-tissue infections', Br. J. Dermatol., vol. 183, no. 1, pp. 105–113, Jul. 2020, doi: 10.1111/bjd.18611.

[50] C. Rodriguez et al., 'Fatal Encephalitis Caused by Cristoli Virus, an Emerging Orthobunyavirus, France', Emerg. Infect. Dis., vol. 26, no. 6, pp. 1287–1290, Jun. 2020, doi: 10.3201/eid2606.191431.

[51] C. Rodriguez et al., 'Viral genomic, metagenomic and human transcriptomic characterization and prediction of the clinical forms of COVID-19', PLoS Pathog., vol. 17, no. 3, p. e1009416, Mar. 2021, doi: 10.1371/journal.ppat.1009416.

[52] 'VirMet'. Accessed: May 07, 2021. [Online]. Available: https://github.com/medvir/VirMet

[53] S. Morfopoulou et al., 'Human Coronavirus OC43 Associated with Fatal Encephalitis', N. Engl. J. Med., vol. 375, no. 5, pp. 497–498, Aug. 2016, doi: 10.1056/NEJMc1509458.

[54] S. H. Lum et al., 'An emerging opportunistic infection: fatal astrovirus (VA1/HMO-C) encephalitis in a pediatric stem cell transplant recipient', Transpl. Infect. Dis., vol. 18, no. 6, pp. 960–964, Dec. 2016, doi: 10.1111/tid.12607.

[55] S. Morfopoulou et al., 'Deep sequencing reveals persistence of cell-associated mumps vaccine virus in chronic encephalitis', Acta Neuropathol. (Berl.), vol. 133, no. 1, pp. 139–147, Jan. 2017, doi: 10.1007/s00401-016-1629-y.

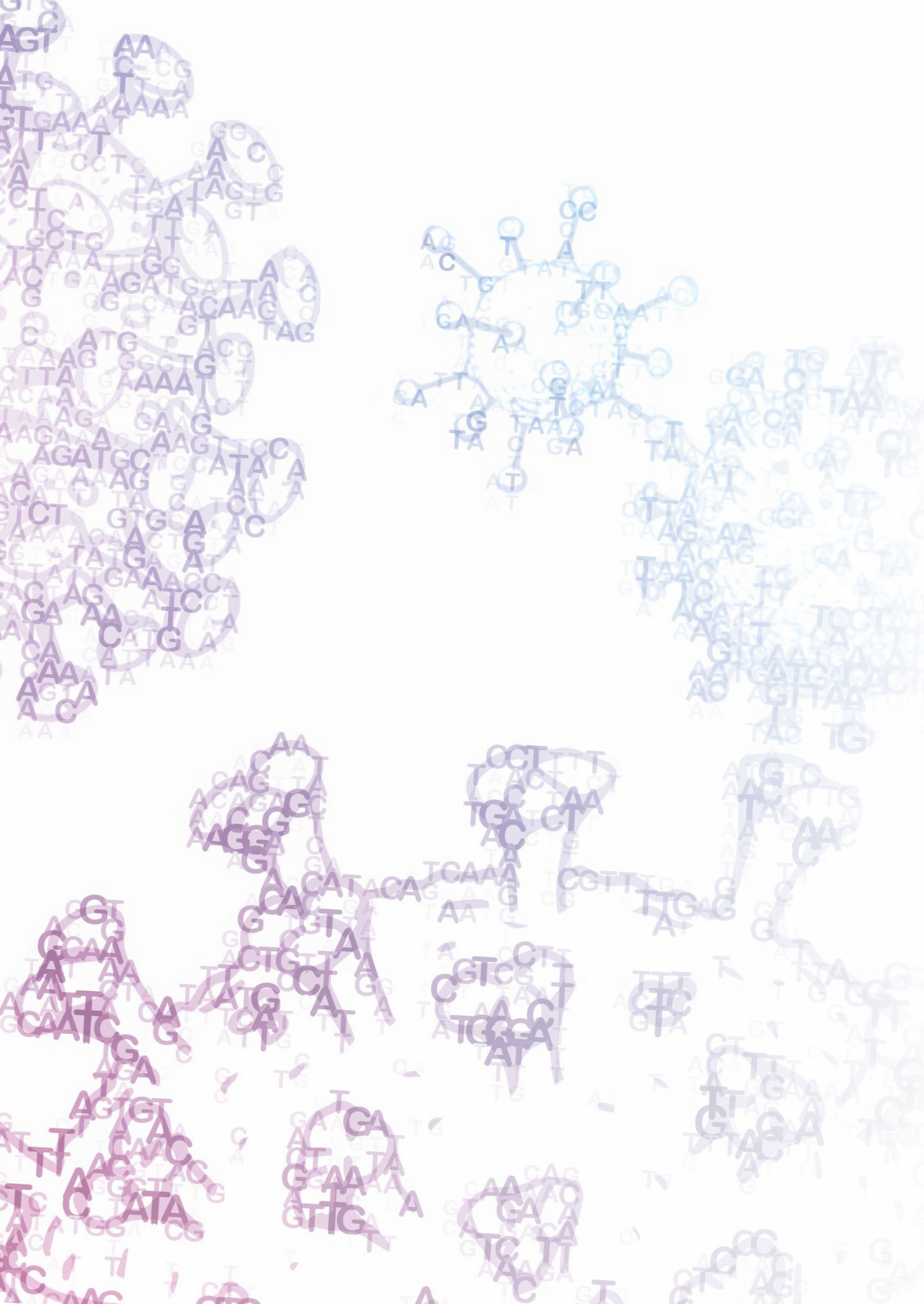
[56] K. Mongkolrattanothai and J. Dien Bard, 'The utility of direct specimen detection by Sanger sequencing in hospitalized pediatric patients', Diagn. Microbiol. Infect. Dis., vol. 87, no. 2, pp. 100–102, Feb. 2017, doi: 10.1016/j.diagmicrobio.2016.10.024.

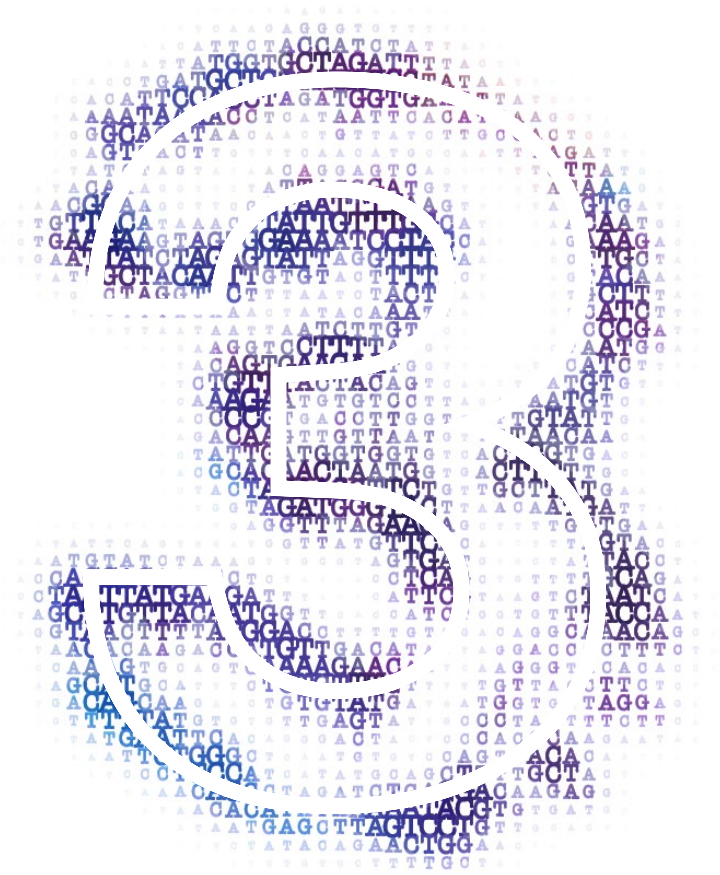
[57] M. Alawi et al., 'DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples', Sci. Rep., vol. 9, no. 1, p. 16841, Nov. 2019, doi: 10.1038/s41598-019-52881-4.

[58] M. Christopeit et al., 'Suspected encephalitis with Candida tropicalis and Fusarium detected by unbiased RNA sequencing', Ann. Hematol., vol. 95, no. 11, pp. 1919–1921, Oct. 2016, doi: 10.1007/s00277-016-2770-3.

[59] N. Fischer et al., 'Evaluation of Unbiased Next-Generation Sequencing of RNA (RNA-seq) as a Diagnostic Method in Influenza Virus-Positive Respiratory Samples', J. Clin. Microbiol., vol. 53, no. 7, pp. 2238–2250, Jul. 2015, doi: 10.1128/JCM.02495-14.

* A reference of importance





Chapter 3 Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics

Ellen C. Carbo[a], Emilie P. Buddingh[b], Evita Karelioti[c], Igor Sidorov[a], Mariet C.W. Feltkamp[a], Peter A. von dem Borne[d], Jan J.G.M. Verschuuren[e], Aloys C.M. Kroes[a], Eric C.J. Claas[a], Jutte J.C. de Vries[a]

a Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands
b Willem-Alexander Children's Hospital, Department of Pediatrics, Leiden University Medical Center, Leiden, The Netherlands
c GenomeScan B.V., Leiden, The Netherlands
d Department of Hematology, Leiden University Medical Center, Leiden, The Netherlands
e Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands

Published in Journal of Clinical Virology

Abstract

Metagenomic sequencing is a powerful technique that enables detection of the full spectrum of pathogens present in any specimen in a single test. Hence, metagenomics is increasingly being applied for detection of viruses in clinical cases with suspected infections of unknown etiology and a large number of relevant potential causes. This is typically the case in patients presenting with encephalitis, in particular when immunity is impaired by underlying disorders.

In this study, viral metagenomics has been applied to a cohort of hematological patients with encephalitis of unknown origin.

Because viral loads in cerebrospinal fluid of patients with encephalitis are generally low, the technical performance of a metagenomic sequencing protocol with viral enrichment by capture probes targeting all known vertebrate viral sequences was studied. Subsequently, the optimized viral metagenomics protocol was applied to a cohort of hematological patients with encephalitis of unknown origin.

Viral enrichment by capture probes increased the viral sequence read count of metagenomics on cerebrospinal fluid samples 100 – 10.000 fold, compared to unenriched metagenomic sequencing.

In five out of 41 (12%) hematological patients with encephalitis, a virus was detected by viral metagenomics which had not been detected by current routine diagnostics. BK polyomavirus, hepatitis E virus, human herpes virus-6 and Epstein Barr virus were identified by this unbiased metagenomic approach.

This study demonstrated that hematological patients with encephalitis of unknown origin may benefit from early viral metagenomics testing as a single step approach.

Highlights

- A metagenomics protocol employing virus capture probes was validated and retrospectively applied to 41 hematological adult and pediatric patients presenting with encephalitis of unknown aetiology
- Viral enrichment by capture probes increased sensitivity of viral metagenomics on cerebrospinal fluid samples 100 – 10.000 fold, compared to unenriched metagenomic sequencing
- In 12% of hematological patients with encephalitis of unknown origin, a virus was detected by viral metagenomics, which was not found by routine diagnostics
- Viral metagenomics represents a valuable addition to the diagnostics repertoire for hematological patients with suspected CNS infection

Introduction

Encephalitis is an important clinical condition with high morbidity and mortality and therefore necessitates a proper and timely diagnosis and pathogen identification [1]. However, up to 63% of the encephalitis cases remain undiagnosed [1] and as a result, no targeted treatment can be initiated, no specific prognostic information can be obtained, and in outbreak settings no effective preventive measures can be taken.

Metagenomic next-generation sequencing has the potential to detect the full spectrum of viral pathogens in a single test. An increasing number of case reports have described the application of metagenomics to clinical cases of encephalitis of unknown origin in both immunocompetent and immunocompromised patients [2-15]. Immunocompromised patients are most at risk of infection with unexpected and novel pathogens and may present with insidious clinical symptoms [2,16]. Recent prospective studies evaluated the use of viral metagenomics for undiagnosed cases in parallel with conventional diagnostics over a period of one year or longer [17,18]. Only a reduced portion of immunocompromised patients was represented in these studies, mainly corresponding to HIV and solid organ transplants. To date, no metagenomic cohort studies have been published focusing on hematological patients with encephalitis.

Cerebrospinal fluid (CSF) remains the most common sample type obtained for diagnostics in cases of encephalitis, though brain biopsies tend to have a higher diagnostic yield for metagenomics [2,10,19,20] as viral loads are lower in CSF. Moreover, metagenomic analysis is greatly affected by an extremely low pathogen-to-host genome ratio. Consequently, a lower sensitivity of metagenomic sequencing has been reported, when compared with conventional PCR-based molecular assays [21-26]. Host cell depletion is one way to increase the relative abundance of viral nucleic acids in metagenomic sequencing, this approach has not demonstrated enough benefits when analyzing clinical samples [26]. In contrast, virus genome enrichment by means of capture probes has been shown to significantly enhance virus detection when sequencing for example respiratory samples [27-30].

In this study, the technical performance of a metagenomic sequencing protocol using capture probes targeting all viral taxa known to infect vertebrates was determined when applied to CSF samples. This technical performance study was followed by a retrospective cohort study with hematologic adult and pediatric patients with encephalitis of unknown etiology.

Methods

Patient and sample selection

For the technical validation study, fifteen CSF samples of patients with encephalitis of known etiology previously sent to the Clinical Microbiological Laboratory (CML) of the Leiden University Medical Center (LUMC, The Netherlands) in the period of 2012-2017. Samples were selected based on positive real-time PCR findings of all the common viruses known to cause encephalitis; both samples with low and high viral loads were selected. These samples were tested by means of a lab-developed metagenomic protocol with and without viral capture probes. Additionally, three tissue biopsies from enteral origin were tested since brain biopsy availability was limited.

Following the technical validation, a cohort of 41 adult and pediatric hematological LUMC patients presenting with clinical symptoms of (meningitis-)encephalitis by the treating clinician, based on a combination of clinical, biological and radiological data, was selected for retrospective analysis. Their CSF samples and brain tissue (one patient) were previously sent to the CML for routine diagnostics in the period of 2011-2019 and selected based on negative real-time PCR results for viral and bacterial pathogens, the latter by culture and PCR.

Ethical approval

This study was approved by the medical ethics review committee of the Leiden University Medical Center (CME number B19.021)

Routine real-time PCR testing (PCR)

In the absence of relevant travel history, the laboratory-developed molecular real-time PCR panel for detection of pathogens in CSF consists of herpes simplex virus type 1 and 2 (HSV1/2), varicella zoster virus, enterovirus and parechovirus. In immunocompromised patients, the panel is expanded with Epstein Barr virus, human cytomegalovirus, JC virus and human herpesvirus type 6 (HHV-6), upon clinical request. These real-time PCRs are performed with internal controls for nucleic acid extraction and real-time PCR inhibition as published previously [31-37]. The initial diagnostic results were confirmed in this study by retesting (see table 1) to ensure the sample integrity after storage at -80°C.

Metagenomic next-generation sequencing (mNGS)

The metagenomics protocol used has previously been described and optimized for simultaneous detection of RNA and DNA targets [38,39]. In short, 20.000-50.000 copies of internal controls, equine arteritis virus (EAV) for RNA and phocid herpesvirus-1 (PhHV) for DNA viruses were spiked into the clinical samples. Subsequently, nucleic acids were extracted directly from 200 µl CSF sample using the MagNApure 96 DNA and Viral NA Small volume extraction kit on the MagNAPure 96 system (Roche Diagnostics, Almere, The Netherlands) with 100 µL output eluate. Extraction buffer only was used as negative control (for extraction, library preparation, and sequencing), this negative control will be given the same treatment from nucleotide extraction until sequencing to detect and rule out contamination. From each sample 50 ul of eluate was used as input and concentrated using the SpeedVac vacuum concentrator (Eppendorf). Samples were dissolved in 10 µl of master mix for fragmentation (consisting of NEB next First Strand Synthesis, random primers and nuclease free water). RNA library preparation was performed using NEBNext Ultra II Directional RNA Library prep kit for Illumina with several in-house adaptations [39] to the manufacturers protocol in order to enable simultaneous detection of both DNA and RNA in a single tube per sample. Poly A mRNA capture isolation, rRNA depletion and DNase treatment steps were omitted, and diluted full size Y-shaped, dual indexed adaptors (1.5 uM) were used. For comparison, library preparation by means of the NEBNext Ultra II DNA Library preparation kit was performed with preceding cDNA and second strand synthesis step. Resulting amplified libraries were used as input material for capture of specific target regions or were subjected to sequence analysis without further processing.

Clustering and metagenomic sequencing using the NovaSeq6000 sequencing system (Illumina, San Diego, CA, USA) was performed according to manufacturer's protocols. Approximately 10 million 150 bp paired-end reads were obtained per sample.

Viral capture probe enrichment

SeqCap EZ Hypercap probes (Roche), designed to cover the genomes of 207 viral taxa known to infect vertebrates including humans were utilized. A complete list of the viral taxa included can be found in the supplementary tables of the manuscript by Briese et al [40]. The quality and quantity of the amplified libraries before capture were determined using the Fragment Analyzer (Agilent) and Qubit (Invitrogen) respectively. For capturing, 250 ng of four amplified DNA libraries were combined in a single pool resulting in a combined mass of 1 µg. For enrichment of the DNA sample library pools, the SeqCap EZ HyperCap Workflow User's Guide (Roche) was followed with several in-house adaptations to the manufacturers protocol. Briefly, human Cot DNA and blocking oligos (Integrated DNA Technologies) were added to each library pool to block non-specific cross hybridization. The target regions were captured by hybridizing each pool of four sample libraries with the SeqCap EZ probe pool [40] overnight. The HyberCap Target Enrichment kit and Hyber Cap Bead kit were used for washing and recovery of the captured DNA. Finally, post-capture PCR amplification was performed using KAPA HiFi HotStart ReadyMix (2X) and Illumina NGS primers (5 µM), followed by DNA purification using AMPure XP beads. Quality and quantity of the post-capture multiplexed libraries were determined by Fragment Analyzer (Agilent) or Bioanalyzer (Agilent). The coefficient of variation as measure for reproducibility for the whole procedure was approximately 5% between runs.

Bioinformatic analysis

Primary data analysis and results Image analysis, base calling, and quality check was performed with the Illumina data analysis pipeline RTA3.4.4 and bcl2fastq v2.20 (Illumina). After quality pre-processing, sequencing reads were taxonomically classified with the pipeline Centrifuge [41] using an index database constructed from NCBI’s RefSeq and taxonomy databases (accessed April 4th, 2019). Reads with multiple best matches were uniquely assigned to the lowest common ancestor (k = 1 Centrifuge setting; previously validated [39]. Negative control sequence reads were subtracted from patient sample reads by Recentrifuge 0.28.7 [42]. Metagenomic findings were confirmed by a second pipeline, GenomeDetective [43] version 1.111 (accessed December 2018 — January 2019) accounting for horizontal genome coverage (%) and confirmatory real-time PCR. Read counts were normalized for total read count and genome size.

Results

Technical performance on PCR-positive CSF samples

The results of the comparison of the metagenomic protocol with and without viral enrichment using capture probes for real-time PCR positive clinical CSF samples are shown in table 1. The metagenomic protocol without enrichment failed to detect the target viruses in three out of 18 cases. In contrast, the metagenomic protocol with enrichment for vertebrate viruses by capture probes detected all viruses that had been detected by real-time PCR. The target virus read counts were increased

100-10.000 fold after viral enrichment. Plots of horizontal coverage of viral sequences, with and without viral capture probes, are shown in Figure 1.

Retrospective study: clinical cohort

Following the validation of the use of the viral capture probes, the metagenomic protocol was used for the clinical application study on samples of pediatric and adult hematological patients with encephalitis of unknown etiology. In total 46 samples (42 CSF samples, one brain biopsy, three blood samples) of 41 patients, including 17 children, were tested. Viral metagenomic sequencing resulted in virus detection in four CSF samples and one brain biopsy (5/41, 12%, Table 2). The clinical symptoms, underlying condition, imaging findings and treatment are shown in Table 3. In these five cases, the virus detected by means of metagenomics had not been targeted by the routine PCR assays that were performed initially.

Table 1. Comparison of the metagenomic protocol with and without viral capture probes in a panel of PCR positive CSF samples.

Patient	Sample type	PCR result	Initial Cq-value/load (diagnostics)	Retested Cq-value/ load (current study)
1	CSF	Enterovirus	27	27
2	CSF	Enterovirus	30	34
3	CSF	Herpes simplex virus type 1	25	NT
4	CSF	Herpes simplex virus type 1	30	NT
5	CSF	HIV type 1	302.500c/mL[c]	NT
6	CSF	Varicella zoster virus	27	28
7	CSF	Varicella zoster virus	30	28
8	CSF	Varicella zoster virus	31	31
9	CSF	Epstein-Barr virus	4.8 $\log_{10}$ IU/mL	4.3 $\log_{10}$ IU/mL
10	CSF	Epstein-Barr virus	3.8 $\log_{10}$ IU/mL	4.1 $\log_{10}$ IU/mL
11	CSF	Enterovirus	33	34
12	Biopsy	Human cytomegalovirus	22	23
13	Biopsy	Human cytomegalovirus	22	26
14	Biopsy	Human cytomegalovirus	24	28
15	CSF	Human cytomegalovirus, resistent	27	NT
16	CSF	CSF: negative but biopsy astrovirus PCR positive	Neg	Neg
17	CSF	Human herpes virus type 6	32	26
18	CSF	Human herpes virus type 6	35	34

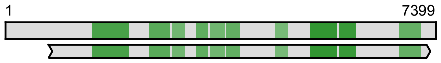
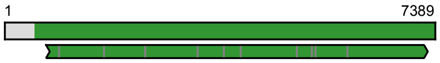
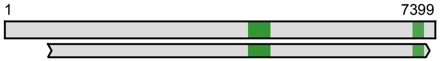
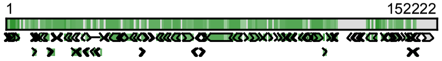
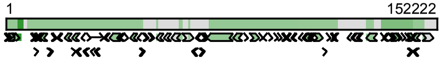
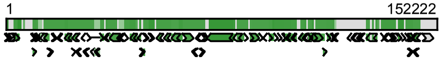
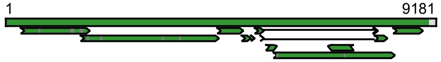
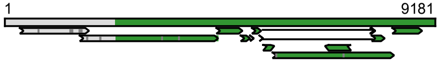
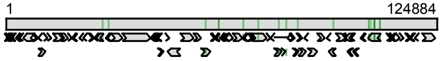
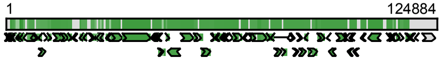
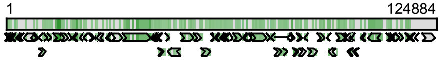
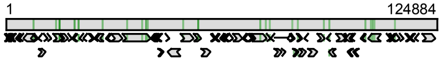
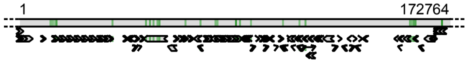
mNGS; metagenomic next-generation sequencing, CSF; cerebrospinal fluid, NT; not tested

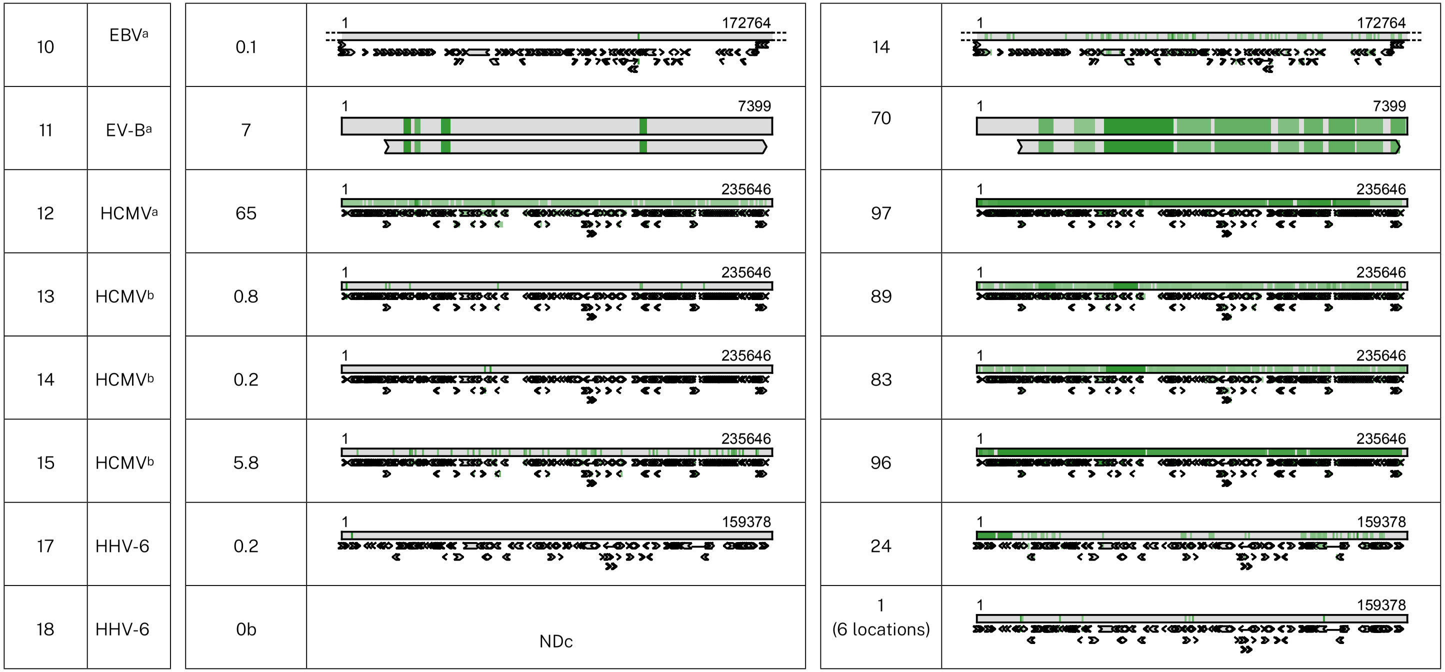
[a] NEBNext Ultra II Directional RNA Library preparation kit with in-house adaptations for total NA sequencing (see methods)

[b] NEBNext Ultra II DNA Library preparation kit preceded by cDNA and 2nd strand synthesis for total NA sequencing (see methods)

[c] Insufficient material available for retesting

mNGS results, without viral probes (read count, Centrifuge)		mNGS results, with viral probes (read count, Centrifuge) [a]	Increase in read count (-fold)
Adapted RNA prep.[a]	DNA prep. incl. cDNA[b]		
0	367	515.069	1.404
0	0	12.368	>12.368
6.616	4.842	3.302.218	499
NT	144	913.662	6.345
2.281	187	38.749.926	16.988
286	0	334.368	1.169
NT	36	131.138	3.643
NT	3	10.241	3.412
NT	4	8.172	2.043
0	90	28.044	312
0	0	15.829	>15.829
2.228	8.000	2.047.002	256
NT	193	169.154, 113.777 (duplicate)	876
NT	96	160.639	1.673
NT	22.350	3.577.617	160
0	0	0	Not applicable
26	306	168.837	552
NT	0	1.283	>1.283

Patient #	Virus	WITHOUT capture probes		WITH capture probes	
		Genome coverage, %	Distribution of reads	Genome coverage, %	Distribution of reads
1	EV-B[a]	44		93	
2	EV-B[a]	0	ND[c]	8	
3	HSV-1[a]	77		73	
4	HSV-1	0[b]	ND[c]	78	
5	HIV-1[a]	98		74	
6	VZV[a]	4		82	
7	VZV	0[b]	ND[c]	55	
8	VZV	0[b]	ND[c]	5	
9	EBV	0[b]	ND[c]	5	



10	EBV[a]	0.1	1 … 172764	14	1 … 172764
11	EV-B[a]	7	1 … 7399	70	1 … 7399
12	HCMV[a]	65	1 … 235646	97	1 … 235646
13	HCMV[b]	0.8	1 … 235646	89	1 … 235646
14	HCMV[b]	0.2	1 … 235646	83	1 … 235646
15	HCMV[b]	5.8	1 … 235646	96	1 … 235646
17	HHV-6	0.2	1 … 159378	24	1 … 159378
18	HHV-6	0[b]	ND[c]	1 (6 locations)	1 … 159378

Figure 1 **Horizontal genome coverage of PCR target viruses in technical performance study without (left) and with viral capture probes (right). Top bar represents nucleotide alignment, bottom bar(s) represents amino acid alignment, green zone: matching sequences. Sample 16 is not included because of negative PCR results.**

EV-B, enterovirus type B; HCMV, human cytomegalovirus; HSV, human simplex virus; VZV, varicellovirus; HIV, human immunodeficiency virus; ND, not detected

[z] NEBNExt Ultra II Directional RNA Library preparation kit with in-house adaptations for total NA sequencing (see table 1 and methods)

[b] NEBNExt Ultra II DNA Library preparation kit preceded by cDNA synthesis (see table 1 and methods)

[c] Not detected (GenomeDetective)

Table 2 Findings by viral metagenomic sequencing of 41 pediatric and adult hematological patients in CSF and brain biopsy samples.

Patient	Sample type	Initially requested molecular diagnostics (real-time PCR, Cq-value)	mNGS results (viral probe capture)	Read count (Centrifuge/ Genome Detective)[a]	Genome coverage (Genome Detective)		Confirmatory testing (PCR)
					%	Depth	Target PCR Cq/load, retested value/load
1, child	Brain tissue, post-mortem	Brain biopsy: HSV-1/2, JC virus, enterovirus, parechovirus, HCMV, EBV, HHV6, M. tuberculosis, and T. whipplei: negative. Positive: adenovirus Cq 36, EBV Cq 37, HHV-6 Cq 33 CSF: Adenovirus, EBV, HHV-6, and VZV: negative	Biopsy: BK polyomavirus CSF: negative	140/ 857	63	33	BKPyV PCR positive Cq 22/ 4.031.000, 549.400 c/ml
2, adult	CSF	HSV-1/2, VZV, enterovirus, parechovirus, HCMV, EBV, JC virus: negative	Human herpes-virus type 6 [a]	29.398/ 225.466	32 (>30 regions)	576	HHV-6 PCR positive Cq 26, 29
3, adult	CSF	HSV1/2, VZV, HCMV, toxoplasma: negative	Human herpes-virus type 6	1.117/ 82.961	5 (>25 regions)	1330	HHV-6 PCR positive Cq 30, 28
4, adult	CSF	HSV1/2, VZV, HCMV, EBV, adenovirus, HHV6, BK, JC, enterovirus, and parechovirus: negative	Hepatitis E virus [a]	61/ 2767	1	2690	HEV PCR positive Cq 36, 37
5, adult	CSF	HSV-1/2, VZV, HCMV, JC virus, adenovirus, enterovirus, parechovirus, M. pneumoniae, L. monocytogenes , M. tuberculosis, and HHV-6: negative	Epstein-Barr virus [a]	26618/ 602.782	46	990	EBV PCR positive Cq 28/ 21.380c/ml

Cq-value; quantification cycle value, mNGS; metagenomic next-generation sequencing, ped.; pediatric patient, HSV-1/2; herpes simplex virus type 1/2, HCMV; human cytomegalovirus, EBV; Epstein Barr virus, HHV-6; human herpes virus type 6, VZV; varicella zoster virus; BKPyV; BK polyomavirus, CSF; cerebrospinal fluid

[a] Initially not tested for by PCR but diagnosed with a delay of up to 2 weeks

Table 3 **Clinical data of the patients with additional mNGS findings (see Table 2).**

	Age, sex	Underlying condition	Clinical signs	Brain MRI findings	Pathology and CSF findings	Other sample types & target	Antiviral treatment	Neurological outcome
1, child	5, M	AML, 26 d post-SCT no engraftment, neutropenic: leucocytes <0.10x10^9/L	Somnolent, pupils dilated, opisthotonus, panuveitis of the left eye	Bilateral asymmetrical hyperintense lesions, compatible with demyelinisation in the context of PML	Post mortem brain biopsy: inflammation, granulomatous, lymphocytes. No AML CSF: leu 0/3µL, segm 0/3µL, ery 55/3µL, glu 2.0 mmol/L, prot 2.97 g/L	BK virus PCR on CSF and plasma: negative, Urine: Cq 13	Foscarnet for clinically suspected HHV6 encephalitis	Deceased from fatal encephalitis 54 d post-SCT
2, adult	30, M	AML, 23 d post-cord blood transplant, neutropenic, leucocytes 0.08x10^9/L	Comatose, status epilepticus	Symmetrical hyperintensity bilateral temporal, mid frontal, and hippocampus; limbic encephalitis	CSF: leu 0/field, ery 0/field. Bacterial/mycological diagnostics negative (CSF, blood)	HHV-6 PCR on plasma Cq 29-37 (<< CSF load)	ACV, 6d after disease onset (HHV-6 PCR+): GCV, duration 4 weeks	Partial neuro-logical recovery, persistent cognitive damage
3, adult	51, F	Hodgkin, 7 d post autologous SCT, leucocytes 6.3x10^9/L	Dysartria, apathia, pupil anisocoria, headache, insult, mutism, nuchal rigidity	Dural and multifocal sulci hyperintensity; meningo-encephalitis	CSF: leu 98/3µL, segm 0/3µL, ery 0/3µL, glu 3.71 mmol/L, prot 0.75 g/L	No plasma/serum available	ACV empirically	Neurological recovery
4, adult	52, M	Multiple myeloma, 14 d post-allogeneic SCT, neutropenic: leucocytes 1.18x10^9/L	Progressive tetraparesis	No enhancement, no tumor EMG: axonal polyneuropathy	CSF: leu 2/3µL, segm 0/3µL, ery 2/3µL, glu 3.6 mmol/L, prot 0.13 g/L, no myeloma cells	HEV PCR on plasma: Cq 25-33 (retrospectively tested)	Ribavirin, IVIG	Progression, cognitive symptoms
5, adult	78, M	7 y post-pancreas island transplant, leucocytes 2.44x10^9/L	Decreased consciousness, epileptic insults	Bilateral temporal and (sub)cortical enhancement; encephalitis, no signs of lymphoma/PTLD	CSF: leu 25x10^6/L, segm <3x10^6/L, ery <500x10^6/L, glu 5.38 mmol/L, prot 3.47 g/L, no lymphoma cells Intestinal biopsies: EBV negative	Plasma EBV PCR: 168 and 448 c/ml (<< CSF load)	Rituximab 3 gifts	Partial neurological recovery; som-nolence

mNGS; metagenomic next-generation sequencing, ped.; pediatric patient, AML; acute myeloid leukemia, SCT; stem cell transplant, PML; progressive multifocal leukoencephalopathy, HHV-6; human herpes virus type 6, ACV; acyclovir, PTLD; post-transplant lymphoproliferative disease, leu; leucocytes, segm; segmented cells, ery; erythrocytes, glu; glucose, prot; total protein

Discussion

In this study, a metagenomic sequencing protocol employing virus capture probes was shown to be highly sensitive and of added value for detection of viruses when applied to a cohort of hematologic adult and pediatric patients with encephalitis of unknown origin. When compared to conventional molecular assays, viral metagenomic sequencing resulted in additional findings in 12% of the cases, including some unexpected viruses initially not tested for. In none of these cases, the diagnosis was made by PCR in the acute phase of the disease.

An increase in the number of case reports involving the experimental use of metagenomic sequencing for diagnosing encephalitis in immunocompetent [3-6,8,11-13] and immunocompromised [7,9,10,14,15] patients is evident in recent literature [2,44]. In these reported cases, the causes of encephalitis detected by metagenomic sequencing were novel, previously unknown viruses. However, well-established causes were reported with similar frequency which could have been identified by conventional molecular techniques, if only requested [2]. Other agents that were involved were known human pathogens that previously not had been observed as a causative agent of encephalitis [2]. Given the bias towards publication of cases with novel viruses, it is expected that when performing cohort studies, novel viruses will be less prominent, in line with our study. It must be noted that detection of novel viruses using a protocol employing virus capture probes is dependent on the amount of sequence similarity between novel and known viruses. None of the recent retrospective [45] and prospective [17,18,46,47] cohort studies on metagenomic sequencing focused on neutropenic hematological patients, whom are likely at increased risk of infectious causes of encephalitis.

The clinical significance of detection of possibly latent and low level persistent viruses in CSF may be difficult to determine. Cohort studies do not provide the best support for causal relationships and the presence of the viruses detected in CSF in this study needs further investigation. For example, encephalitis caused by BK polyomavirus (BKPyV) has been indeed described in a series of case reports [48]. BK virus-associated progressive multifocal leukoencephalitis has previously only been reported in five cases [49]. In the current case, BKPyV was detected in brain tissue, which is considered the best support for diagnosing BKPyV virus encephalitis [49]. The absence of BK viremia in our case suggests localized reactivation of BKPyV in the central nervous sytem (CNS).

Likewise, positive findings of potentially latent viruses such as HHV-6 and EBV should be interpreted in the context of clinical presentation and sample type. HHV-6 DNA can be detected in the blood of approximately 50% of the hematopoietic stem cell transplant recipients [50], while the reported incidence of HHV-6 encephalitis is only 1% [51]. The presence of high viral loads in CSF when compared to blood, as seen in our cases of HHV-6 and EBV reactivation, is suggestive for localized CNS reactivation.

Hepatitis E virus (HEV) infection is associated with neurological dysfunctions, such as encephalitis and Guillain-Barré syndrome. This is supported by both clinical and laboratory studies, detecting HEV RNA in brain tissues of animals after experimental infection [52]. Neurological manifestations of hepatitis E virus infections are more frequently found in immunocompetent patients, suggesting pathophysiological mechanisms involving the immune response [53]. This may be the case in our patient given the lower viral load in CSF.

Though brain biopsies tend to have a higher diagnostic yield of metagenomics [2,10,19,20], the most commonly collected sample type in cases of encephalitis is CSF. Given the commonly low viral loads in CSF, optimal sensitivity is essential but challenging due to the high amount of background sequences [21-26]. Technical validation studies of viral metagenomic protocols using CSF samples with known pathogens [25,26,54,55] are essential to gain insight in its analytical performance including sensitivity. Virus enriched sequence analysis after probe capture has been shown to enhance virus detection significantly in respiratory samples [27-30,56]. The current study confirms an increased sensitivity in both CSF and tissue samples. Efficacy of targeted enrichment is affected by the representation of the viral sequences in the database and probe design [29], which may have caused differences in efficacy between RNA and DNA viruses in the current study. After technical validation, periodic updates of the probe panel with novel sequences would be advisable, though novel viruses commonly share homologous sequences present in the large list of vertebrate viruses targeted by the probes. Extension of the probe panel towards mixed bacteria-virus probe panels would be beneficial for use in routine diagnostics for undiagnosed cases with infectious symptoms.

Summarized, probe enrichment for vertebrate viruses increases sensitivity. The usefulness of viral metagenomics in clinical practice is dependent on several factors, including the technical aspects of the protocol, and the patient population studied. The current study shows that hematological patients may benefit from early, unbiased diagnostics by means of a virus enriched metagenomic sequencing protocol.

Declarations of interest

None

Funding

This work was partially funded by the Leiden University Fund / Den Dulk-Moermans Fonds.

Acknowledgement

We thank Lopje Höcker, Tom Vreeswijk and Margriet Kraakman for their technical assistance.

References

[1] Granerod, J. and N.S. Crowcroft, The epidemiology of acute encephalitis. Neuropsychol Rehabil, 2007. 17(4-5): p. 406-28.

[2] Brown, J.R., T. Bharucha, and J. Breuer, Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. J Infect, 2018. 76(3): p. 225-240.

[3] Chiu, C.Y., et al., Diagnosis of Fatal Human Case of St. Louis Encephalitis Virus Infection by Metagenomic Sequencing, California, 2016. Emerg Infect Dis, 2017. 23(10): p. 1964-1968.

[4] Edridge, A.W.D., et al., Novel Orthobunyavirus Identified in the Cerebrospinal Fluid of a Ugandan Child With Severe Encephalopathy. Clin Infect Dis, 2019. 68(1): p. 139-142.

[5] Fridholm, H., et al., Human pegivirus detected in a patient with severe encephalitis using a metagenomic pan-virus array. J Clin Virol, 2016. 77: p. 5-8.

[6] Hoffmann, B., et al., A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis. N Engl J Med, 2015. 373(2): p. 154-62.

[7] Lipowski, D., et al., A Cluster of Fatal Tick-borne Encephalitis Virus Infection in Organ Transplant Setting. J Infect Dis, 2017. 215(6): p. 896-901.

[8] Mai, N.T.H., et al., Central Nervous System Infection Diagnosis by Next-Generation Sequencing: A Glimpse Into the Future? Open Forum Infect Dis, 2017. 4(2): p. ofx046.

[9] Murkey, J.A., et al., Hepatitis E Virus-Associated Meningoencephalitis in a Lung Transplant Recipient Diagnosed by Clinical Metagenomic Sequencing. Open Forum Infect Dis, 2017. 4(3): p. ofx121.

[10] Naccache, S.N., et al., Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. Clin Infect Dis, 2015. 60(6): p. 919-23.

[11] Perlejewski, K., et al., Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: Unexpected detection of human herpesvirus 1 while searching for RNA pathogens. J Virol Methods, 2015. 226: p. 1-6.

[12] Piantadosi, A., et al., Rapid Detection of Powassan Virus in a Patient With Encephalitis by Metagenomic Sequencing. Clin Infect Dis, 2018. 66(5): p. 789-792.

[13] Tschumi, F., et al., Meningitis and epididymitis caused by Toscana virus infection imported to Switzerland diagnosed by metagenomic sequencing: a case report. BMC Infect Dis, 2019. 19(1): p. 591.

[14] Wilson, M.R., et al., A novel cause of chronic viral meningoencephalitis: Cache Valley virus. Ann Neurol, 2017. 82(1): p. 105-114.

[15] Wilson, M.R., et al., Acute West Nile Virus Meningoencephalitis Diagnosed Via Metagenomic Deep Sequencing of Cerebrospinal Fluid in a Renal Transplant Patient. Am J Transplant, 2017. 17(3): p. 803-808.

[16] Saylor, D., K. Thakur, and A. Venkatesan, Acute encephalitis in the immunocompromised individual. Curr Opin Infect Dis, 2015. 28(4): p. 330-6.

[17] Kufner, V., et al., Two Years of Viral Metagenomics in a Tertiary Diagnostics Unit: Evaluation of the First 105 Cases. Genes (Basel), 2019. 10(9).

[18] Wilson, M.R., et al., Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. N Engl J Med, 2019. 380(24): p. 2327-2340.

[19] Fremond, M.L., et al., Next-Generation Sequencing for Diagnosis and Tailored Therapy: A Case Report of Astrovirus-Associated Progressive Encephalitis. J Pediatric Infect Dis Soc, 2015. 4(3): p. e53-7.

[20] Morfopoulou, S., et al., Deep sequencing reveals persistence of cell-associated mumps vaccine virus in chronic encephalitis. Acta Neuropathol, 2017. 133(1): p. 139-147.

[21] Wylie, K.M., et al., Sequence analysis of the human virome in febrile and afebrile children. PLoS One, 2012. 7(6): p. e27735.

[22] Lim, E.S., et al., Early life dynamics of the human gut virome and bacterial microbiome in infants. Nat Med, 2015. 21(10): p. 1228-34.

[23] Thorburn, F., et al., The use of next generation sequencing in the diagnosis and typing of respiratory infections. J Clin Virol, 2015. 69: p. 96-100.

[24] Junier, T., et al., Viral Metagenomics in the Clinical Realm: Lessons Learned from a Swiss-Wide Ring Trial. Genes (Basel), 2019. 10(9).

[25] Miller, S., et al., Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res, 2019. 29(5): p. 831-842.

[26] Oechslin, C.P., et al., Limited Correlation of Shotgun Metagenomics Following Host Depletion and Routine Diagnostics for Viruses and Bacteria in Low Concentrated Surrogate and Clinical Samples. Front Cell Infect Microbiol, 2018. 8: p. 375.

[27] O'Flaherty, B.M., et al., Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing. Genome Res, 2018. 28(6): p. 869-877.

[28] Wylie, K.M., et al., Detection of Viruses in Clinical Samples by Use of Metagenomic Sequencing and Targeted Sequence Capture. J Clin Microbiol, 2018. 56(12).

[29] Wylie, T.N., et al., Enhanced virome sequencing using targeted sequence capture. Genome Res, 2015. 25(12): p. 1910-20.

[30] Yang, Y., et al., Targeted Sequencing of Respiratory Viruses in Clinical Specimens for Pathogen Identification and Genome-Wide Analysis. Methods Mol Biol, 2018. 1838: p. 125-140.

[31] Kalpoe, J.S., et al., Clinical relevance of quantitative varicella-zoster virus (VZV) DNA detection in plasma after stem cell transplantation. Bone Marrow Transplant, 2006. 38(1): p. 41-6.

[32] van der Beek, M.T., et al., Rapid susceptibility testing for herpes simplex virus type 1 using real-time PCR. J Clin Virol, 2013. 56(1): p. 19-24.

[33] van Doornum, G.J., et al., Diagnosing herpesvirus infections by real-time amplification and rapid culture. J Clin Microbiol, 2003. 41(2): p. 576-80.

[34] Benschop, K., et al., Rapid detection of human parechoviruses in clinical samples by real-time PCR. J Clin Virol, 2008. 41(2): p. 69-74.

[35] Read, S.J. and J.B. Kurtz, Laboratory diagnosis of common viral infections of the central nervous system by using a single multiplex PCR screening assay. J Clin Microbiol, 1999. 37(5): p. 1352-5.

[36] Kalpoe, J.S., et al., Validation of clinical application of cytomegalovirus plasma DNA load measurement and definition of treatment criteria by analysis of correlation to antigen detection. J Clin Microbiol, 2004. 42(4): p. 1498-504.

[37] Lankester, A.C., et al., Epstein-Barr virus (EBV)-DNA quantification in pediatric allogenic stem cell recipients: prediction of EBV-associated lymphoproliferative disease. Blood, 2002. 99(7): p. 2630-1.

[38] van Rijn, A.L., et al., The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease. PLoS One, 2019. 14(10): p. e0223952.

[39] van Boheemen, S., et al., Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients. J Mol Diagn, 2019.

[40] Briese, T., et al., Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. mBio, 2015. 6(5): p. e01491-15.

[41] Kim, D., et al., Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res, 2016. 26(12): p. 1721-1729.

[42] Marti, J.M., Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLoS Comput Biol, 2019. 15(4): p. e1006967.

[43] Vilsker, M., et al., Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics, 2019. 35(5): p. 871-873.

[44] Schubert, R.D. and M.R. Wilson, A tale of two approaches: how metagenomics and proteomics are shaping the future of encephalitis diagnostics. Curr Opin Neurol, 2015. 28(3): p. 283-7.

[45] Eibach, D., et al., Viral metagenomics revealed novel betatorquevirus species in pediatric inpatients with encephalitis/meningoencephalitis from Ghana. Sci Rep, 2019. 9(1): p. 2360.

[46] Haston, J.C., et al., Prospective Cohort Study of Next-Generation Sequencing as a Diagnostic Modality for Unexplained Encephalitis in Children. J Pediatric Infect Dis Soc, 2019.

[47] Turner, P., et al., The aetiologies of central nervous system infections in hospitalised Cambodian children. BMC Infect Dis, 2017. 17(1): p. 806.

[48] Chittick, P., J.C. Williamson, and C.A. Ohl, BK virus encephalitis: case report, review of the literature, and description of a novel treatment modality. Ann Pharmacother, 2013. 47(9): p. 1229-33.

[49] Melis, M., et al., BK-virus progressive multifocal leukoencephalitis in a patient with systemic lupus erythematosus. Neurol Sci, 2018. 39(9): p. 1613-1615.

[50] Yamane, A., et al., Risk factors for developing human herpesvirus 6 (HHV-6) reactivation after allogeneic hematopoietic stem cell transplantation and its association with central nervous system disorders. Biol Blood Marrow Transplant, 2007. 13(1): p. 100-6.

[51] Hill, J.A., et al., Cord-blood hematopoietic stem cell transplant confers an increased risk for human herpesvirus-6-associated acute limbic encephalitis: a cohort analysis. Biol Blood Marrow Transplant, 2012. 18(11): p. 1638-48.

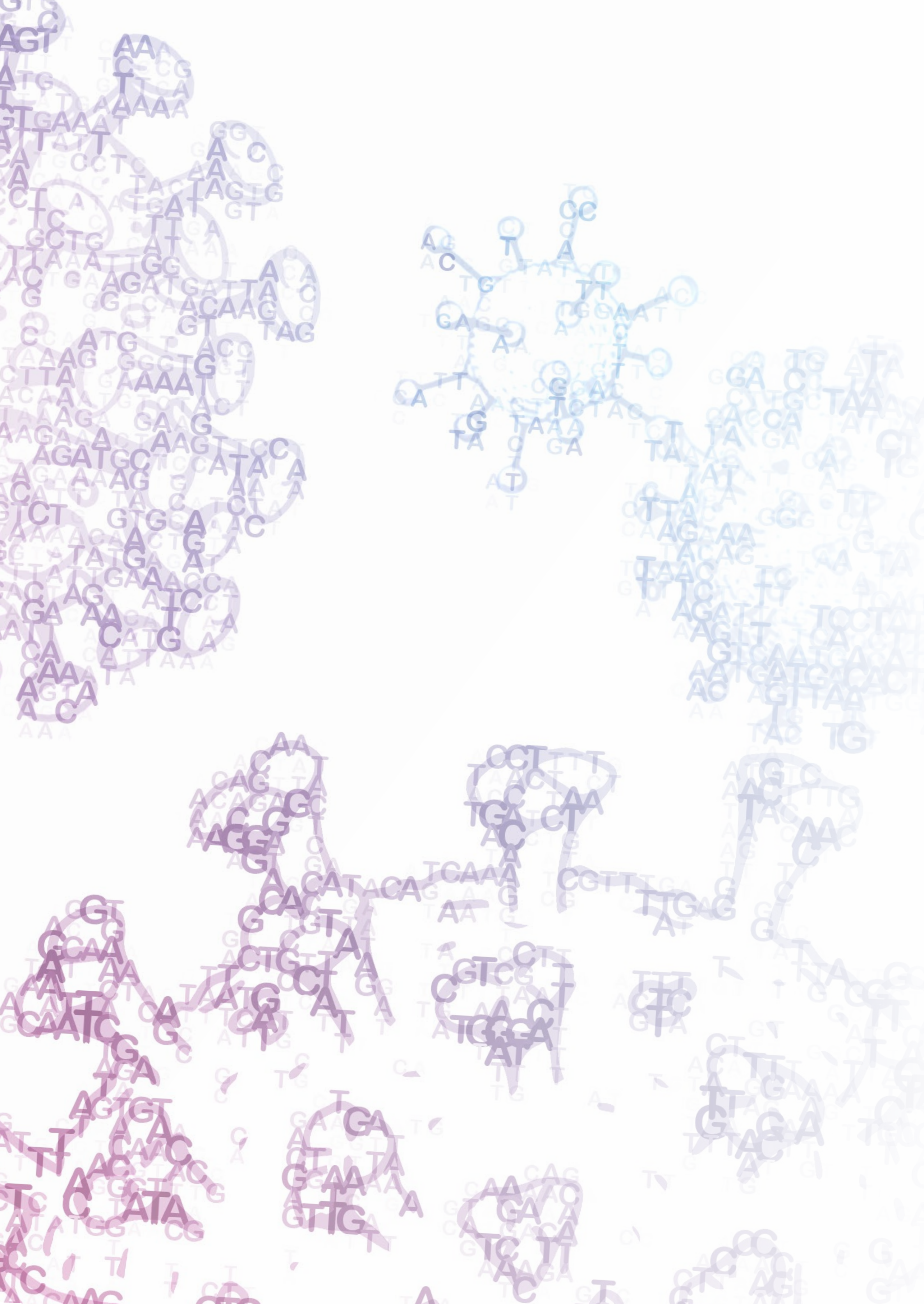
[52] Zhou, X., et al., Hepatitis E Virus Infects Neurons and Brains. J Infect Dis, 2017. 215(8): p. 1197-1206.

[53] Abravanel, F., et al., Acute hepatitis E in French patients and neurological manifestations. J Infect, 2018. 77(3): p. 220-226.

[54] Edridge, A.W.D., et al., Viral Metagenomics on Cerebrospinal Fluid. Genes (Basel), 2019. 10(5).

[55] Bukowska-Osko, I., et al., Sensitivity of Next-Generation Sequencing Metagenomic Analysis for Detection of RNA and DNA Viruses in Cerebrospinal Fluid: The Confounding Effect of Background Contamination. Adv Exp Med Biol, 2016.

[56] Singanallur, N.B., et al., Probe capture enrichment next-generation sequencing of complete foot-and-mouth disease virus genomes in clinical samples. J Virol Methods, 2019. 272: p. 113703.



Chapter 4 Viral metagenomic sequencing in a cohort of international travellers returning with febrile illness

Alhena Reyes [a,b], Ellen C. Carbo [a*], Johan Sippo van Harinxma thoe Slooten [a], Margriet E.M. Kraakman [a], Igor A. Sidorov [a], Eric C.J. Claas [a], Aloys C.M. Kroes [a], Leo G. Visser [c], Jutte J.C. de Vries [a]

**Corresponding author*

Affiliations:
a Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands
b Current affiliation: Microbiology Department, Hospital Universitario 12 de Octubre, Madrid, Spain
c Department of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands

Published in Journal of Clinical Virology

Abstract

Background: Diagnosis of infections in returning international travellers can be challenging because of the broad spectrum of potential infectious etiologies potentially involved. Viral metagenomic next-generation sequencing (mNGS) has the potential to detect any virus present in a patient sample and is increasingly being used for difficult to diagnose cases. The aim of this study was to analyze the performance of mNGS for viral pathogen detection in the clinical setting of international travellers returning with febrile illness.

Methods: Thirty-eight serum samples from international travellers returning with febrile illness and presenting at the outpatient clinic of the Leiden University Medical Center in the Netherlands in the time period 2015-2016 were selected retrospectively. Samples were processed for viral metagenomic sequencing using a probe panel capturing all known vertebrate viruses. Bioinformatic analysis was performed using Genome Detective software for metagenomic virus detection. Metagenomic virus findings were compared with viral pathogen detection using conventional methods.

Results: In 8 out of the 38 patients (21%), a pathogenic virus was detected by mNGS. All viral pathogens detected by conventional assays were also detected by mNGS: dengue virus (n=4 patients), Epstein-Barr virus (n=2), hepatitis B virus (n=1). In addition, mNGS resulted in additional pathogenic findings in 2 patients (5%): dengue virus (n=1), and hepatitis C virus (n=1). Non-pathogenic viruses detected were: GB virus C (n=1) and torque teno viruses (n=3). High genome coverage and depth using capture probes enabled typing of the dengue viruses detected.

Conclusions: Viral metagenomics has the potential to assist the detection of viral pathogens and co-infections in one step in international travellers with a febrile syndrome. Furthermore, viral enrichment by probes resulted in high genome coverage and depth which enabled dengue virus typing.

Keywords

Viral metagenomic; pathogen detection; travellers; fever; capture probes; serum

Introduction

Accurate diagnosis of travel-associated febrile illness in the returning traveller can be challenging, because of the broad spectrum of viral etiologies potentially involved [1,2]. Identification of potential viral pathogens is important for clinical management and epidemiological reasons.

Metagenomic next-generation sequencing (mNGS) has the potential to detect any known or new pathogen in one single run, in contrast to conventional targeted methods such as PCR. In addition, molecular techniques targeting specific pathogens are dependent upon matching specific primers, leaving variant pathogens unidentified. Finally, emerging pathogens that have not associated with a specific clinical syndrome before, for example encephalitis caused by astroviruses, will be included in the metagenomic width of detection [3].

Since the amount of pathogen in a sample is relatively low and human background is high, several strategies have been applied to increase the sensitivity of mNGS based on physical or enzymatic pre-processing of samples for human DNA depletion [3,4]. Another strategy is hybridization enrichment, with the application of a virome probe panel that targets all known vertebrate viruses to increase the sensitivity of virus detection and characterization [5,6]. Viral enrichment by capture probes increased viral sequence read counts in cerebrospinal fluid samples 100 – 10.000 fold, compared to unenriched sequencing [7]. These probe capture panels can also be helpful in the detection of divergent and novel viruses up to approximately 40-58% different at nucleotide level from the genome references used in the probe library design [5-8].

Viral mNGS is increasingly being applied directly on different types of samples from patients for pathogen detection in undiagnosed cases, both in retrospective studies [9] and prospective ones [10] with a wide range of proportion of additional findings. A striking example is the rapid and impactful metagenomic analysis of SARS-CoV-2 in 2019 [11]. The aim of the current study was to investigate the utility of viral enhanced metagenomic sequencing (mNGS) as a diagnostic tool for viral infections in the returning traveller with febrile illness.

Material and Methods

Study design

Retrospectively, a cohort of international travellers with febrile illness upon their return was studied. Patients presenting at the Leiden University Medical Center (LUMC, the Netherlands) from January 2015 to March 2017 with fever after recent international trip and informed consent [12] were enrolled. Serum samples were obtained upon presentation at the first-aid department or outpatient clinic and were tested for dengue antigenemia, malaria, and other infections on clinical suspicion, at the Clinical Microbiology Laboratory of the LUMC as routine diagnostic practice. Patients with proven viral respiratory infections, viral or bacterial gastro-enteritis, or malaria have been excluded from viral metagenomics analysis. Of the included travellers (n=38) serum samples were utilized to perform viral mNGS sequencing independently of conventional test results and diagnosis.

Ethical approval

Approval was obtained from the ethical committee from the LUMC (P11.165 NL 37682.058.11, and Biobank Infectious Diseases protocol 2020-03 & 2020-04 B20.002).

Metagenomic next-generation sequencing (mNGS)

The procedure for metagenomic detection using a viral probe capture panel for clinical samples has been validated previously [7]. Prior to nucleic acid extraction, serum samples were spiked with fixed amounts of non-human pathogenic viruses as internal sequencing RNA and DNA controls: Equine Arteritis Virus (EAV) and Phocid alpha-herpesvirus (PhHV-1). Nucleic acid extraction was performed using MagNApure 96 DNA and Viral NA Small volume extraction kit on the MagNA Pure 96 instrument (Roche, Germany), with 200 µl of serum sample input and 100 µl output eluate. STAR Buffer was used as negative control for the entire workflow from nucleotide isolation throughout sequencing. Library preparation was carried out with the NEBNext® Ultra II Directional RNA Library Prep kit for Illumina® and NEBNext® Multiplex Oligos for Illumina® (unique dual index primers pairs, E6440) with 10 ul of pre-concentrated eluate and following a modified version of the protocols for use with “purified mRNA or rRNA Depleted RNA” as described previously [13,14]. Nuclease free water was used as a library preparation control (upstream negative control). Libraries were combined in pools of three libraries from samples plus one negative control for viral capture probe enrichment.

Viral capture probe enrichment

SeqCap EZ Hypercap probes (Roche), designed to cover the genomes of 207 viral taxa known to infect vertebrates including humans were utilized. A complete list of the viral taxa included can be found in the supplementary tables of the manuscript by Briese et al. [6] The quality and quantity of the amplified libraries before and post-capture were determined using the Agilent 2100 Bioanalyzer (Agilent technologies, Palo Alto, CA, USA) and Qubit (Thermo Fisher, Waltham, MA, USA). For capturing, four amplified DNA libraries were combined in a single pool. For enrichment of the DNA sample library pools, the SeqCap EZ HyperCap Workflow User's Guide (Roche) was followed with several in-house adaptations to the manufacturers protocol as described previously [7]. Subsequently, the target regions were captured by hybridizing each pool of four sample libraries with the SeqCap EZ probe pool [6] overnight. The HyperCap Target Enrichment kit and Hyper Cap Bead kit were used for washing and recovery of the captured DNA. Finally, post-capture PCR amplification was performed using KAPA HiFi HotStart ReadyMix (2X) and Illumina NGS primers (5 uM), followed by DNA purification using AMPure XP beads. Final products were sequenced using the Novaseq6000 platform (Illumina, San Diego, California, USA), obtaining up to 10 million (median 9.9 million reads, IQR 7.5 million) of 150bp paired-end reads per patient sample (GenomeScan B.V. Leiden, the Netherlands).

Bioinformatic analysis

Primary data analysis, bcl conversion and demultiplexing was performed with bcl2fastq (Illumina). After quality pre-processing, including filtering out low-complexity and low-quality reads, the remaining sequencing reads were taxonomically classified and subtyped with metagenomic pipeline Genome Detective (www.genomedetective.com) [15] and classification tool Centrifuge [16] (v1.0.3, GeneBank taxonomy v2019-04-04), including analysis of the proportion of sequence reads assigned to the human genome. The variables collected for virus hits were: number of total, human and viral reads, horizontal coverage (%), number of contigs aligned over the genome, and virus types. Reads were normalized to get the number of reads per Kb genome per Million total reads (RPKM) using the following formula: RPKM= (target reads*1000000*1000)/(total reads after quality check* genome size in base pairs). The following criteria were applied for defining a positive result: horizontal coverage of three or more genome locations without the virus being detected in the negative controls of simultaneous runs, and a positive confirmatory conventional test result. The mNGS detection of a pathogenic virus was subject to confirmatory analysis: RT-PCR on the original sample depending on the virus and reference diagnostic test availability. Bacteriophages and human retroviruses known to integrate in human chromosomes were not taken into account.

Phylogenetic analysis

Typing and phylogenetic analysis based on whole genome sequences was performed using the Genome Detective Typing Tool [15] for dengue virus.

Results

Cohort

One hundred and thirteen returning travellers visited the outpatient clinic of the LUMC from January 2015 to March 2016 with fever. After exclusion of patients with proven respiratory infections and viral or bacterial gastro-enteritis, 38 serum samples from returning travellers with febrile illness were processed for metagenomic sequencing. The mean age of the 38 travellers was 44,2 years (range 13.3 – 71.9). Seventeen (45%) travellers returned from South or Sub-Saharan African countries, 14 (37%) from Central-Southeast Asia, and seven (18%) from Central-South America.

mNGS findings in relation to conventional diagnosis

Thirty-eight serum samples were analyzed by mNGS. The percentage of sequence reads assigned to the human genome was on average 74% (range 5-96%, data not shown).

Six patients had a viral infection diagnosed by conventional serologic methods during the time of the visit at the outpatient clinic: three dengue virus primary/secondary infections, two Epstein-Barr virus primary infections and one chronic hepatitis B virus infection. All viral infections detected by conventional methods were also been detected by metagenomic sequencing (sensitivity 6/6, 100%). Table 1 shows the number of reads, genome coverages and dengue genotypes found in relation with the conventional diagnostic test performed. Genome coverage bars are shown in Figure 1 A.

Additional mNGS findings

Using mNGS, two additional pathogenic viruses were detected (Table 2, Figure 1 B): one additional dengue virus and one hepatitis C virus. These infections were

Table 1. **mNGS results from previously diagnosed viral infections.**

mNGS virus finding#	Patient (age range, country or countries visited)	mNGS finding	Conventional positive tests	Diagnosis	# of reads Genome Detective (15) / Centrifuge (16)	# of normalized reads Genome Detective (15) / Centrifuge (16) (RPKM)	Coverage	# of contigs
1	30-39 y, Brasil	Dengue virus 1 Genotype V	NS1 Ag+, IgG+/IgM+	Dengue virus infection	824,061/ 270,449	1,172/ 1,835	82%	11
2	40-49 y, Malaysia	Dengue virus 3 Genotype I	NS1 Ag+ IgM-/IgG-	Dengue virus infection	1,050,036/ 430,486	59,994/ 74,293	99%	1
3	20-29 y, Sri Lanka	Dengue virus 1 Genotype I	NS1 Ag+ IgM+/IgG–	Dengue virus infection	173,019/ 78,729	876/ 989	88%	7
4	40-49 y, South Africa	Human gamma-herpesvirus 4 (EBV)	EBV VCA IgM+/IgG- EBNA IgG - load: 1.020 IU/mL	Primary EBV infection	463,808/ 307,178	89/ 69	31%	98
5	20-29 y, Myanmar, Thailand	Human gamma-herpesvirus 4 (EBV)	EBV VCA IgM+/IgG+ EBNA IgG -	Primary EBV infection	68,999/ 79,652	64/ 29	8%	43
6	30-39 y, Thailand, Cambodia	Hepatitis B virus	HBsAg+, anti-HBc+, HBe Ag+ load: 8.3 log10 IU/ml	Hepatitis B infection, chronic	206,014,159/ 61,273,548	184,249/ 311,644	100%	1

NS1 Ag; non-structural antigen 1, EBV; Epstein-Barr virus, HBV; Hepatitis B virus

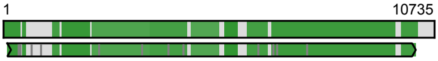
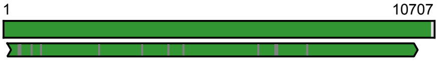
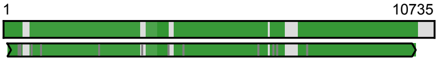
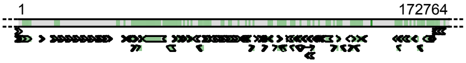
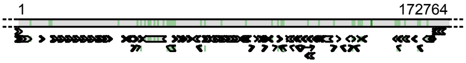
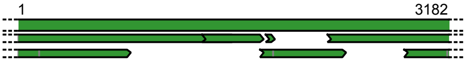
RPKM; Reads per Kb genome per Million total reads confirmed as true positives by qPCR. In the patient in whom the dengue virus infection was diagnosed mNGS, the original dengue NS1 antigen screening test was negative at the time of the visit at the outpatient clinic.

The following non-pathogenic viruses were detected: GB virus C (one patient, 1.821.303 sequence reads) and torque teno viruses in three patients: type 6 (147 reads), 24 (892 reads) and 18 (139 reads), coverage bars are represented in Figure 1B. The GB virus C was found in co-infection with dengue virus.

Virus typing

The four dengue virus infections could also be typed. Three of the infections were classified as serotype 1 (genotype I, IV and V) and one as serotype 3 (genotype I). The HCV positive finding was classified by Genome Detective as genotype 2 with 87.8% horizontal genome alignment and an 80.9% nucleotide identity to reference NC_009823.1 strain. In Supplementary Figure 1 is shown that the mNGS data enabled phylogenetic analysis.

A

mNGS virus finding #	Virus	Genome coverage, %	Distribution of reads
1	Dengue virus 1	82	
2	Dengue virus 3	99	
3	Dengue virus 1	88	
4	Human gamma-herpes virus 4 (EBV)	31	
5	Human gamma-herpesvirus 4 (EBV)	8	
6	Hepatitis B virus	100	

B

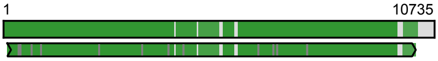
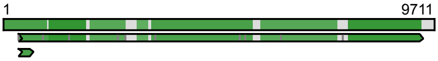
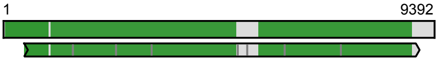
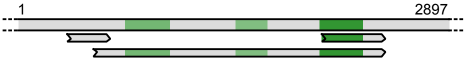
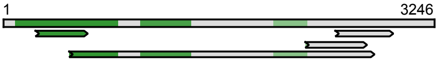
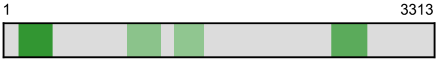
	Virus	Genome coverage, %	Distribution of reads
7	Dengue virus 1	92	
8	Hepatitis C virus genotype 2	88	
9	GB virus C (Pegivirus species)	89	
10	Torque teno mini virus 6	28	
11	Torque teno virus 24	43	
12	Torque teno virus 18	31	

Figure 1. A and B Horizontal genome coverage of mNGS virus findings in patients with conventional diagnosis (A) and in patients without etiology by conventional assays (B).

Top bar represents nucleotide alignment, bottom bar(s) represents amino acid alignment, green zone: matching sequences. EBV; Epstein-Barr virus

Table 2. **Additional findings by mNGS, coverage, and confirmatory test results.**

mNGS virus finding #	Patient (age range, country or countries visited)	Clinical symptoms	Conventional diagnostics	mNGS finding	# of reads Genome Detective (15) / Centrifuge (16)	# of normalized reads Genome Detective (15) / Centrifuge (16) (RPKM)	Coverage	# of contigs	Confirmatory test results (performed after mNGS)
7	20-29 y, Indonesia	Fever since day of return, unknown etiology	Malaria PCR-, smear-, Dengue NS1 Ag- Leptospirosis Ab-/PCR-	Dengue virus 1 genotype IV	1,718,292/ 814,613	63,303/ 68,849	92%	6	Dengue PCR+, Anti-dengue IgG+/ IgM-, suggestive for re-infection
8	60-69 y Suriname, French Guiana	Fever and joint pain 3d after return, Unknown etiology	Chikungunya IgG-/IgM-, Dengue IgG+/gM-, Ag-: past dengue infection Blood culture: negative	Hepatitis C virus genotype 2	939,937/ 1,456,226	8,820/ 2,897	88%	7	Anti-VHC+ Load: 5.640 IU/mL NS5B region: Genotype 2

RPKM; Reads per Kb genome per Million total reads

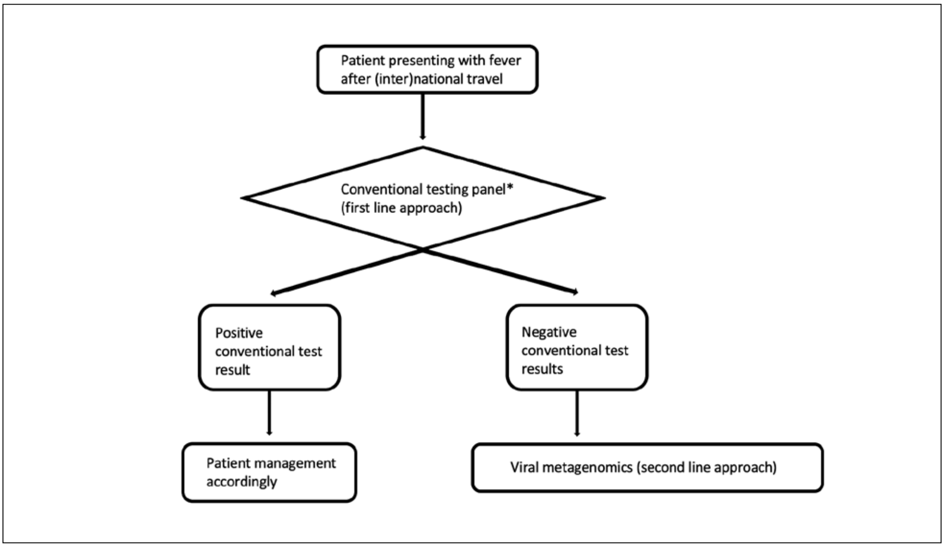

Patient presenting with fever after (inter)national travel
Conventional testing panel* (first line approach)
Positive conventional test result
Negative conventional test results
Patient management accordingly
Viral metagenomics (second line approach)

Figure 2. Decision flowchart with suggested position of viral mNGS in the diagnosis of travellers with fever, enabling the detection of viruses not included in the first line testing panel, such as novel viruses.

This position is based on the evidence, protocols and techniques available to date.

* Not the scope of this review, guideline recommendations differ based on the travel destination, exposures, duration of travel, country of origin and the presence of additional specific symptoms.

Discussion

In the current study, viral mNGS was successful in detecting six previously diagnosed infections and revealed two new findings (5%) in 38 serum samples from returning travellers with febrile illness. In similar studies using mNGS in returned travellers, new and diverse findings have been reported but none of them used a capture panel for pathogen detection [1,17]. Application of capture probes results in higher coverage of the genomes detected and in more reliable sequences because of an increased sequencing depth [6,7]. As a result, subsequent typing and phylogenetic analysis could be performed using the consensus genome sequences after de novo genome assembly.

The diagnosis dengue virus was rejected in one patient after a negative dengue NS1 antigen rapid test upon outpatient visitation, whereas dengue sequences were detected by mNGS and afterwards confirmed by PCR. Dengue virus antigen tests are known for their lower sensitivity after one week of onset of disease and potentially in case of secondary infections with lower loads, suggesting a higher efficacy of molecular diagnostics in these cases [18].

Genotypes of dengue virus were available after sequence analysis using a dengue virus typing tool based on E gene sequences (1,485bp). Dengue types differ in more than 30% in their whole genome nucleotide sequences [19]. Dengue virus typing is of use for epidemiological surveillance of infecting strains. Furthermore, dengue virus typing can be of use for differentiating isolates in case of secondary infections, while severity of primary or secondary infections are related to serotypes [20]. In our study, all the dengue virus types could be properly identified as a result of increased genome coverage and depth due to effective enrichment.

The detection of hepatitis C virus in serum of a patient highlights the potential of mNGS to diagnose infectious diseases beyond the differential diagnostic list of expected pathogens. The patient was a 60-69-years-old traveller presenting with fever at the outpatient clinic three days after the return from Suriname, French Guiana. Patient was tested for dengue antigenemia, and Zika and Chikungunya antibodies which were all negative. There was no suspicion of HCV infection since signs and symptoms were not consistent and transaminases levels were in the range of normality. Although HCV infection is not a cause of febrile illness in the returning traveller, the detection of unsuspected pathogens is crucial for clinical management, therapy administration and prevention of transmission.

GB virus C, formerly known as hepatitis G virus, is a lymphotropic RNA virus of the Pegivirus family, it is related to hepatitis C virus and was thought to cause chronic hepatitis in the past [21,22], however nowadays is considered non-pathogenic.

The use of a virus capture panel has been reported to increase significantly the number of reads and coverages generated in sequencing platforms [7]. It must be noted that EBV coverage was likely to be adversely affected by its genome structure with two very long and multiple short repeat elements. The high sensitivity of NGS when combined with the viral capture panel does not only enable the finding of clinically unsuspected pathogens, but also may provide a negative predictive value within the range of use for clinical practice [7,14]. The increase in viral sequence reads also enables subsequent typing or detection of potential antiviral resistance. A recent review highlighted the sensitivity of hybridization-based enrichment techniques for viral genomes screening and proposed its deployment as an alternative diagnostic tool when traditional methods fail to detect a pathogen, even when viral genomes differs <40% from probe sequences [21].

In conclusion, the application of viral metagenomics in this study provided the additional detection of two unsuspected viral pathogens and one first report coinfection. Metagenomic sequencing has the potential to diagnose a viral febrile syndrome in the returned traveller with the use of a single test. The use of a broad viral capture panels makes this method more sensitive and generates enough reads and coverages for reliable pathogen identification and typing. Implementation of viral metagenomic protocols in diagnostic laboratories is currently modest in size due to several factors including considerable costs, complexity of bioinformatic analysis, laborious protocols, and the time to result in comparison with syndromic PCR panels. While sequencing costs remain declining, cloud-based user-friendly bioinformatic software and formal external quality assessment have become available, a handful of virological diagnostic laboratories currently have implemented mNGS within the scope of their accreditation as an approach for undiagnosed cases (Figure 2). Implementation of established metagenomic protocols in developing countries would be beneficial for the detection of unexpected viruses of local origin. A milestone for implementation of viral metagenomics in both low and high-income countries would be easy access to cloud-based user-friendly bioinformatic analysis software.

Procedures to manage adventitious findings such as HIV should be in place: at the moment of mNGS request, the untargeted nature of this approach is communicated. It is likely that, gradually, the use and experience with this technique will become more widespread and will stimulate the ongoing development and optimization of metagenomic sequencing for diagnostic use.

Funding

We want to thank the Spanish Society of Microbiology and Infectious Diseases (SEIMC) for the mobility grant provided for contributing to this work in the clinical microbiology laboratory of LUMC, The Netherlands.

Conflicts of interest

None

References

[1] Schlagenhauf P, Weld L, Goorhuis A, Gautret P, Weber R, von Sonnenburg F, et al. Travel-associated infection presenting in Europe (2008-12): an analysis of EuroTravNet longitudinal, surveillance data, and evaluation of the effect of the pre-travel consultation. Lancet Infect Dis. 2015;15(1):55-64.

[2] Thwaites GE, Day NPJ. Approach to Fever in the Returning Traveler. N Engl J Med. 2017;376(18):1798.

[3] Brown JR, Morfopoulou S, Hubb J, Emmett WA, Ip W, Shah D, et al. Astrovirus VA1/HMO-C: an increasingly recognized neurotropic pathogen in immunocompromised patients. Clin Infect Dis. 2015;60(6):881-8.

[4] Lopez-Labrador FX, Brown JR, Fischer N, Harvala H, Van Boheemen S, Cinek O, et al. Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure. J Clin Virol. 2020;134:104691.

[5] Wylie TN, Wylie KM, Herter BN, Storch GA. Enhanced virome sequencing using targeted sequence capture. Genome Res. 2015;25(12):1910-20.

[6] Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, et al. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. mBio. 2015;6(5):e01491-15.

[7] Carbo EC, Buddingh EP, Karelioti E, Sidorov IA, Feltkamp MCW, Borne P, et al. Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics. J Clin Virol. 2020;130:104566.

[8] Carbo EC, Sidorov IA, Zevenhoven-Dobbe JC, Snijder EJ, Claas EC, Laros JFJ, et al. Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness. J Clin Virol. 2020;131:104594.

[9] Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019;20(6):341-55.

[10] Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. N Engl J Med. 2019;380(24):2327-40.

[11] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-3.

[12] Neumayr A, Munoz J, Schunk M, Bottieau E, Cramer J, Calleri G, et al. Sentinel surveillance of imported dengue via travellers to Europe 2012 to 2014: TropNet data from the DengueTools Research Initiative. Euro Surveill. 2017;22(1).

[13] van Boheemen S, van Rijn AL, Pappas N, Carbo EC, Vorderman RHP, Sidorov I, et al. Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients. J Mol Diagn. 2020;22(2):196-207.

[14] van Rijn AL, van Boheemen S, Sidorov I, Carbo EC, Pappas N, Mei H, et al. The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease. PLoS One. 2019;14(10):e0223952.

[15] Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics. 2019;35(5):871-3.

[16] Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721-9.

[17] Jerome H, Taylor C, Sreenu VB, Klymenko T, Filipe ADS, Jackson C, et al. Metagenomic next-generation sequencing aids the diagnosis of viral infections in febrile returning travellers. J Infect. 2019;79(4):383-8.

[18] Hunsperger EA, Munoz-Jordan J, Beltran M, Colon C, Carrion J, Vazquez J, et al. Performance of Dengue Diagnostic Tests in a Single-Specimen Diagnostic Algorithm. J Infect Dis. 2016;214(6):836-44.

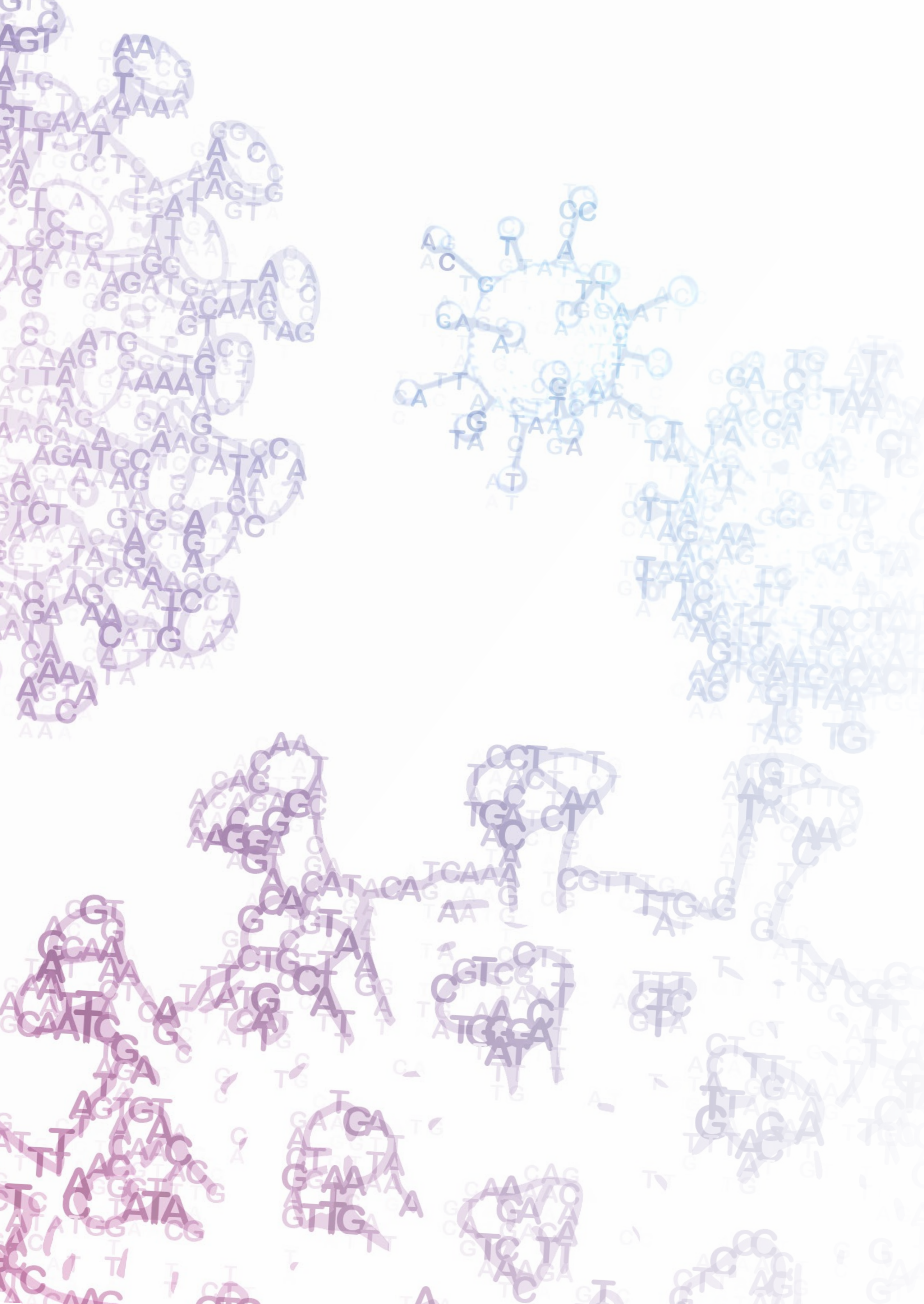
[19] Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013;496(7446):504-7.

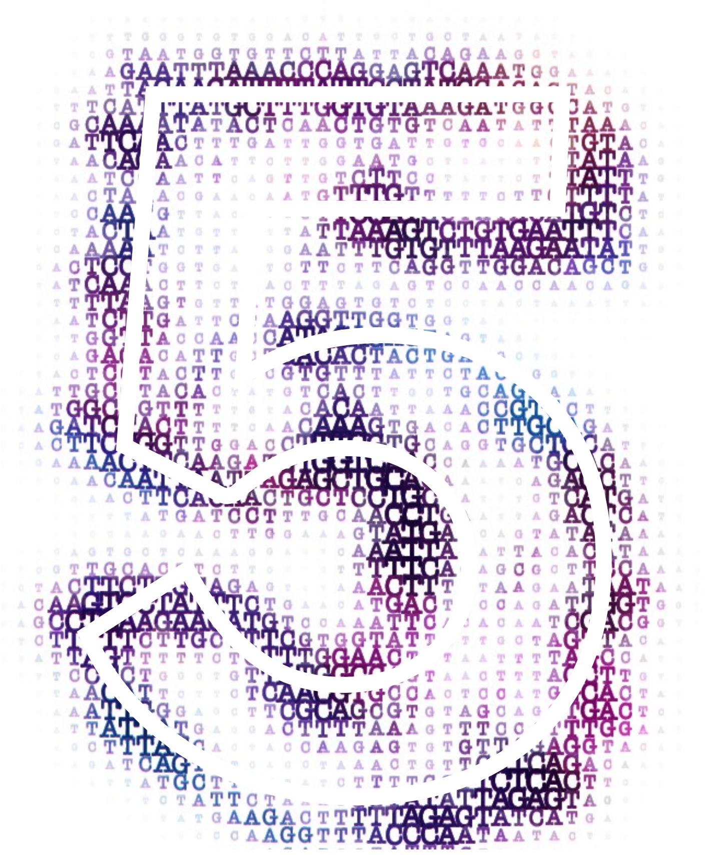
[20] Soo KM, Khalid B, Ching SM, Chee HY. Meta-Analysis of Dengue Severity during Infection by Different Dengue Virus Serotypes in Primary and Secondary Infections. PLoS One. 2016;11(5):e0154760.

[21] Bhattarai N, Stapleton JT. GB virus C: the good boy virus? Trends Microbiol. 2012;20(3):124-30.

[22] Lauck M, Bailey AL, Andersen KG, Goldberg TL, Sabeti PC, O'Connor DH. GB virus C coinfections in west African Ebola patients. J Virol. 2015;89(4):2425-9.

[23] Kiselev D, Matsvay A, Abramov I, Dedkov V, Shipulin G, Khafizov K. Current Trends in Diagnostics of Viral Infections of Unknown Etiology. Viruses. 2020;12(2).





Chapter 5 Performance of five metagenomic classifiers for virus pathogen detection using respiratory samples from a clinical cohort

Ellen C. Carbo[a,*], Igor A. Sidorov[a], Anneloes L. van Rijn-Klink[a], Nikos Pappas[b,1], Sander van Boheemen[a,2], Hailiang Mei[b], Pieter S. Hiemstra[c], Tomas M. Eagan[d], Eric C.J. Claas[a], Aloys C.M. Kroes[a], Jutte J.C. de Vries[a]

a Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands
b Sequencing Analysis Support Core, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
c Department of Pulmonology, Leiden University Medical Center, Leiden, The Netherlands
d Department of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway
1 Present affiliation: Theoretical Biology and Bioinformatics, Department of Biology, Science for Life, Utrecht University, Utrecht, The Netherlands
2 Present affiliation: Department of Viroscience, Erasmus Medical Center, Rotterdam, The Netherlands

Published in Pathogens

Abstract

Viral metagenomics is increasingly applied in clinical diagnostic settings for detection of pathogenic viruses. While several benchmarking studies have been published on the use of metagenomic classifiers for abundance and diversity profiling of bacterial populations, studies on the comparative performance of the classifiers for virus pathogen detection are scarce. In this study, metagenomic data sets (n = 88) from a clinical cohort of patients with respiratory complaints were used for comparison of the performance of five taxonomic classifiers: Centrifuge, Clark, Kaiju, Kraken2, and Genome Detective. A total of 1144 positive and negative PCR results for a total of 13 respiratory viruses were used as gold standard. Sensitivity and specificity of these classifiers ranged from 83 to 100% and 90 to 99%, respectively, and was dependent on the classification level and data pre-processing. Exclusion of human reads generally resulted in increased specificity. Normalization of read counts for genome length resulted in a minor effect on overall performance, however it negatively affected the detection of targets with read counts around detection level. Correlation of sequence read counts with PCR Ct-values varied per classifier, data pre-processing (R^2 range 15.1–63.4%), and per virus, with outliers up to 3 $\log_{10}$ reads magnitude beyond the predicted read count for viruses with high sequence diversity. In this benchmarking study, sensitivity and specificity were within the ranges of use for diagnostic practice when the cut-off for defining a positive result was considered per classifier.

Keywords:

viral metagenomics; bioinformatics; pathogen detection; next-generation sequencing

1. Introduction

In the era of next-generation sequencing (NGS), clinical metagenomics, the analysis of all microbial genetic material in clinical samples, is being introduced in diagnostic laboratories and revolutionizing the diagnostics of infectious diseases [1,2,3,4]. As opposed to running a series of pathogen targeted diagnostic PCR assays to identify suspected pathogens, one single metagenomic run enables the detection of all potential pathogens in a clinical sample [5,6]. The use of this method, also known as shotgun high-throughput sequencing, has resulted in the detection of several pathogens missed by current routine diagnostic procedures [1,7]. For a large part the clinical application of metagenomic sequencing for pathogen detection has focused on patients with encephalitis [1,8,9,10,11,12]. However, patients with clinical syndromes suspected from an infectious disease but with negative conventional test results are increasingly considered as candidates for metagenomic testing. With sequencing costs decreasing and the significance of detection of unexpected, novel viruses being underscored by the currently pandemic SARS-CoV-2 [13], metagenomics is increasingly moving towards implementation in diagnostic laboratories.

Performance testing is typically part of the implementation procedure in diagnostic laboratories to ensure the quality of diagnostic test results. Accurate bioinformatic identification of viral pathogens depends on both the classification algorithm and the database [14,15,16]. Metagenomic sequencing in the past has been mainly oriented at profiling of bacterial genomes in the context of microbiome comparisons in research settings, and most bioinformatic tools currently available have been designed for that specific purpose [17,18]. Some of the previously bacterial oriented classifiers are now being used for other domains, including viruses. However, viral metagenomics for pathogen detection has specific challenges such as the low abundancy of viral sequences for some targets, and incomplete or inaccurate reference sequences. The high diversity of viral sequences due to the high mutation rate of RNA viruses further complicates accurate detection and identification [19]. While the number of benchmarking studies published on the use of metagenomic classifiers for bacterial abundancy profiling is increasing, studies on the performance of classifiers for virus pathogen detection remain scarce. Publications on the performance of the computational analysis of viral metagenomics are usually limited to in silico analysis of artificial sequence data [14,20,21] or mock samples [22,23]. Though both sensitivity and specificity can be deduced when using simulated datasets, they usually do not represent the complexity of data sets from clinical samples which typically contain sequences from wet lab reagents that have been referred to as the 'kitome' [22,24,25]. These factors can affect the sensitivity and specificity of the overall procedure and may result in incorrect diagnoses. In contrast, performance studies that use real-world samples are usually hindered by the huge number of negative metagenomic findings in the absence of gold standard results for validation. Therefore, the performance parameters typically reported are recall (sensitivity), precision (positive predictive value), and F1 (the harmonic mean of recall and precision); while specificity is usually not assessed because negative findings by metagenomics are poorly defined.

Here, we perform a comparison of five taxonomic classifiers: Centrifuge [26], Clark [18], Kaiju [27], Kraken 2 [28], and Genome Detective [29]. The classifiers were tested using metagenomic shotgun sequencing data obtained from a cohort of chronic obstructive pulmonary disease patients (COPD) with a clinical exacerbation and therefore suspected of a respiratory infection. For these samples, 1144 PCR test results were used as gold standard to infer both sensitivity and specificity of the classifiers. For each classifier, we present appropriate benchmark scores for virus classification in the diagnostic setting.

2. Materials and Methods

2.1. Clinical Samples and PCR Results

Clinical respiratory samples were used to obtain metagenomic data sets. In total 88 nasal washings were taken from 63 patients with COPD suspected for respiratory infection as previously described [30]. Each sample was tested using a respiratory PCR panel resulting in 1144 real-time positive and negative PCR results for 13 viral respiratory targets as previously described [30]. The respiratory viruses addressed by this respiratory panel and cohort prevalence are shown in Table 1.

2.2. Metagenomic Next-Generation Sequencing (mNGS)

The metagenomic datasets used for comparison were generated as described before [30]. In short, clinical samples were spiked with equine arteritis virus (EAV) and phocine herpesvirus 1 (PhHV-1), as internal positive controls for RNA and DNA detection per sample, throughout the entire workflow. Negative and positive washings were used as respectively environmental and positive run controls. Subsequently, extraction of nucleic acids was performed using the Magnapure 96

DNA and Viral NA Small volume extraction kit on the MagnaPure 96 system (Roche, Basel, Switzerland). Library preparation was performed utilizing the NEBNext Ultra II Directional RNA Library prep kit for Illumina (New England Biolabs, Ipswich, MA, USA) using single, unique adaptors and a protocol optimized for processing RNA and DNA simultaneously in a single tube [25]. Sequencing was performed on an Illumina NextSeq 500 sequencing system (Illumina, San Diego, CA, USA) at GenomeScan BV (Leiden, The Netherlands), obtaining approximately 10 million 150 bp paired-end reads per sample.

2.3. Pre-Processing of Data

To exclude variability based on pre-processing procedures, the identical procedure was followed prior to analysis of the sequence data by all classifiers in the current comparison. Illumina 150 bp paired-end sequence reads were demultiplexed by standard Illumina software followed by trimming, adapter clipping, and filtering of low-complexity reads using Trimmomatic [v. 0.36] [31]. This was performed for all classifiers, regardless of quality filtering options that have been previously used in combination with specific classifiers in literature. Human reads were excluded after mapping them to the human genome GRCh38 [32] using Bowtie2 with standard settings [33]. Unmapped reads were used for further analysis for the classification tests excluding human reads.

2.4. Metagenomic Classifiers

Bioinformatic metagenomics tools designed for taxonomic classification were selected for benchmarking based on the following criteria: applicable for viral metagenomics for pathogen detection; available either as download or webserver; and it is either widely used or showed potential of diagnostics implementation in the future. Some tools considered were excluded due to lack of support or details on how to use the tool, or non-functioning webservers. An overview of characteristics of the selected classifiers can be found in Table 2.

Table 1. Overview of respiratory PCR panel targets and their test results.

PCR target viruses	Family	Genus	Species	Alternative naming	# PCR positive samples	# PCR negative samples	PCR Ct-values (range)
HRV	Picorna-viridae	Enterovirus	Rhinovirus A, B, C, Enterovirus D		14	74	19-38
PIV1, PIV3	Paramyxo-viridae	Respiro-virus	Human respiro-virus 1	Human parainfluenza virus 1	-	88	-
			Human respiro-virus 3	Human parainfluenza virus 3	2	86	26-36
PIV2, PIV4	Paramyxo-viridae	Ortho-rubulavirus	Human orthorubu-lavirus 2	Human parainfluenza virus 2	-	88	-
			Human orthorubu-lavirus 4	Human parainfluenza virus 4	1	87	24
INF	Orthomyxo-viridae	Alpha-influenzavirus	Influenza A virus Influenza B virus		3 -	85 88	29-36 -
ACoV	Corona-viridae	Alpha-coronavirus	Human coronavirus NL63 Human coronavirus 229E		2 -	86 88	32 -
BCoV	Corona-viridae	Betacorona-virus	Human corona-virus HKU1, Betacoronavirus 1; Human coronavirus OC43		2	86	27
HMPV	Pneumo-viridae	Metapneu-movirus	Human metapneumo-virus		-	88	-
RSV	Pneumo-viridae	Orthopneu-movirus	Human orthopneumo-virus		-	88	-
Total			**Total PCR results: 1,144 (13 targets tested in 88 samples)**		**24**	**1,120**	**19-38**

Table 2. **Overview of characteristics of the classifiers evaluated.**

	Centrifuge [26]	**Clark** [18]	**Kaiju** [27]	**Kraken 2** [28]	**Genome Detective** [29]
License	**Open source**	**Open source**	**Open source**	**Open source**	**Commercial/ free to use web application**
Version	1.0.4	1.2.6.1	1.7.3	2.0.8-beta	1.126
Sequencing technology compatibility	Short/long reads	Short/long reads	Short/long reads	Short/long reads	Short reads (long reads experimentally)
Pre-processing	No	No	No	No	Yes
Type of alignment	NT	NT	AA	NT	NT/AA (DIAMOND [38]) including de novo assembly
Algorithm characteristics	Exact matches of 22 bp with target with default 5 labels per sequence, LCA optional	Exact matches of 31 bp with target with highest number of hits	Maximum exact matches (MEM) of AA, up to 5 mismatched optional*. LCA in case of multiple hits	Exact matches of 35 bp. LCA in case of multiple hits	Combined results of NT and AA hits based on scoring. LCA in case of multiple hits
Database (compression)	Compressed index NT database of only unique sequences	Compressed index NT database of only unique sequences	No compression, AA database	Compressed index NT database	No compression, viral subset of Swiss-Prot UniRef90 protein database

NT; nucleotide, AA; amino acid; LCA, lowest common ancestor

* Greedy-5 mode was used in the current study

2.5. Reference Database

For comparison of classification performance, a single database was used as starting point for the classifiers Centrifuge, Clark, Kaiju, and Kraken 2: viral genomes from NCBI/RefSeq [34] (downloaded on 27 December 2020). Genome Detective was used as a service, and it uses its own database that was generated on 3 March 2020 (version 1.130) by Genome Detective.

2.6. Metagenomic Classifiers and Characteristics

2.6.1. Centrifuge

Classification with Centrifuge (version 1.0.4) [26] is based on exact matches of at least 22 base pair nucleotide sequences with the reference index, using k-mers of user-defined length. Centrifuge by default allows five classification labels per sequence read. For a realistic comparison, in the current study, this setting was adapted to maximum one label per sequence (the lowest common ancestor) to mimic results of Kraken2 and other classifiers where only one label per sequence read is given. Preceding classification, Centrifuge builds small reference indexes based on adapted versions of the Burrows–Wheeler transform (BWT) [35] and the Ferragina–Manzini (FM) index [36] resulting in a compressed index of only unique genomic sequences.

2.6.2. Clark

Clark (version 1.2.6.1) [18] is a taxonomic classifier based on reduced k-mers using nucleotide-level classification. It uses a compressed index database containing unique target specific k-spectrum of target sequences. For the current comparison the default execution mode was used.

2.6.3. Kaiju

Kaiju (version 1.7.3) [27] is a taxonomic classifier that assigns sequence reads using amino acid-level classification. Sequence reads are translated into six possible open reading frames and split into fragments according to the detected stop codons. Classification with Kaiju can be performed using two settings, both based on an adjusted backward alignment search algorithm of BWT [35]. For the current comparison study, the greedy mode was used providing high sensitivity because it allows up to five mismatches to further increase the highest scoring matches. In this mode Kaiju assesses six possible ORF’s using the amino acid scores of Blosum62 [37] to obtain the highest scoring match.

2.6.4. Kraken 2

Kraken 2 (version 2.0.8-beta) [28] is a classifier designed to improve the large memory requirements of the former version of Kraken [17], resulting in a reduction of in general 85% of the size of the index database. Kraken 2 uses a probabilistic, compact hash table to map minimizers to the lowest common ancestors (LCA), and stores only minimizers from the reference sequence library in its index reference [28].

2.6.5. Genome Detective

Genome Detective [29] is a commercially available bioinformatic pipeline that includes the entire workflow from automated quality control, de novo assembly of reads and classification of viruses. After automated adapter trimming and filtering low-quality reads using Trimmomatic [31], viral reads are selected based on Diamond [38] protein alignment using as reference protein sequences from Swissprot Uniref 90 [39]. Viral reads are sorted in buckets, after which all sequences in one bucket are de novo assembled into contigs using SPAdes [40] or metaSPAdes [41]. Subsequently, contigs are processed by BLASTx and BLASTn [42] against databases containing NCBI Refseq [34] sequences and some additional virus sequences. Potential hits represented by the contigs are assigned to individual species using the Advanced Genome Aligner [43], and coverage the viral genomes is calculated. For analysis using Genome Detective sequence reads were first pre-processed with Trimmomatic [31] manually, similar for other tools (see Pre-processing of data), prior to automated filtering by the Genome Detective pipeline.

2.7. Performance, Statistical Analysis, and ROC

Sensitivity and specificity were calculated for the classifiers based on the application of PCRs (designed for detection of 13 targets) for 88 samples with 24 PCR positive and 1120 PCR negative results. Receiver Operating Characteristic (ROC) curves were generated for results of classification at species, genus, and family levels, by varying the number of sequence-read counts used as cut-off for defining a positive result (resolution: 1000 steps from one read to the maximum number of sequence reads for each PCR target per sample). Area under the curve (AUC), the ROC distance to the closest error-free point (0,1, informedness) curve, positive and negative predictive values were calculated. Furthermore, correlation (R^2) of sequence read counts with PCR cycle threshold (Ct) value were analyzed.

3. Results

3.1. Performance: Sensitivity, Specificity, and ROC

The performance of the selected taxonomic classifiers Centrifuge, Clark, Kaiju, Kraken 2, and Genome Detective for metagenomic virus pathogen detection was assessed using datasets from 88 respiratory samples with 24 positive and 1,120 negative PCR results available as gold standard. To exclude variability based on different default databases provided with the classifiers, a single database of reference genome sequences was used in combination with a standardized dataset for all classifiers. Raw NGS reads were filtered and classified, both prior and after the exclusion of human sequence reads, and after exclusion of human reads combined with normalization of reads based on the target viral genome length. ROC curves are shown for all classifiers, for assignments at species, genus and family level for the NGS data in Figure 1, and Supplementary Table S1. Detection parameters (ROC distance to the upper left corner of the graph, sensitivity and selectivity, and AUC) at three taxonomic levels calculated for the NGS data, before and after exclusion of human reads, with or without normalization of assigned reads by corresponding genome sequence lengths are additionally shown in Figure 2. Overall, sensitivity, specificity, and AUC ranged from 83 to 100%, 90 to 99%, and 91 to 98%, respectively, and varied per level of taxonomic classification, per classifier, and with the exclusion of human reads prior to classification. Classification at species and genus levels tended to result in lower sensitivity and higher ROC distances, but higher selectivity when compared with family level classification, for most of the classifiers evaluated. Extraction of human sequence reads prior to classification resulted in comparable sensitivity at all levels of assignment for all classifiers except CLARK for which sensitivity plummeted at species and genus levels. Selectivity was mainly increased after extraction of human reads, for classification at all levels, except for Kaiju and Kraken2, for which decreased selectivity values at family level were observed. Extraction of human reads reduced the differences in selectivity between the classifiers that were observed at genus and family level prior to extraction. The ROC distances were overall smallest, and the AUC highest, when using amino-acid based classifier Kaiju, the latter at species and family levels and was comparable with Kraken2 at genus level. Normalization of assigned read counts by corresponding genome length resulted in minor changes in performance when considering 1 read as the threshold for defining positive results. Sensitivity was dramatically reduced to 13–33% at species level after read normalization when a threshold of 10 reads was applied, while sensitivity was 75–88% without read normalization in combination with a threshold of 10 reads, (Supplementary Table S1). This indicates that normalization of reads can negatively affect the detection of targets with read counts around detection level.

Overall, Kaiju outperformed all classifiers when ROC distance, AUC, and sensitivity were considered, but had consistently lower selectivity values than Centrifuge and Genome Detective.

In this patient cohort, with an incidence of 21% (24/88 samples) of respiratory viruses, the positive and negative predictive values at species levels were 42–67% and 99–100%, respectively (see Supplementary Table S1).

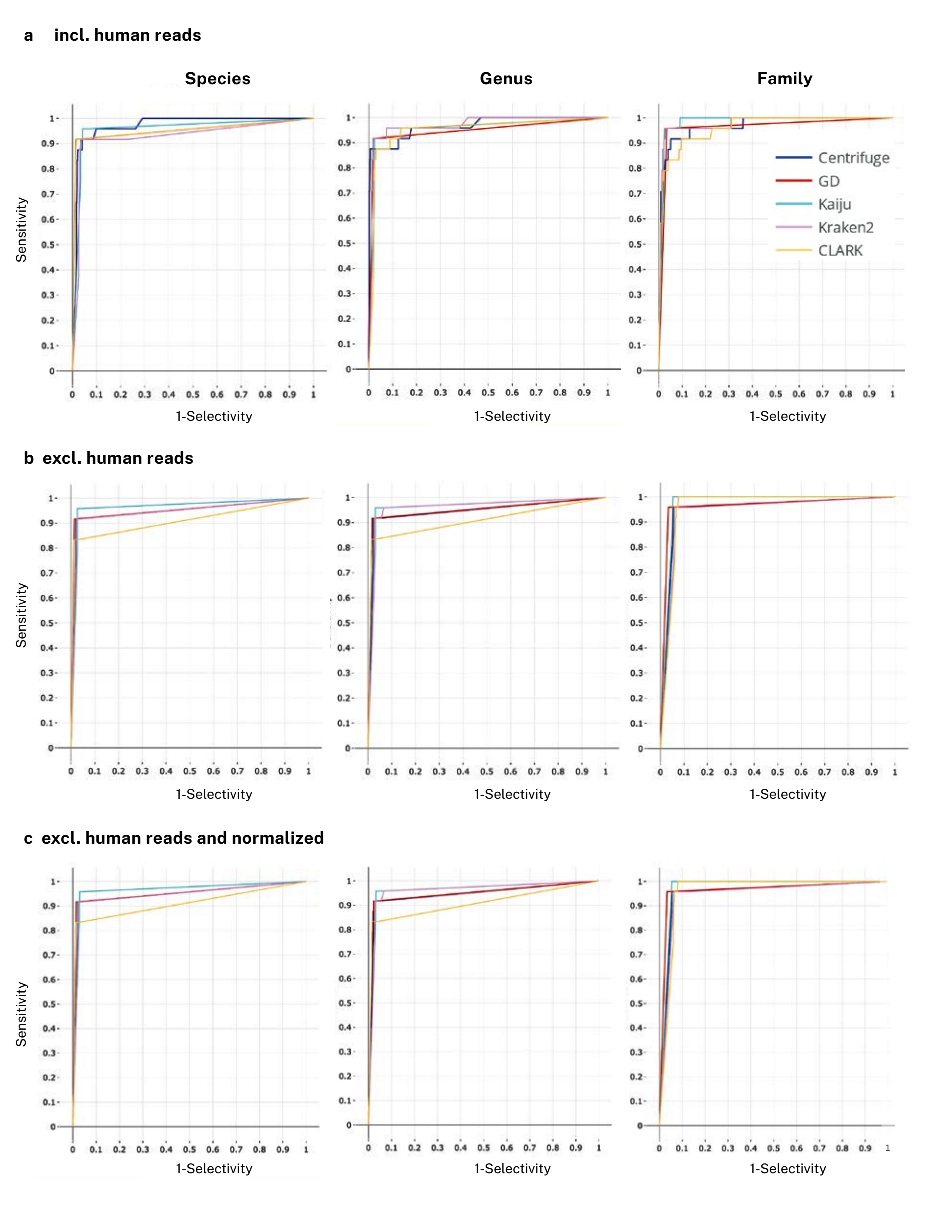


Figure 1. ROC curves.

Calculated based on reads of taxonomic assignment at three. taxonomic levels (species, genus, and family) by the five classifiers, based on PCR-targets, (a), without extraction of human reads and (b), after extraction of human reads, (c), after extraction of human reads and normalization of reads by corresponding genome lengths (resolution of 1000 steps from one read to the maximum number of sequence reads for each PCR target per sample).

3.2. Correlation Read Counts and Ct-Values

The correlation between sequence read counts at Ct-value for the corresponding PCR target viruses for all classifiers is shown in Figure 3 and Supplementary Table S2. Correlation (R2, %), linear regression slope and intercept varied per virus species, per taxonomic classifier, and was dependent on the extraction of human reads. Correlation ranged from 15.1% for CLARK (no exclusion of human reads, species level) to 62.7% for Kaiju-based classification at species level (after exclusion of human reads with normalization of assigned reads by corresponding genome sequence lengths). The most consistent results (when comparing R^2 prior and after human reads exclusion, and after normalization) were demonstrated by Kaiju and Genome Detective with overall outperformance of Kaiju classifier at all classification levels (61.8–62.7% versus 42.3–43.9% for Centrifuge). Reads assigned to rhinoviruses were most common outliers in relation to Ct-value and varied up to 3 $\log_{10}$ reads difference from the predicted read count (LR), possibly resulting from their high divergence within species. This was in contrast to read counts of other viruses (for example influenza viruses), which were closer to the predicted correlation line. Extraction of human sequence reads resulted in an increase in R^2 for CLARK classifier at species and family level, a decrease for Centrifuge and Kraken at all levels, and resulted in minor changes for amino acid-based classifiers Genome Detective and Kaiju at all levels. Decrease in absolute or relative number of total reads after pre-processing (extraction of human reads in combination with normalization of assigned reads by corresponding genome lengths) led to a decrease in intercept values for all classifiers.

These data support that a more accurate taxonomic classification assists semi-quantitative performance of metagenomic classification tools.

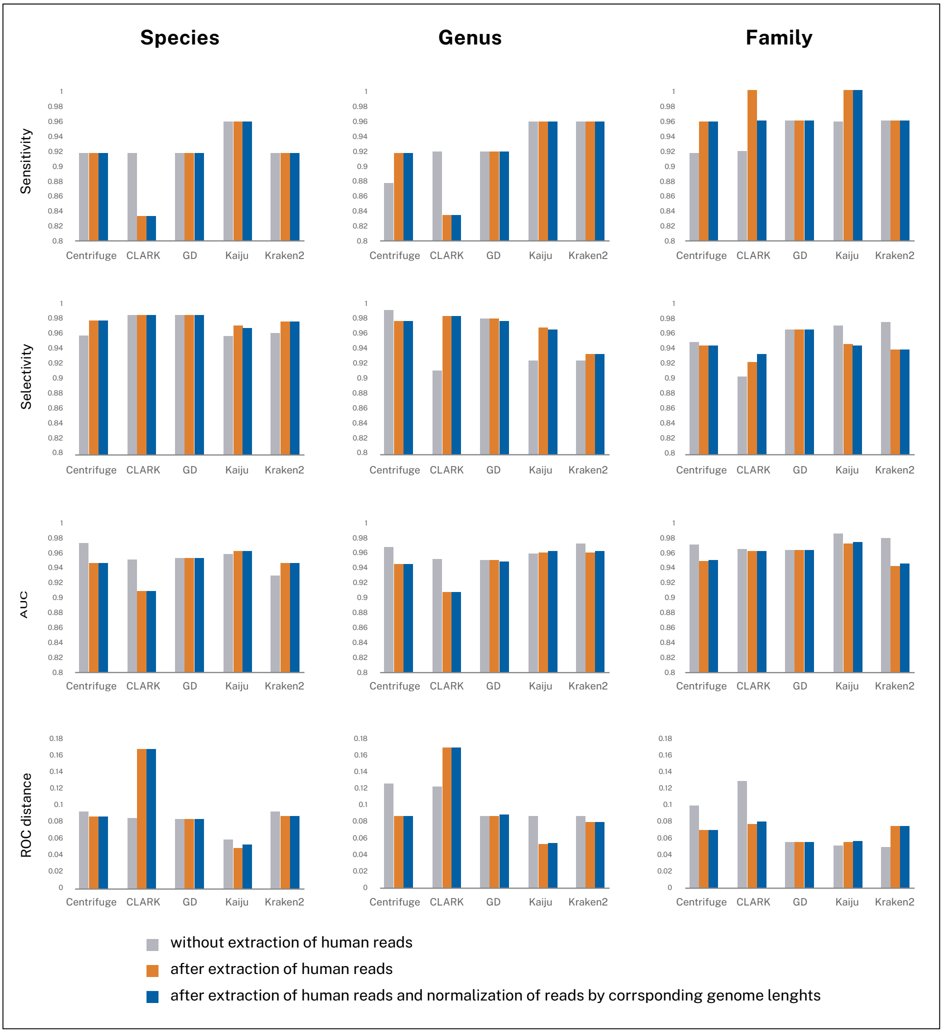


Figure 2. Sensitivity, selectivity, AUC, and ROC distance.

Calculated based on assignment at three taxonomic levels (species, genus, and family) by the five classifiers for three types of pre-processing of the NGS datasets, a, without extraction of human reads and b, after extraction of human reads, c, after extraction of human reads and normalization of reads by corresponding genome lengths.

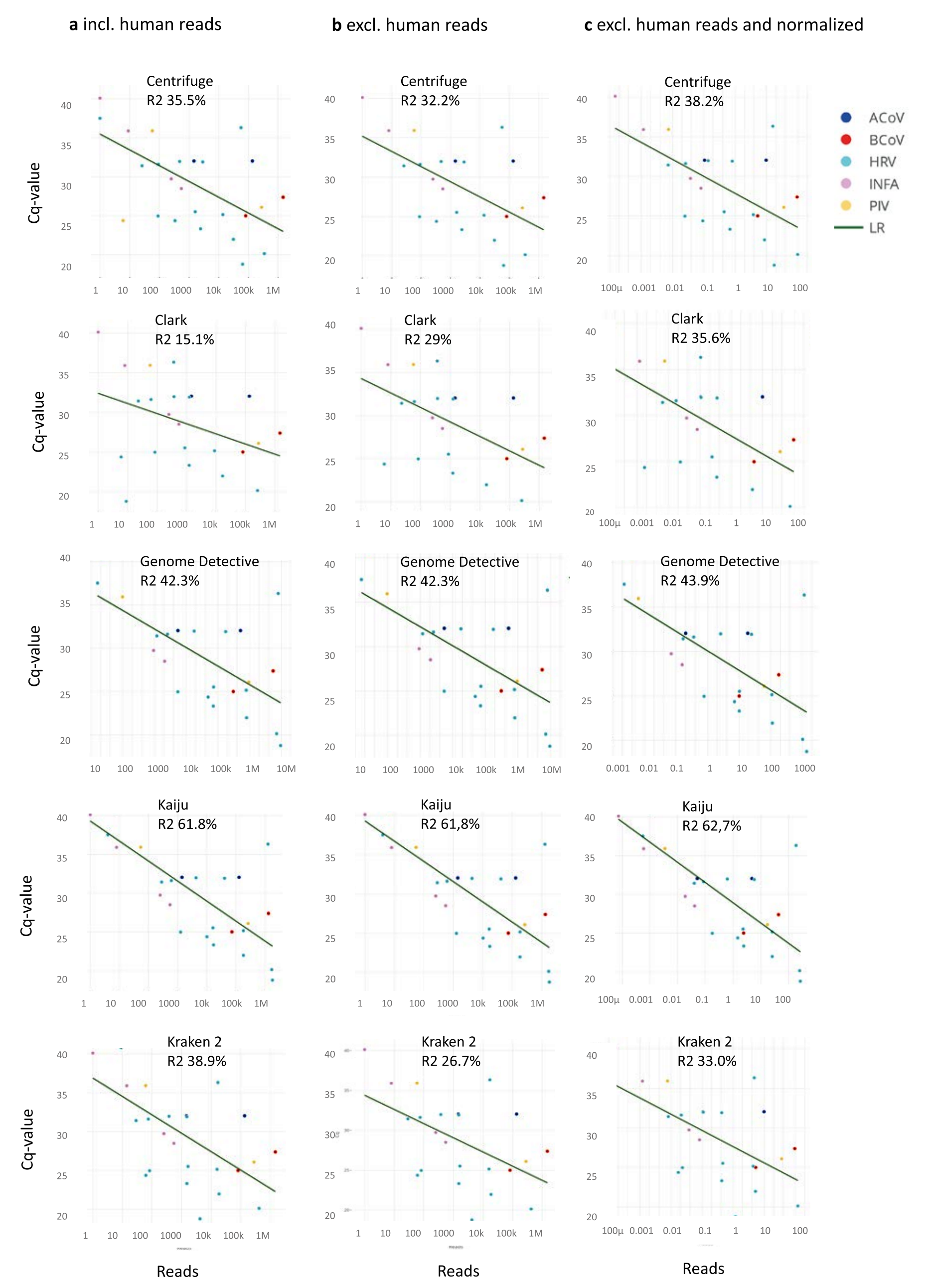


Figure 3. Correlation between the number of sequence reads assigned (species level) and Ct-values of virus-specific PCRs for the five taxonomic classifiers evaluated, (a), without extraction of human reads and (b), after extraction of human reads, (c), after normalization of reads by corresponding genome lengths.

4. Discussion

In this study, we compared the performance of five taxonomic classification tools for virus pathogen detection, using datasets from well-characterized clinical samples. In contrast to previously reported comparisons with datasets from real samples, both sensitivity and specificity could be assessed using a unique set of 1144 PCR results as gold standard. A uniform database was created to exclude variability based on differences in availability of genomes in databases provided with the classifiers. In general, sensitivity and specificity were within ranges applicable to diagnostic practice. Exclusion of human reads generally resulted in increased specificity. Normalization of read counts for genome length negatively affected the detection of targets with read counts around detection level. The correlation of sequence read counts with PCR Ct-values was highest for viruses with relatively lower sequence diversity.

Previous studies have benchmarked metagenomic profilers, mainly for the use of bacterial profiling and DNA-to-DNA and DNA-to-protein classification methods were among the best-scoring methods in comparison with DNA-to-marker (16S) methods [22,27,44,45,46,47,48]. In a study with simulated bacterial datasets comparing the performance of CLARK, Kraken and Kaiju, sensitivity and precision were 75% and 95% and decreased when a lower number of reference genomes was available for the specific target [27]. As the same reference database was used by all classifiers in this study, the only determining factors would be the index database built from the reference database and the classification algorithm. DNA-to-DNA methods have been applied in hundreds of published microbiome studies (e.g., Kraken: 1438 citations; Kraken 2: 204 citations, by March 2021, according to their official websites [48]). Centrifuge was designed as a follow-up of Kraken with enhanced features, though misclassifications have also been reported in a comparison with simulated datasets [22]. DNA-to-protein methods are generally more sensitive to novel and highly variable sequences due to lower mutation rates of amino acid compared to nucleotide sequences [22,27] as was seen in our study when classifying rhinoviruses by Kaiju. The difference was especially visible in genera with limited availability of genomes in reference databases [27].

Misclassification of human genomic sequence reads has been reported for most DNA classifiers [22]. Protein-based classifiers had higher misclassification ranges of human genome sequences (up to 15%), partially due to the larger number of target sequences in their default databases [22]. Inclusion of the human genome in the reference database, which is by default the case for Centrifuge and KrakenUniq [49] reduced the rate of misclassification to negligible [22]. This finding is supported in our study, as exclusion of human sequence reads prior to classification reduced misclassifications for all classifiers. In general, reduction of false-positive hits can be achieved by assembly of sequences (for example, by Genome Detective), thus reducing the number of hits based on short nucleotide sequences used by k-mer based methods. Inclusion of genome coverage of mapped reads, as adopted by Genome Detective and KrakenUniq [49], also can reduce false-positive hits.

One of the strengths of this study, the use of one single wet lab and sequencing procedure, in order to enable comparison of the bioinformatic analyses, is also a limitation of the study. The sensitivity and specificity results will likely vary when the classifiers are used in combination with a different wet lab methodology. Therefore, no conclusions can be drawn on the absolute numbers, sensitivity and specificity, of other workflows that include the classifiers, since every step in the entire workflow can influence the overall performance.

To our knowledge, a limited number of studies on the benchmarking of tools for viral metagenomics for pathogen detection have been published. In a Switzerland-wide ring trial based on spiked plasma samples, median F1 scores ranged from 70 to 100% for the different pipelines, though since the entire workflow was analyzed, and thus no conclusions on specific classifiers could be drawn [15]. A series of tools and programs were analyzed in a COMPARE virus proficiency test using a single in silico dataset [14]. For Kraken discrepant classification results that were observed, this was likely due to differences in the databases used by the participants. A recent European benchmark of 13 bioinformatic pipelines currently in use for metagenomic virus diagnostics used datasets from clinical samples [16] analyses using Centrifuge and Genome Detective software resulted in sensitivities of 93% and 87%, respectively.

In conclusion, sensitivity and specificity of the classifiers evaluated in this study was within the ranges that may be applied in clinical diagnostic settings. Performance testing for viral metagenomics for pathogen detection is intrinsically different from benchmarking of bacterial profiling and should incorporate parameters that are inherent to clinical diagnostic use such as specificity calculations, sensitivity for divergent viruses and variants, and importantly, a determined cut-off for defining a positive result for each workflow. Taking these factors into account during validation and implementation of viral metagenomics for pathogen detection contributes to optimal performance and applicability in clinical diagnostic settings.

Supplementary Materials

The following supporting information can be downloaded at:
Https://www.mdpi.com/article/10.3390/pathogens11030340/s1

Table S1: Overview of performance characteristics for the classifiers benchmarked in this study, at species, genus, and family level. Table S2: Correlation between the number of sequence reads assigned and Ct-values of virus-specific PCRs, for the five taxonomic classifiers evaluated, without extraction of human reads, after extraction of human reads, and after normalization of reads by corresponding genome size.

Funding

This research received no external funding.

Institutional Review Board Statement

Ethical approval for metagenomic sequencing of the clinical cohorts was obtained from the medical ethics review committee of the Leiden University Medical Center, The Netherlands, (CME number B16.004 and date of approval 30 May 2016).

Data Availability Statement

NGS data used in this study have been submitted (after removal of human reads) to the NCBI's Sequence Read Archive (http://www.ncbi.nlm.nih.gov; accession number SRX6713943-SRX6714030).

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] M. R. Wilson et al., ‘Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis’, N. Engl. J. Med., vol. 380, no. 24, pp. 2327–2340, Jun. 2019, doi: 10.1056/NEJMoa1803396.

[2] F. X. López-Labrador et al., ‘Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure’, J. Clin. Virol., vol. 134, p. 104691, Jan. 2021, doi: 10.1016/j.jcv.2020.104691.

[3] J. J. C. de Vries et al., ‘Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting’, J. Clin. Virol., vol. 138, p. 104812, May 2021, doi: 10.1016/j.jcv.2021.104812.

[4] E. C. Carbo, I. Blankenspoor, J. J. Goeman, A. C. M. Kroes, E. C. J. Claas, and J. J. C. De Vries, ‘Viral metagenomic sequencing in the diagnosis of meningoencephalitis: a review of technical advances and diagnostic yield’, Expert Rev. Mol. Diagn., vol. 21, no. 11, pp. 1139–1146, Nov. 2021, doi: 10.1080/14737159.2021.1985467.

[5] C. Y. Chiu and S. A. Miller, ‘Clinical metagenomics’, Nat. Rev. Genet., vol. 20, no. 6, pp. 341–355, Jun. 2019, doi: 10.1038/s41576-019-0113-7.

[6] W. Gu, S. Miller, and C. Y. Chiu, ‘Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection’, Annu. Rev. Pathol. Mech. Dis., vol. 14, no. 1, pp. 319–338, Jan. 2019, doi: 10.1146/annurev-pathmechdis-012418-012751.

[7] A. Reyes et al., ‘Viral metagenomic sequencing in a cohort of international travellers returning with febrile illness’, J. Clin. Virol., vol. 143, p. 104940, Oct. 2021, doi: 10.1016/j.jcv.2021.104940.

[8] J. R. Brown, T. Bharucha, and J. Breuer, ‘Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases’, J. Infect., vol. 76, no. 3, pp. 225–240, Mar. 2018, doi: 10.1016/j.jinf.2017.12.014.

[9] E. C. Carbo et al., ‘Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics’, J. Clin. Virol., p. 104566, Jul. 2020, doi: 10.1016/j.jcv.2020.104566.

[10] C. Y. Chiu et al., ‘Diagnosis of Fatal Human Case of St. Louis Encephalitis Virus Infection by Metagenomic Sequencing, California, 2016’, Emerg. Infect. Dis., vol. 23, no. 10, pp. 1964–1968, Oct. 2017, doi: 10.3201/eid2310.161986.

[11] M. Christopeit et al., ‘Suspected encephalitis with Candida tropicalis and Fusarium detected by unbiased RNA sequencing’, Ann. Hematol., vol. 95, no. 11, pp. 1919–1921, Nov. 2016, doi: 10.1007/s00277-016-2770-3.

[12] A. W. D. Edridge et al., ‘Novel Orthobunyavirus Identified in the Cerebrospinal Fluid of a Ugandan Child With Severe Encephalopathy’, Clin. Infect. Dis., vol. 68, no. 1, pp. 139–142, Jan. 2019, doi: 10.1093/cid/ciy486.

[13] E. C. Carbo et al., ‘Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness’, J. Clin. Virol., vol. 131, p. 104594, Oct. 2020, doi: 10.1016/j.jcv.2020.104594.

[14] A. Brinkmann et al., ‘Proficiency Testing of Virus Diagnostics Based on Bioinformatics Analysis of Simulated In Silico High-Throughput Sequencing Data Sets’, J. Clin. Microbiol., vol. 57, no. 8, pp. e00466-19, /jcm/57/8/JCM.00466-19.atom, Jun. 2019, doi: 10.1128/JCM.00466-19.

[15] Junier et al., 'Viral Metagenomics in the Clinical Realm: Lessons Learned from a Swiss-Wide Ring Trial', Genes, vol. 10, no. 9, p. 655, Aug. 2019, doi: 10.3390/genes10090655.

[16] De Vries, J.J.; Brown, J.R.; Fischer, N.; Sidorov, I.A.; Morfopoulou, S.; Huang, J.; Munnink, B.B.O.; Sayiner, A.; Bulgurcu, A.; Rodriguez, C.; et al. Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. J. Clin. Virol. 2021, 141, 104908

[17] D. E. Wood and S. L. Salzberg, 'Kraken: ultrafast metagenomic sequence classification using exact alignments', Genome Biol., vol. 15, no. 3, p. R46, 2014, doi: 10.1186/gb-2014-15-3-r46.

[18] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, 'CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers', BMC Genomics, vol. 16, no. 1, p. 236, Dec. 2015, doi: 10.1186/s12864-015-1419-2.

[19] P. Simmonds et al., 'Virus taxonomy in the age of metagenomics', Nat. Rev. Microbiol., vol. 15, no. 3, pp. 161–168, Mar. 2017, doi: 10.1038/nrmicro.2016.177.

[20] S. Nooij, D. Schmitz, H. Vennema, A. Kroneman, and M. P. G. Koopmans, 'Overview of Virus Metagenomic Classification Methods and Their Biological Applications', Front. Microbiol., vol. 9, p. 749, Apr. 2018, doi: 10.3389/fmicb.2018.00749.

[21] A. Escobar-Zepeda et al., 'Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics', Sci. Rep., vol. 8, no. 1, p. 12034, Dec. 2018, doi: 10.1038/s41598-018-30515-5.

[22] S. H. Ye, K. J. Siddle, D. J. Park, and P. C. Sabeti, 'Benchmarking Metagenomics Tools for Taxonomic Classification', Cell, vol. 178, no. 4, pp. 779–794, Aug. 2019, doi: 10.1016/j.cell.2019.07.010.

[23] N. Couto et al., 'Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens', Sci. Rep., vol. 8, no. 1, p. 13767, Dec. 2018, doi: 10.1038/s41598-018-31873-w.

[24] M. Asplund et al., 'Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries', Clin. Microbiol. Infect., vol. 25, no. 10, pp. 1277–1285, Oct. 2019, doi: 10.1016/j.cmi.2019.04.028.

[25] S. van Boheemen et al., 'Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients', J. Mol. Diagn., vol. 22, no. 2, pp. 196–207, Feb. 2020, doi: 10.1016/j.jmoldx.2019.10.007.

[26] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg, 'Centrifuge: rapid and sensitive classification of metagenomic sequences', Genome Res., vol. 26, no. 12, pp. 1721–1729, Dec. 2016, doi: 10.1101/gr.210641.116.

[27] P. Menzel, K. L. Ng, and A. Krogh, 'Fast and sensitive taxonomic classification for metagenomics with Kaiju', Nat. Commun., vol. 7, no. 1, p. 11257, Sep. 2016, doi: 10.1038/ncomms11257.

[28] D. E. Wood, J. Lu, and B. Langmead, 'Improved metagenomic analysis with Kraken 2', Genome Biol., vol. 20, no. 1, p. 257, Dec. 2019, doi: 10.1186/s13059-019-1891-0.

[29] M. Vilsker et al., ‘Genome Detective: an automated system for virus identification from high-throughput sequencing data’, Bioinformatics, vol. 35, no. 5, pp. 871–873, Mar. 2019, doi: 10.1093/bioinformatics/bty695.

[30] A. L. van Rijn et al., ‘The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease’, PLOS ONE, vol. 14, no. 10, p. e0223952, Oct. 2019, doi: 10.1371/journal.pone.0223952.

[31] A. M. Bolger, M. Lohse, and B. Usadel, ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, Bioinformatics, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[32] ‘GRCh38’. [Online]. Available: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/

[33] B. Langmead and S. L. Salzberg, ‘Fast gapped-read alignment with Bowtie 2’, Nat. Methods, vol. 9, no. 4, pp. 357–359, Apr. 2012, doi: 10.1038/nmeth.1923.

[34] N. A. O’Leary et al., ‘Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation’, Nucleic Acids Res., vol. 44, no. D1, pp. D733–D745, Jan. 2016, doi: 10.1093/nar/gkv1189.

[35] Burrows, M.; Wheeler, David J., ‘A block-sorting lossless data compression algorithm.’, vol. Digital Equipment Corporation, no. Technical Report 124, 1994.

[36] P. Ferragina and G. Manzini, ‘Opportunistic data structures with applications’, in Proceedings 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, CA, USA, 2000, pp. 390–398. doi: 10.1109/SFCS.2000.892127.

[37] S. Henikoff and J. G. Henikoff, ‘Amino acid substitution matrices from protein blocks.’, Proc. Natl. Acad. Sci., vol. 89, no. 22, pp. 10915–10919, Nov. 1992, doi: 10.1073/pnas.89.22.10915.

[38] B. Buchfink, C. Xie, and D. H. Huson, ‘Fast and sensitive protein alignment using DIAMOND’, Nat. Methods, vol. 12, no. 1, pp. 59–60, Jan. 2015, doi: 10.1038/nmeth.3176.

[39] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, ‘UniRef: comprehensive and non-redundant UniProt reference clusters’, Bioinformatics, vol. 23, no. 10, pp. 1282–1288, May 2007, doi: 10.1093/bioinformatics/btm098.

[40] A. Bankevich et al., ‘SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing’, J. Comput. Biol., vol. 19, no. 5, pp. 455–477, May 2012, doi: 10.1089/cmb.2012.0021.

[41] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, ‘metaSPAdes: a new versatile metagenomic assembler’, Genome Res., vol. 27, no. 5, pp. 824–834, May 2017, doi: 10.1101/gr.213959.116.

[42] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, ‘Basic local alignment search tool’, J. Mol. Biol., vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[43] K. Deforche, ‘An alignment method for nucleic acid sequences against annotated genomes’, Bioinformatics, preprint, Oct. 2017. doi: 10.1101/200394.

[44] K. Mavromatis et al., ‘Use of simulated data sets to evaluate the fidelity of metagenomic processing methods’, Nat. Methods, vol. 4, no. 6, pp. 495–500, Jun. 2007, doi: 10.1038/nmeth1043.

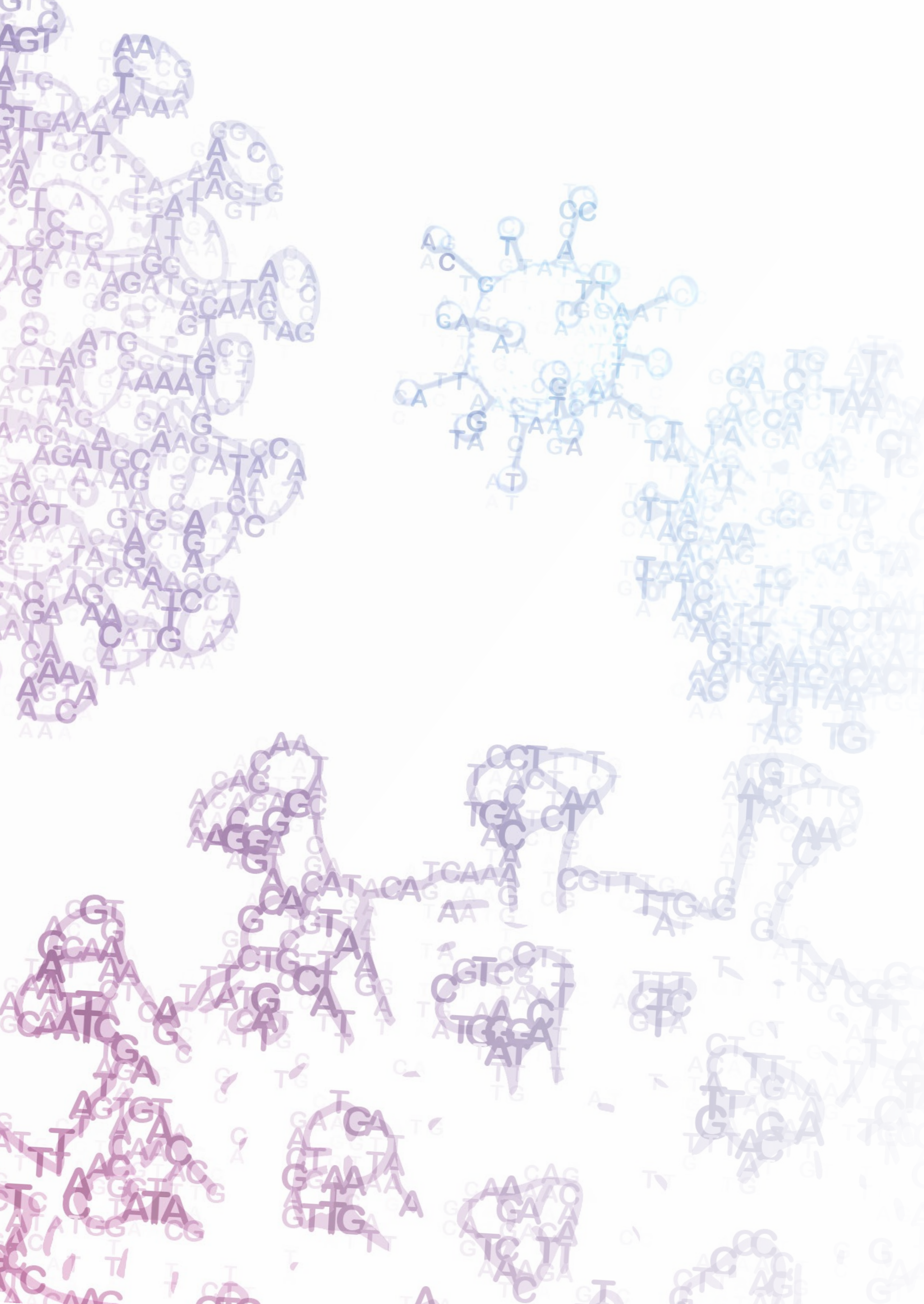
[45] F. Meyer, A. Bremges, P. Belmann, S. Janssen, A. C. McHardy, and D. Koslicki, ‘Assessing taxonomic metagenome profilers with OPAL’, Genome Biol., vol. 20, no. 1, p. 51, Dec. 2019, doi: 10.1186/s13059-019-1646-y.

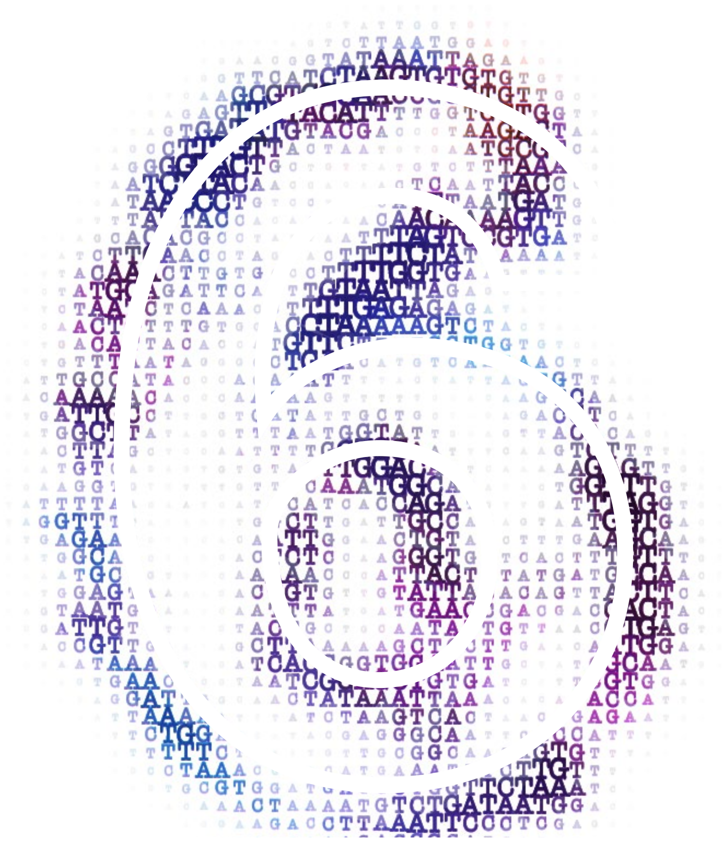
[46] A. Sczyrba et al., ‘Critical Assessment of Metagenome Interpretation — a benchmark of metagenomics software’, Nat. Methods, vol. 14, no. 11, pp. 1063–1071, Nov. 2017, doi: 10.1038/nmeth.4458.

[47] A. B. R. McIntyre et al., ‘Comprehensive benchmarking and ensemble approaches for metagenomic classifiers’, Genome Biol., vol. 18, no. 1, p. 182, Dec. 2017, doi: 10.1186/s13059-017-1299-7.

[48] Z. Sun et al., ‘Challenges in benchmarking metagenomic profilers’, Nat. Methods, vol. 18, no. 6, pp. 618–626, Jun. 2021, doi: 10.1038/s41592-021-01141-3.

[49] F. P. Breitwieser, D. N. Baker, and S. L. Salzberg, ‘KrakenUniq: confident and fast metagenomics classification using unique k-mer counts’, Genome Biol., vol. 19, no. 1, p. 198, Dec. 2018, doi: 10.1186/s13059-018-1568-0.





Chapter 6 Longitudinal monitoring of DNA viral loads in transplant patients using quantitative metagenomic next-generation sequencing

Ellen C. Carbo [*#], Anne Russcher [*], Margriet E.M. Kraakman, Caroline S. de Brouwer,
Igor A. Sidorov, Mariet C.W. Feltkamp, Aloys C.M. Kroes, Eric C.J. Claas, Jutte J.C. de Vries

Clinical Microbiological Laboratory, department of Medical Microbiology,
Leiden University Medical Center, Leiden, the Netherlands;
** Both authors contributed equally to this work*
# Corresponding author

Published in Pathogens

Abstract

Introduction: Immunocompromised patients are prone to reactivations and (re-)infections of multiple DNA viruses. Viral load monitoring by single-target quantitative PCRs (qPCR) is the current cornerstone for virus quantification. In this study, a metagenomic next-generation sequencing (mNGS) approach was used for the identification and load monitoring of transplantation-related DNA viruses. Methods: Longitudinal plasma samples from six patients that were qPCR-positive for cytomegalovirus (CMV), Epstein-Barr virus (EBV), BK polyomavirus (BKV), adenovirus (ADV), parvovirus B19 (B19V), and torque teno-virus (TTV) were sequenced using the quantitative metagenomic Galileo Viral Panel Solution (Arc Bio, LLC, Cambridge, MA, USA) reagents and bioinformatics pipeline combination. Qualitative and quantitative performance was analysed with a focus on viral load ranges relevant for clinical decision making. Results: All pathogens identified by qPCR were also identified by mNGS. BKV, CMV, and HHV6B were additionally detected by mNGS, and could be confirmed by qPCR or auxiliary bioinformatic analysis. Viral loads determined by mNGS correlated with the qPCR results, with inter-method differences in viral load per virus ranging from 0.19 log10 IU/mL for EBV to 0.90 log10 copies/mL for ADV. TTV, analysed by mNGS in a semi-quantitative way, demonstrated a mean difference of 3.0 log10 copies/mL. Trends over time in viral load determined by mNGS and qPCR were comparable, and clinical thresholds for initiation of treatment were equally identified by mNGS. Conclusions: The Galileo Viral Panel for quantitative mNGS performed comparably to qPCR concerning detection and viral load determination, within clinically relevant ranges of patient management algorithms.

Keywords: viral metagenomics; pathogen detection; quantification; next-generation sequencing; load monitoring

1. Introduction

Opportunistic viral infections frequently occur after solid organ or hematopoietic cell transplantation, with associated morbidity and mortality of up to 40% [1]. Successful prevention and early detection of viral infections including reactivations are the cornerstones of transplant patient management. For effective pre-emptive and therapeutic treatment strategies, accurate viral load quantification is essential. Typically, in immunocompromised hosts, multiple viruses can reactivate simultaneously, which makes comprehensive identification of replicating pathogenic viruses essential. Currently, the monitoring of opportunistic viral infections in transplant patients is most frequently performed by multiple single-plex quantitative PCRs.

Metagenomic next-generation sequencing (mNGS) is increasingly being applied for the identification of pathogens in undiagnosed cases suspected of infection [2,3,4]. Quantification of viral loads utilising mNGS remains a challenge [5,6,7,8]. Complicating factors are the varying amount of background sequences from the host and from bacterial origin, technical bias affecting target sequence depth, unselective attribution of reads, and the number of calibration curves that are needed simultaneously when using untargeted sequencing for viral load calculations. Reports comparing mNGS with qPCR demonstrated a correlation with normalised sequence read counts but never as accurate as qPCR for viral load prediction [5]. Other previous research concerning the quantification of shotgun sequence read counts focused mainly on differential expression of RNA [9,10,11,12].

Recently, the Galileo Viral Panel (Arc Bio, LLC, Cambridge, MA, USA) has been designed as a quantitative mNGS approach for ten transplant-related DNA viruses [13,14]. This all-inclusive approach encompasses the library preparation kit, controls, calibration reagents, and cloud-based user-friendly software for bioinformatic analysis. Previous data on the performance of this mNGS approach demonstrated that the analytical performance was comparable to qPCR results with regard to the limits of detection, limits of quantification, and inter-assay variation [13,14].

In this study, we analysed the performance of the Galileo Viral Panel for viral load quantification in transplant patients over time. Subsequent samples from six transplant patients with proven infections or reactivations with transplantation-related DNA viruses (adenovirus, ADV; BK polyomavirus, BKV; cytomegalovirus, CMV; Epstein-Barr virus, EBV; human herpesvirus type 6A, HHV-6A; human herpesvirus type 6B, HHV-6B; herpes simplex type 1, HSV-1; herpes simplex type 2, HSV-2; JC polyomavirus, JCV; varicella-zoster virus, VZV; parvovirus B19, B19V; and torque teno virus, TTV) were analysed in comparison with qPCR. Accuracy of viral load quantification by mNGS was studied in relation to thresholds that had been used for the initiation of treatment or tapering of immunosuppression. Furthermore, we investigated the additional detection of DNA viruses identified by the broad mNGS approach, for which no targeted qPCR had initially been ordered.

2. Methods

2.1. Patients and Sample Selection

Six adult immunocompromised patients (one allogeneic stem cell transplant patient, four kidney transplant patients, and one patient with hematological malignancy) were retrospectively selected based on available follow-up EDTA plasma samples that previously tested positive for one or more transplantation-related DNA viruses. Samples had previously (July 2008–December 2019) been sent to the Clinical Microbiological Laboratory (CML) of the Leiden University Medical Center (LUMC, The Netherlands) for viral load monitoring as part of routine patient care. Routine patient diagnostics consisted of several collection points, resulting in positive qPCR’s with a wide range of viral loads. CMV/EBV were routinely screened for in plasma post transplantation. BKV was screened in urine post renal transplantation; when positive it was also screened for in plasma. ADV and B19V were not routinely screened for but ordered at the discretion of the treating physician based on symptomatology. TTV viral load had been tested retrospectively by qPCR in the context of a different study. Patient plasma samples were stored at −80 °C until mNGS analysis.

2.2. Ethical Approval

Approval was obtained from the ethical committee from the LUMC (P11.165 NL 37682.058.11, and Biobank Infectious Diseases protocol 2020-03 & 2020-04 B20.002).

2.3. Extraction of Nucleic Acids; Internal Controls

Patient plasma samples were spiked with an internal control (baculovirus, Arc Bio, LLC) before extraction. Nucleic acids were extracted from 200 μL plasma using the MagNApure 96 DNA and Viral NA Small volume extraction kit on the MagNAPure

96 system (Roche Diagnostics, Almere, The Netherlands) with 100 µL output eluate. The eluate was concentrated using vacuum centrifugation by a SpeedVac vacuum concentrator (Thermo Scientific, Waltham, MA, USA) to a volume of 26 µL.

2.4. Library Preparation and Sequencing

Sequence libraries were prepared using the Galileo Viral Panel sequencing kit (Arc Bio, LLC, Cambridge, MA, USA) according to the manufacturer's instructions. The protocol was based on enzymatic fragmentation at 37 °C for 5 min, followed by end repair and A-tailing at 65 °C for 30 min. Subsequently, fragments were ligated using unique dual-index adapters (ArcBio) at 20 °C for 15 min and purified using magnetic Kapa Pure Beads (Roche, Basel, Switzerland). No RNase treatment was included in the procedure, and human DNA was depleted using human depletion reagents at 45 °C for 2 h followed by 45 °C for 15 min, after which libraries were amplified using library amplification primers for 45 °C for 30 s, by 14 cycles of 98 °C for 10 s and 65 °C for 75 s and 65 °C for 5 min. The final library preparation products were purified using magnetic Kapa Pure Beads (Roche) and quantified using a Qubit fluorometer (Thermo Fisher, Waltham, MA, USA) followed by equally pooling using the Arc Bio calculation pooling tool. After a final quantity and quality check using a Bioanalyser (Agilent, Santa Clara, CA, USA), samples were sequenced using the NovaSeq 6000 sequencing system (Illumina, San Diego, CA, USA) at GenomeScan B.V. (Leiden, The Netherlands). For sequencing, S4 flowcells were used and samples were sequenced in two runs, where each pool consisted of around 12% of the lane capacity. Ten million reads per library were aimed for; the total reads per sample can be found in Table S1.

2.5. Calibration Samples

Initial calibration runs were performed testing the multi-analyte mixture (MAM) of whole-virus particles at viral loads of 0, 1000, 5000, 10,000, and 100,000 copies/mL or IU/mL plasma, in quintuple (Arc Bio, LLC) for the following 10 viruses: hADV-C1, BKV, CMV, EBV, HHV-6A, HHV6B, HSV-1, HSV-2, JCV, and VZV. For TTV and B19V, no Arc Bio calibrator panels were available, and therefore the Galileo Signal values were plotted against the calibrator plot of other viruses that demonstrated optimal agreement with the viral load (JCV and VZV, respectively), representing a semi-quantitative result.

2.6. Bioinformatic Analysis

After demultiplexing of the sequence reads using bcl2fastq (version 2.2.0) (Illumina, San Diego, CA, USA), FASTQ files were uploaded to the Galileo Analytics web application [13,15] which automatically processes data for quality assessment and pathogen detection using a custom database of DNA viruses involved in transplant-associated infections: ADV, CMV, EBV, HHV-6A, HHV-6B, HSV-1, HSV-2, JCV, VZV, B19V, and TTV. Human reads were removed before uploading the fastq files to the web application after mapping them to the human reference genome GRCh38 with Bowtie2 version 2.3.4 [6]. The analytics web application aligns sequence reads to the genomes of the DNA viruses in their calibration kit, scores these read alignments based on complexity, uniqueness, and alignment scores, and reports this in a signal value. The signal value is normalised for read counts across libraries, correcting for differences in genome lengths and technical bias, based on the spiked-in normalisation controls. The signals reported are related to the genomic depth and the observed amount of viral DNA being present in a sample, belonging to non-confounding genomic regions [13]. The sample signals were visualised in linear calibration curves (Figure S1).

2.7. Analysis of Performance and Additional Findings

Performance of the metagenomic Galileo Viral Panel assay was assessed in comparison with routine qPCR, analysing both qualitative and quantitative detection. Additional findings by mNGS were confirmed by additional qPCR analysis. In case no remaining sample was available, the Galileo Analytics software results were compared with results from the analysis using alternative bioinformatic tools: metagenomic taxonomic classifier Centrifuge (1.0.4-beta) [16] and de novo assembly-based viral metagenomic analysis software Genome Detective [17].

3. Results

3.1. Calibration Curves

After metagenomic sequencing, the viral loads were calculated for each virus by the Galileo Analytics web application. Signals of both the calibrators and patient plasma samples were plotted in load graphs (Figure S1) and the corresponding viral load of the patient samples was extrapolated. As no calibrator panels for B19V and TTV virus were available, these signals were plotted against other calibration curves of viruses that demonstrated the optimal agreement with the known viral load for semi-quantitative detection. All calibration sample signals correlated well with the titre (R^2 range 0.84–0.92).

3.2. Viral Load by mNGS Versus qPCR

In total, six patients were tested by qPCR and mNGS for quantification of different viruses at subsequent time points. The agreement between the methods for qualitative detection was 100% for the viruses targeted by PCR. Quantitative results per patient are shown in Table 1, and Figure 1 depicts viral loads by mNGS versus qPCR per target virus. CMV and EBV viral loads demonstrated the highest agreement, with a maximum difference in viral load of 0.70 $\log_{10}$ IU/mL. Mean differences in viral loads were 0.43 for CMV and 0.19 $\log_{10}$ IU/mL for EBV. Genotyping had not been performed for ADV (patient 1) and TTV (patient 4) in the context of routine care but resulted in the human adenovirus 1 and TTV-like mini virus, respectively, using mNGS data (based on de novo genome assembly followed by blastn). Viral loads were higher when quantified with mNGS with a mean difference of 0.90 $\log_{10}$ c/mL. For BKV, viral loads by mNGS were lower in comparison with qPCR, with a mean difference of 1.32 $\log_{10}$ c/mL. When taking into account viral loads measured above the limit of quantification of 2.5 $\log_{10}$ c/mL, as applied in our diagnostic qPCR for BKV, the mean difference is 0.62 $\log_{10}$ c/mL and a trend towards a better agreement with higher viral loads could be observed. Semi-quantitative detection of B19V and TTV viruses by mNGS resulted in mean differences of, respectively, 0.39 $\log_{10}$ IU/mL and 3.0 $\log_{10}$ c/mL in comparison with qPCR.

Table 1. Viral load quantification by qPCR and mNGS per patient sample.

Patient-sample	Viral load qPCR	Viral load qPCR (log10)	Viral load mNGS	Viral load mNGS (log10)	ΔqPCR-mNGS (log10)
Virus: ADV					
P1-S1	675 c/mL	2,83 c/mL	1277 c/mL	3,11 c/mL	0,28 c/mL
P1-S2	4517	3,65	66273	4,82	1,17
P1-S3	34740	4,54	287844	5,46	0,92
P1-S4	136900	5,14	1435130	6,16	1,02
P1-S5	60540	4,78	777172	5,89	1,11
Virus: BKV					
P2-S1	796 c/mL	2,90 c/mL	3 c/mL	0,48 c/mL	-2,42 c/mL
P2-S2	614	2,79	3	0,48	-2,31
P2-S3	233700	5,37	9011	3,95	-1,41
P2-S4	2401000	6,38	1857785	6,27	-0,11
P2-S5	71480	4,85	32321	4,51	-0,34
Virus: CMV					
P3-S1	2370 IU/mL	3,37 IU/mL	6246 IU/mL	3,80 IU/mL	0,42 IU/mL
P3-S2	122800	5,09	275657	5,44	0,35
P3-S3	10680	4,03	22242	4,35	0,32
P3-S4	4915	3,69	11366	4,06	0,36
P3-S5	9156	3,96	46231	4,66	0,70
Virus: EBV					
P3-S1	2083 IU/mL	3,32 IU/mL	4581 IU/mL	3,66 IU/mL	0,34 IU/mL
P3-S2	12970	4,11	1573	4,20	0,09
P3-S3	17710	4,25	14549	4,16	-0,09
P3-S4	10500	4,02	15077	4,18	0,16
P3-S5	7723	3,89	14844	4,17	0,28
Virus: TTV*					
P4-S1	140 c/mL	2,15 c/mL	4 c/mL	0,60 c/mL	-1,54 c/mL
P4-S2	2400000	6,38	5142	3,71	-2,67
P4-S3	5,7E+09	9,76	319074	5,50	-4,25
P4-S4	2,4E+08	8,38	46261	4,67	-3,71
Virus: B19V *					
P5-S1	1,34 *1011 IU/mL	11,13 IU/mL	2,07 *1011 IU/mL	11,32 IU/mL	0,19 IU/mL
P5-S2	1407365	6,15	1235416	6,09	-0,06
P5-S3	45846	4,66	41787	4,62	-0,04
Virus: B19V *					
P6-S1	4,07 *1010 IU/mL	10,61 IU/mL	4,37 *1011 IU/mL	11,64 IU/mL	1,03 IU/mL
P6-S2	5309308	6,73	9376953	6,97	0,25
P6-S3	8569	3,93	49601	4,70	0,76

* B19V and TTV results were considered semi-quantitative since no Arc Bio calibration samples were available for these targets.

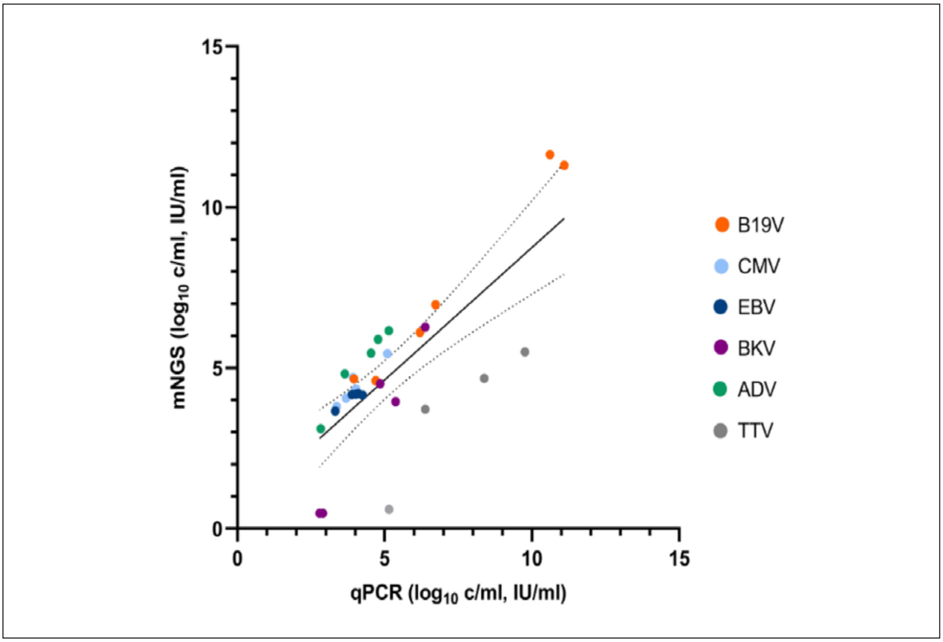


Figure 1. Viral loads as predicted by Galileo Viral Panel mNGS versus qPCR.
(copies/mL for ADV, BK, and TTV, and IU/mL for CMV, EBV, and B19V). B19V and TTV results were considered semi-quantitative, as no Galileo calibration panels were available for these targets.

3.3. Longitudinal Patient Follow-Up and Clinical Decision Making

Table 2 gives an outline of patient characteristics and provides clinical information on underlying conditions and complications during the sampling period. Furthermore, for each patient, the viral loads over time were plotted in graphs with clinical information, symptomatology, relevant laboratory parameters, and treatment (Figure 2). For CMV, EBV, and BKV, in our clinical practice, specific viral load thresholds are used to decide whether immunosuppression should be tapered and/or antiviral therapy should be administered. Viral load quantification around these thresholds demonstrated good agreement in identifying these clinical decision-making breakpoints. In Patient 3, the antiviral treatment with Foscarnet was started for CMV-reactivation when viral load measured by qPCR exceeded 4.0 log10IU/mL. By mNGS, this critical threshold for treatment initiation was correctly identified with a viral load by mNGS of 5.44 log10 IU/mL. In the same patient, rituximab was administered when the EBV load by qPCR was repeatedly above the threshold of 4.0 log10 IU/mL, consistently quantified thrice above 4.0 log10 IU/mL before administration of rituximab, both by qPCR and mNGS.

For B19V, ADV, and TTV, no predefined thresholds were used for changing the treatment regimen. For all viruses, the observed trends in load over time in each patient were comparable for qPCR and mNGS, despite the semi-quantitative nature of the B19V mNGS assay. Effect of treatment (anti-viral drugs, immunoglobulins, and/or tapering of immunosuppressive drugs) in patients was estimated by follow-up of viral loads by qPCR. For B19V in Patients 5 and 6, the effect of intravenous immunoglobulins (IVIG) could be assessed by the decreasing viral load in the weeks after administration, as also observed by mNGS. For ADV, in patient 1, antiviral therapy with cidofovir was started when a consistent increase in viral load was detected, both by qPCR and mNGS.

3.4. Additional Findings

For some samples, additional viral reads were detected in the pathogenic mNGS reports that were not initially tested for by qPCR (Table S1). Most additional findings were supported by a secondary bioinformatic analysis using the Centrifuge and Genome Detective: BK (1 patient), CMV (1 patient), HHV-6B (1 patient), and TTV (4 patients, torque teno virus was the deepest level of classification obtained, using mNGS data, with lower than 100% genome coverage). In a few cases, additional findings were not confirmed by a second analysis, leaving some low mNGS signals for CMV, EBV, and HSV. JCV was detected by mNGS in a sample with a high concentration of BKV, which possibly indicated forced alignment contamination due to high sequence homology between JCV and BKV [13,14].

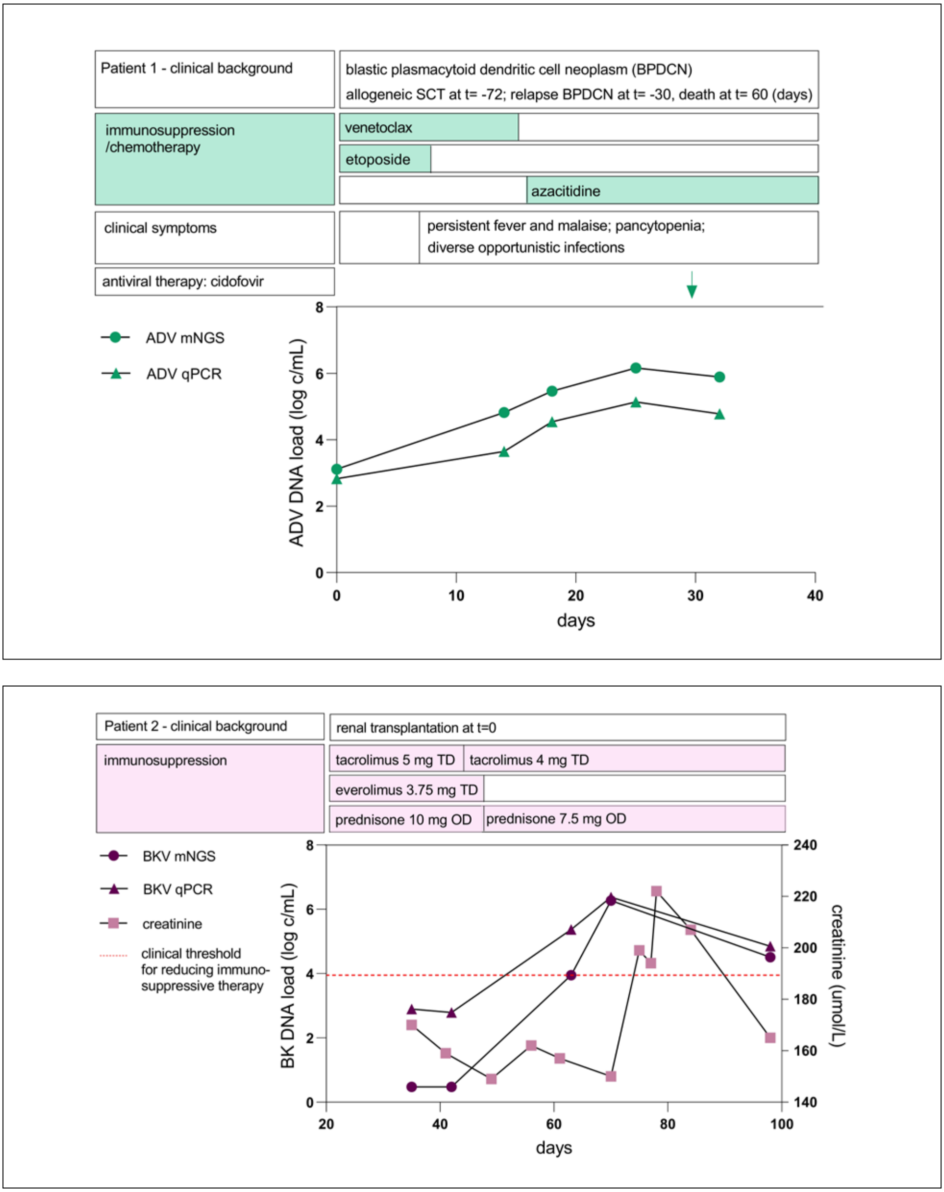


Figure 2. Longitudinal follow-up of DNA viral loads in immunosuppressed patients over time, as predicted by mNGS (Galileo Viral Panel, Arc Bio) versus qPCR.

Clinical information and therapeutic agents are included.

Figure 2. continued

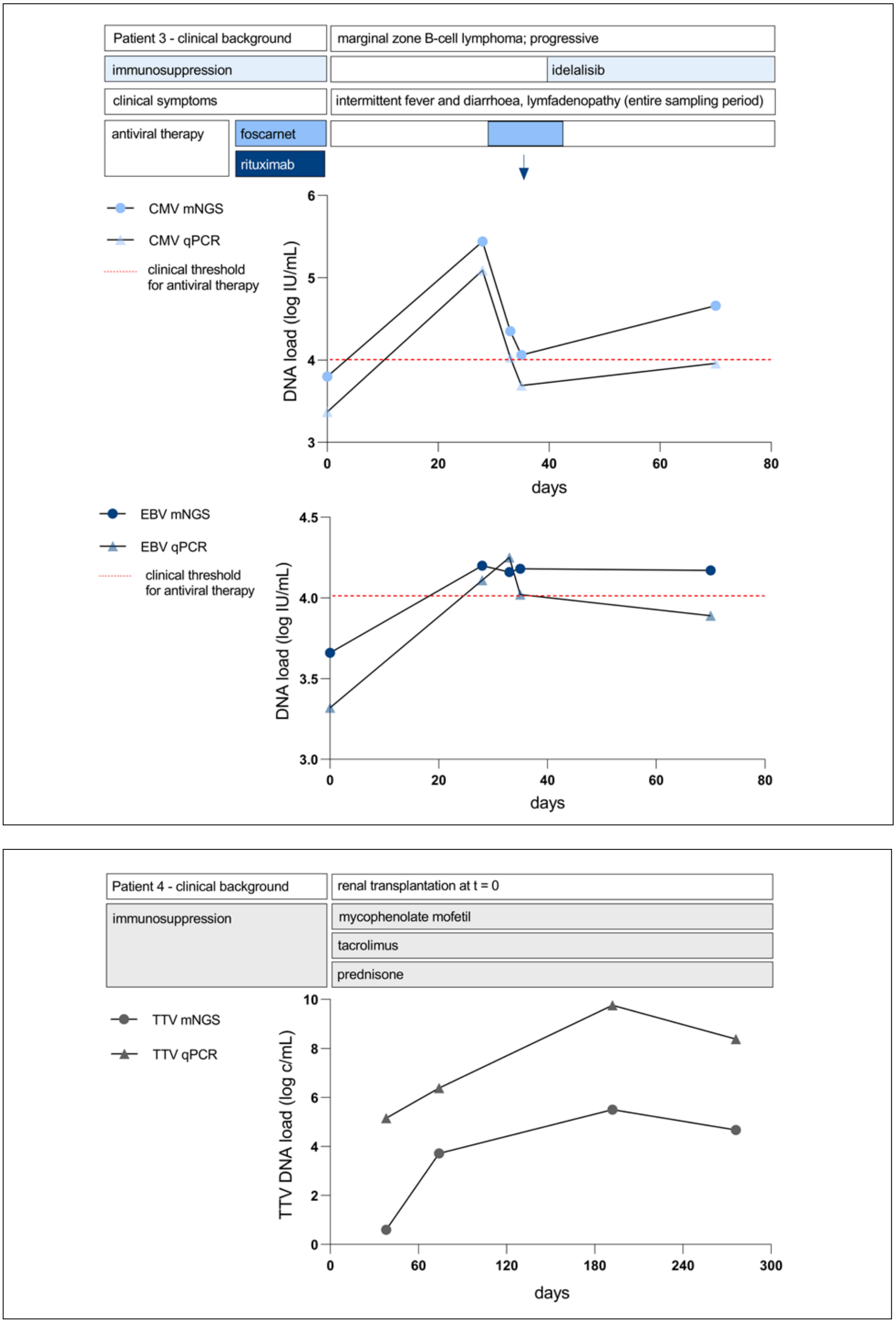


Figure 2. continued

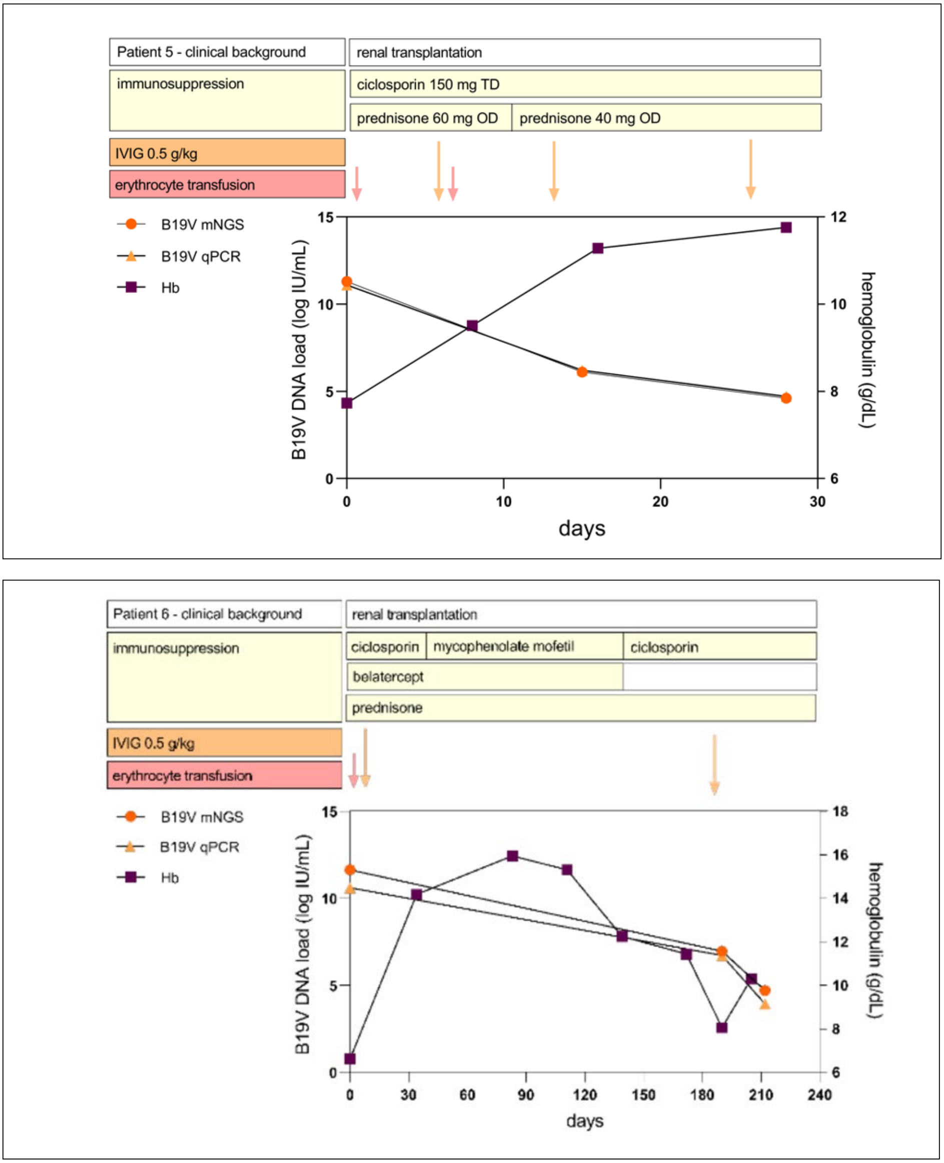


Table 2. **Patient characteristics and clinical background at start of longitudinal follow-up.**

Patient Number	Virus	Age Range	Sex	Underlying Condition	Conditioning Regimen	Transplantation	Other Known Infectious Complications During Sampling Period
1	ADV	60–79	V	Blastic plasmacytoid dendritic cell neoplasm (BPDCN)	Two failed remission-induction regimens; followed by t * = −3: COPADM † t= −2: COPADM	t = 0: Non-myeloablative allogeneic stem cell transplant from unrelated donor; t = 1: relapse BPDCN	1. Probable pulmonal aspergillosis 2. CMV reactivation treated with foscarnet (week before sampling period) 3. Enterococcus faecalis UTI ‡
2	BKV	20–39	M	Chronic renal insufficiency due to TIN ¥, as an extraintestinal manifestation of known colitis ulcerosa or medicine-induced	Alemtuzumab	Pre-emptive living-related renal transplant	1. CMV reactivation
3	CMV, EBV	60–79	V	Marginal zone B-cell lymphoma; established 4 years previously, now progressive	Recent chemotherapy: t = −6: CHOP ¦	Not applicable	1. Escherichia coli UTI ‡ 2. rhinovirus RTI ¶
4	TTV	40–59	V	IgA nephropathy	Basiliximab	Living-related renal transplant	1. Escherichia coli UTI ‡
5	B19V	40–59	M	IgA nephropathy	Basiliximab	pre-emptive living-unrelated renal transplant	
6	B19V	40–59	M	Focal segmental glomerulosclerosis (FSGS)	Not applicable	Non-heart beating renal transplant 4 years previously; 15 years previously living-related renal transplant	

* t = time in months; † COPADM = cyclophosphamide, oncovin (vincristine), prednisone, Adriamycin (doxorubicin), methotrexate;
‡ UTI = urinary tract infection; ¥ TIN = tubulointerstitial nefritis;
¦ CHOP = cyclophosphamide, oncovin (vincristine), Adriamycin, prednisone;
¶ RTI = respiratory tract infection.
For a complementary longitudinal overview of symptomatology, including laboratory parameters and treatment, see Figure 2.

4. Discussion

In this study, the performance of a quantitative mNGS assay for the longitudinal follow-up of DNA viral loads was analysed in six immunocompromised patients. Viral loads determined by mNGS were comparable with loads determined by qPCR, and differed less than 1 $\log_{10}$ for DNA viruses with calibration panels available, in line with previous studies [13,14]. In the current study, the performance of viral loads assessed by mNGS was also evaluated with regard to clinical decision making. In the management of reactivating viruses in immunocompromised patients, local and international guidelines use viral load breakpoints to decide whether antiviral therapy should be administered or whether immunosuppression should be tapered [18,19,20,21,22]. Viral loads under investigation in this study were determined by qPCR as part of routine patient care. When local clinical breakpoints were considered for each virus, mNGS performed comparably to qPCR to identify the clinically relevant breakpoints. B19V is not considered to be a reactivating virus, but quantification may be helpful to distinguish clinically relevant replicative infection from merely DNA remnants [23]. In the range of these breakpoints, viral loads were adequately determined by mNGS to guide clinical decision making. Additionally, the longitudinal trend was similar in comparison with qPCR, indicating precision of mNGS for clinical quantification and reliable indication of the trend in viral load. Clinical decision making is often guided by follow-up of viral load trends, in addition to the cross-sectional viral load measurements for viral infections without available thresholds. In the future, more research is desired to analyse the performance in the lower ranges to map the limit of quantification (LOQ) of mNGS procedures. It is anticipated that the LOQ is somewhat higher than the LOQ of qPCR, given the generally higher limit of detection in combination with the variability of mNGS, mainly resulting from the varying amounts of background sequences.

The principle of a quantitative catchall approach to detect all transplantation-related viruses in a single run is an attractive feature in the clinical follow-up of the immunocompromised host. Simultaneous reactivation of persistent viruses during immunocompromised episodes is common. Co-infection rates of up to 32% have been described using PCR and, importantly, were associated with higher rates of acute rejection or graft dysfunction [24]. Co-infections may be missed when ordering targeted PCRs, while the catchall approach of mNGS could guarantee that active infections are not overlooked. Indeed, our approach demonstrated a complementary yield of seven reactivating viruses in five patients, which had not been identified earlier by qPCR. Some of these unnoticed viruses are not considered pathogenic, such as TTV. However, the role of TTV in clinical management is still developing, as recent and ongoing research suggests its potential as marker of functional immunity, with an inverse correlation between TTV-load and risk of rejection. Clinical trials exploring its role as a marker for balancing immunosuppressive treatment, with a focus on tacrolimus, are currently being conducted (e.g., ClinicalTrials.gov NCT04198506) [25,26,27,28]. ADV, generally, is not systematically screened for in the severely immunosuppressed adult population. In our patient, although actively diagnosed, ADV-loads were rapidly increasing and a catchall approach could guarantee that such less common infections are not overlooked, especially in the absence of localizing symptoms.

A significant complementary virus identification yield by mNGS in transplant patients of 31/49 plasma samples was also reported by Sam et al. [14], with the majority, being viruses, considered pathogenic. These findings demonstrate that mNGS could improve pathogen detection in clinical practice.

Another advantage of mNGS would be its capacity to genotype viruses and detect mutations associated with antiviral resistance, without the need for additional, time-consuming, target-specific ‘wet’ lab procedures that could delay diagnosis and treatment. As an example, Patient 3 in our study was treated with Foscarnet for persistent CMV reactivation pending the results of mutational analysis after clinical failure of valganciclovir treatment. If the results of mutational analysis had been immediately available, resorting to second-line treatment may have been avoided.

Widespread implementation of mNGS approaches in clinical diagnostic settings has been limited by several factors. The ‘wet’ lab protocols can be time-consuming, costly, and have a relatively long turnaround time, mainly due to the time required for sequencing. With various sequencing techniques still rapidly evolving, the costs and sequencing turnaround time of such protocols are expected to improve considerably in the future [29]. Furthermore, bioinformatic skills are generally needed for validation and implementation as a diagnostic assay. User-friendly, all-in-one mNGS data analysis software packages for cloud-based and automated analysis enable use in laboratories with minimal bioinformatic knowledge and allow access to high-performance computing capacity.

Limitations in this current study are the relatively low number of samples and viruses when considering a metagenomic approach, including two viruses without calibration panels available. This small-scale study provides a proof-of-principle demonstration in a retrospective design demonstrating that the current version of the Research Use Only Galileo Viral Panel enables longitudinal viral load monitoring by mNGS. It is expected that, after these initial studies, indicating high performance in terms of limit of detection and quantification, inter-run precision, and prospective viral load monitoring, the kit and software will be expanded to include more viruses, calibration samples, and potentially fit for different sample types. Furthermore, technical and bioinformatic features might be evolved in future versions of the assay.

Overall, viral metagenomic sequencing is a promising approach not only for DNA virus detection and identification, but also for reliable estimation of the viral load in a clinical setting, and potentially mutational typing for drug sensitivity analysis. Several milestones essential for implementation in diagnostic settings have been met by the specific assay used in this study: the limits of detection, the limits of quantification, precision, and overall technical performance, which were comparable with qPCR assays. Precise quantification was accomplished by read normalisation based on a designed control. These accomplishments pave the way for further developments and optimisation of quantitative metagenomic sequencing for longitudinal viral load monitoring and beyond.

Supplementary Materials

The following are available online at:
https://www.mdpi.com/article/10.3390/pathogens11020236/s1,
Figure S1: Calibration graphs of the six viruses in six patients in this study with associated slope, intercepts and R2 values;
Table S1: Additional findings of the metagenomic Galileo Viral Panel compared to Centrifuge and Genome Detective software.

Funding

This research received no external funding.

Institutional Review Board Statement

Approval for this study involving patient material was obtained from the ethical committee from the LUMC (P11.165 NL 37682.058.11, and Biobank Infectious Diseases protocol 2020-03 & 2020-04 B20.002).

Acknowledgments

We thank Gavin Wall, Naissan Hussainzada, and Meredith Carpenter (Arc Bio) for their technical and logistic support. We thank Aline van Rijn for retrieving clinical data from one of the patients.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] I. G. Sia and R. Patel, 'New Strategies for Prevention and Therapy of Cytomegalovirus Infection and Disease in Solid-Organ Transplant Recipients', Clin. Microbiol. Rev., vol. 13, no. 1, pp. 83–121, Jan. 2000, doi: 10.1128/CMR.13.1.83.

[2] E. C. Carbo et al., 'Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics', J. Clin. Virol., p. 104566, Jul. 2020, doi: 10.1016/j.jcv.2020.104566.

[3] A. Reyes et al., 'Viral metagenomic sequencing in a cohort of international travellers returning with febrile illness', Infectious Diseases (except HIV/AIDS), preprint, May 2021. doi: 10.1101/2021.05.13.21257019.

[4] A. L. van Rijn et al., 'The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease', PLOS ONE, vol. 14, no. 10, p. e0223952, Oct. 2019, doi: 10.1371/journal.pone.0223952.

[5] S. van Boheemen et al., 'Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients', J. Mol. Diagn., vol. 22, no. 2, pp. 196–207, Feb. 2020, doi: 10.1016/j.jmoldx.2019.10.007.

[6] J. J. C. de Vries et al., 'Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples', J. Clin. Virol., p. 104908, Jul. 2021, doi: 10.1016/j.jcv.2021.104908.

[7] F. X. López-Labrador et al., 'Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure', J. Clin. Virol., vol. 134, p. 104691, Jan. 2021, doi: 10.1016/j.jcv.2020.104691.

[8] J. J. C. de Vries et al., 'Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting', J. Clin. Virol., vol. 138, p. 104812, May 2021, doi: 10.1016/j.jcv.2021.104812.

[9] G. Cai et al., 'Accuracy of RNA-Seq and its dependence on sequencing depth', BMC Bioinformatics, vol. 13, no. Suppl 13, p. S5, 2012, doi: 10.1186/1471-2105-13-S13-S5.

[10] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, 'Differential analysis of gene regulation at transcript resolution with RNA-seq', Nat. Biotechnol., vol. 31, no. 1, pp. 46–53, Jan. 2013, doi: 10.1038/nbt.2450.

[11] K. R. Kukurba and S. B. Montgomery, 'RNA Sequencing and Analysis', Cold Spring Harb. Protoc., vol. 2015, no. 11, p. pdb.top084970, Nov. 2015, doi: 10.1101/pdb.top084970.

[12] C. Y. Chiu et al., 'Diagnosis of Fatal Human Case of St. Louis Encephalitis Virus Infection by Metagenomic Sequencing, California, 2016', Emerg. Infect. Dis., vol. 23, no. 10, pp. 1964–1968, Oct. 2017, doi: 10.3201/eid2310.161986.

[13] M. L. Carpenter et al., 'Metagenomic Next-Generation Sequencing for Identification and Quantitation of Transplant-Related DNA Viruses', J. Clin. Microbiol., vol. 57, no. 12, pp. e01113-19, /jcm/57/12/JCM.01113-19.atom, Sep. 2019, doi: 10.1128/JCM.01113-19.

[14] S. S. Sam, R. Rogers, F. S. Gillani, G. J. Tsongalis, C. S. Kraft, and A. M. Caliendo, ‘Evaluation of a Next-Generation Sequencing Metagenomics Assay to Detect and Quantify DNA Viruses in Plasma from Transplant Recipients’, J. Mol. Diagn., vol. 23, no. 6, pp. 719–731, Jun. 2021, doi: 10.1016/j.jmoldx.2021.02.008.

[15] ‘ArcBio Galileo Transplant WebApp’. [Online]. Available: galileo.arcbio.com

[16] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg, ‘Centrifuge: rapid and sensitive classification of metagenomic sequences’, Genome Res., vol. 26, no. 12, pp. 1721–1729, Dec. 2016, doi: 10.1101/gr.210641.116.

[17] M. Vilsker et al., ‘Genome Detective: an automated system for virus identification from high-throughput sequencing data’, Bioinformatics, vol. 35, no. 5, pp. 871–873, Mar. 2019, doi: 10.1093/bioinformatics/bty695.

[18] R. R. Razonable and A. Humar, ‘Cytomegalovirus in solid organ transplant recipients—Guidelines of the American Society of Transplantation Infectious Diseases Community of Practice’, Clin. Transplant., vol. 33, no. 9, Sep. 2019, doi: 10.1111/ctr.13512.

[19] H. H. Hirsch, P. S. Randhawa, and AST Infectious Diseases Community of Practice, ‘BK polyomavirus in solid organ transplantation—Guidelines from the American Society of Transplantation Infectious Diseases Community of Practice’, Clin. Transplant., vol. 33, no. 9, Sep. 2019, doi: 10.1111/ctr.13528.

[20] T. Lion, ‘Adenovirus persistence, reactivation, and clinical management’, FEBS Lett., vol. 593, no. 24, pp. 3571–3582, Dec. 2019, doi: 10.1002/1873-3468.13576.

[21] Manaresi and Gallinella, ‘Advances in the Development of Antiviral Strategies against Parvovirus B19’, Viruses, vol. 11, no. 7, p. 659, Jul. 2019, doi: 10.3390/v11070659.

[22] U. D. Allen, J. K. Preiksaitis, and the AST Infectious Diseases Community of Practice, ‘Post-transplant lymphoproliferative disorders, Epstein-Barr virus infection, and disease in solid organ transplantation: Guidelines from the American Society of Transplantation Infectious Diseases Community of Practice’, Clin. Transplant., vol. 33, no. 9, Sep. 2019, doi: 10.1111/ctr.13652.

[23] M. W. A. Molenaar-de Backer, A. Russcher, A. C. M. Kroes, M. H. G. M. Koppelman, M. Lanfermeijer, and H. L. Zaaijer, ‘Detection of parvovirus B19 DNA in blood: Viruses or DNA remnants?’, J. Clin. Virol., vol. 84, pp. 19–23, Nov. 2016, doi: 10.1016/j.jcv.2016.09.004.

[24] C. Anderson-Smits, E. R. Baker, and I. Hirji, ‘Coinfection rates and clinical outcome data for cytomegalovirus and Epstein-Barr virus in post-transplant patients: A systematic review of the literature’, Transpl. Infect. Dis., vol. 22, no. 6, Dec. 2020, doi: 10.1111/tid.13396.

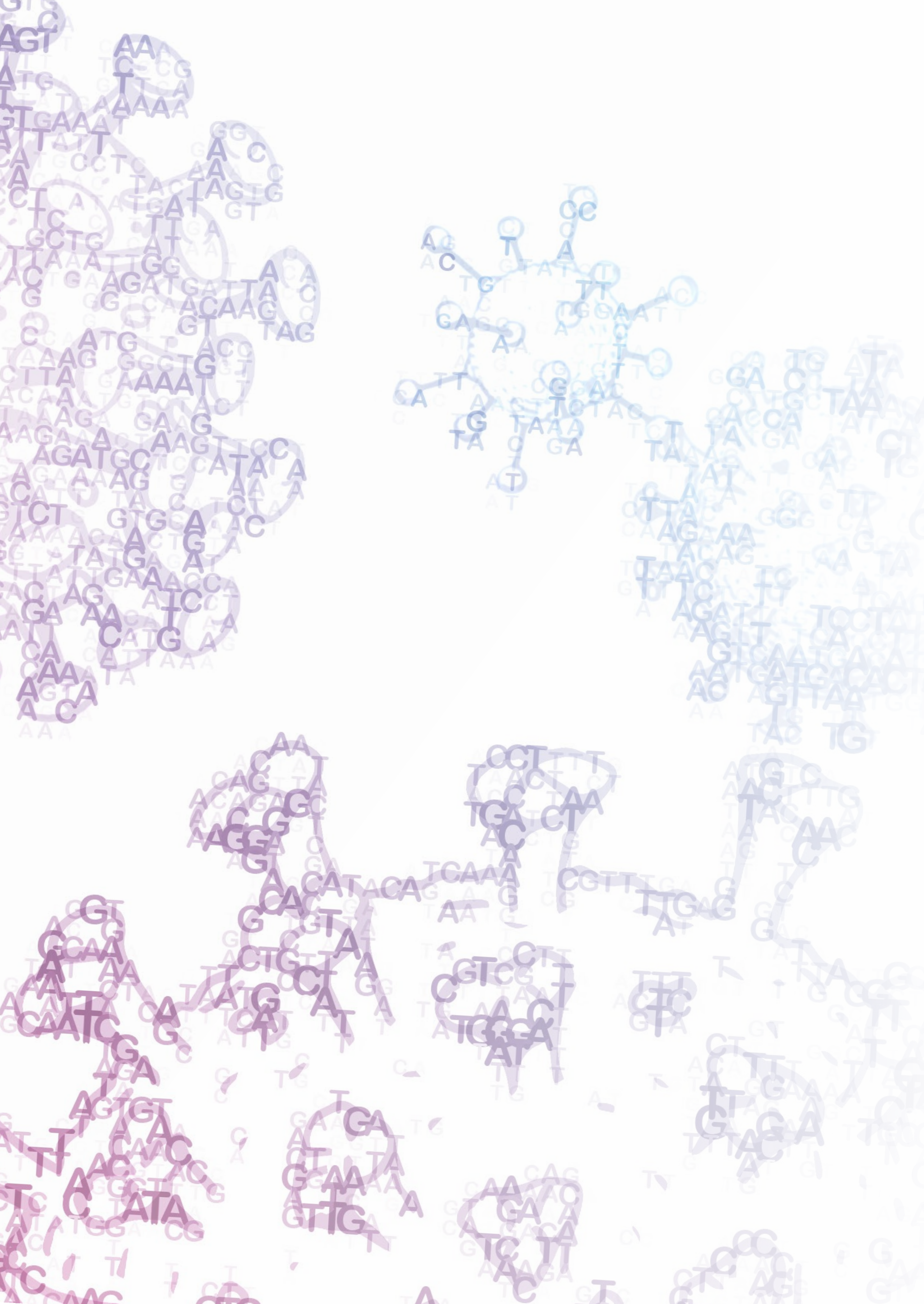
[25] S. Spandole, D. Cimponeriu, L. M. Berca, and G. Mihăescu, ‘Human anelloviruses: an update of molecular, epidemiological and clinical aspects’, Arch. Virol., vol. 160, no. 4, pp. 893–908, Apr. 2015, doi: 10.1007/s00705-015-2363-9.

[26] A. Moustafa et al., ‘The blood DNA virome in 8,000 humans’, PLOS Pathog., vol. 13, no. 3, p. e1006292, Mar. 2017, doi: 10.1371/journal.ppat.1006292.

[27] A. L. van Rijn et al., ‘Torque teno virus loads after kidney transplantation predict allograft rejection but not viral infection’, J. Clin. Virol., vol. 140, p. 104871, Jul. 2021, doi: 10.1016/j.jcv.2021.104871.

[28] Rezahosseini O, Drabe CH, Sørensen SS, Rasmussen A, Perch M, Ostrowski SR, and Nielsen SD. Torque teno virus load as a potential endogenous marker of immune function in solid organ transplantation. Transplantation Reviews. 2019;33(3):137-144., https://doi.org/10.1016/j.trre.2019.03.004

[29] A. von Bubnoff, 'Next-Generation Sequencing: The Race Is On', Cell, vol. 132, no. 5, pp. 721–723, Mar. 2008, doi: 10.1016/j.cell.2008.02.028.



Chapter 7 **Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness**

Ellen C. Carbo[1*], Igor A. Sidorov[1], Jessica C. Zevenhoven-Dobbe[1], Eric J. Snijder[1], Eric C. Claas[1],
Jeroen F.J. Laros[2,3,4], Louis C.M. Kroes[1], Jutte J.C. de Vries[1]

[1] Department of Medical Microbiology, Leiden University Medical Center (LUMC), Leiden, the Netherlands
2 Department Human Genetics, Leiden University Medical Center (LUMC), Leiden, the Netherlands
3 Department of Clinical Genetics, Leiden University Medical Center (LUMC), Leiden, the Netherlands
4 National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands

SCAN ME

Published in Journal of Clinical Virology

Abstract

Introduction: The SARS-CoV-2 pandemic of 2020 is a prime example of the omnipresent threat of emerging viruses that can infect humans. A protocol for the identification of novel coronaviruses by viral metagenomic sequencing in diagnostic laboratories may contribute to pandemic preparedness.

Aim: The aim of this study is to validate a metagenomic virus discovery protocol as a tool for coronavirus pandemic preparedness.

Methods: The performance of a viral metagenomic protocol in a clinical setting for the identification of novel coronaviruses was tested using clinical samples containing SARS-CoV-2, SARS-CoV, and MERS-CoV, in combination with databases generated to contain only viruses of before the discovery dates of these coronaviruses, to mimic virus discovery.

Results: Classification of NGS reads using Centrifuge and Genome Detective resulted in assignment of the reads to the closest relatives of the emerging coronaviruses. Low nucleotide and amino acid identity (81% and 84%, respectively, for SARS-CoV-2) in combination with up to 98% genome coverage were indicative for a related, novel coronavirus. Capture probes targeting vertebrate viruses, designed in 2015, enhanced both sequencing depth and coverage of the SARS-CoV-2 genome, the latter increasing from 71% to 98%.

Conclusion: The model used for simulation of virus discovery enabled validation of the metagenomic sequencing protocol. The metagenomic protocol with virus probes designed before the pandemic, can assist the detection and identification of novel coronaviruses directly in clinical samples.

Keywords

SARS-CoV-2; virus discovery; metagenomics; bioinformatics

1. Introduction

The Severe Acute Respiratory Syndrome Coronavirus type 2 (SARS-CoV-2) pandemic of 2020 demonstrates the devastating effect an emerging virus can have. Although previous pandemics such as the Spanish Flu (1918) and Asian Flu (1957) resulted in a multitude of fatal cases, the SARS-CoV-2 pandemic exhibits an unprecedented impact on public health, the economy and society as a whole. In 2002 and 2012 respectively, the Severe Acute Respiratory Syndrome (SARS [1] and Middle Eastern Respiratory Syndrome (MERS) Coronavirus [2] have emerged as zoonotic infections causing severe respiratory disease, with continued introductions of MERS-CoV remaining a public health threat up to now [3].

Pandemic preparedness comprises strategies and measures to protect human health and lives in anticipation of the worldwide spread of (re)emerging pathogens. Pandemic preparedness plans [4] focus on measures to contain and control the spread of emerging pathogens. Early detection of the pathogen is the mainstay of initiating infection control measures. Global surveillance as a component of the International Health Regulations (IHR) aims at early detection and monitoring of human cases of zoonotic diseases with pandemic potential [5]. Pandemic surveillance plans commonly focus on specific viruses, such as influenza, and depend on targeted detection of these specific viral threats, limiting the detection of unanticipated and novel viruses. The current SARS-CoV-2 pandemic shows the need for unbiased identification of potential pathogens.

Metagenomic Next-Generation Sequencing (mNGS) enables hypothesis-free sequencing of all nucleic acids in a given sample, including genomes of pathogens. All sequences are amplified, followed by classification of sequences based on a reference database. While research applications are more common, mNGS is being introduced in clinical diagnostic laboratories as indicated by recently diagnosed cases of encephalitis [6]. Implementation of mNGS in clinical diagnostics requires validation of metagenomic protocols. Metagenomic protocols and pipelines have been successfully used for detection of known pathogens [6,7,8]. However, detection and identification of novel, previously unknown emerging viruses presents a challenge due to the absence of their genome sequences in reference databases.

In this study, we validated the identification of emerging coronaviruses by a viral metagenomic protocol, using clinical samples with SARS-CoV-2, and samples spiked with cultivated isolates SARS-CoV Frankfurt-1 (SARS-CoV) and MERS-CoV EMC/2012 (MERS-CoV). The validation included analysis of the performance of both an in-house and a commercially available data analysis pipeline, Genome Detective [9]. Identification of coronaviruses was tested using modified databases lacking SARS-CoV-2, SARS-CoV, and MERS-CoV, mimicking the situation at the time of virus discovery. Additionally, the efficacy of detection of novel coronaviruses using capture probes targeting vertebrate viruses [10,11] known before the current pandemic was analyzed using a SARS-CoV-2 clinical sample.

2. Methods

2.1. Sample selection and preparation

Nasopharyngeal swabs were obtained from two patients who tested positive for SARS-CoV-2 by real-time PCR targeting the SARS-CoV-2 E-gene [12] with Cq values of 20 and 30, respectively. These PCRs were performed as part of routine diagnostics at the Clinical Microbiological Laboratory (CML) of the Leiden University Medical Center.

For the SARS-CoV and MERS-CoV analyses, nasopharyngeal material that had tested negative for all respiratory viruses addressed by in-house multiplex PCRs(coronaviruses 229E, HKU1, NL63, OC43, influenza A, B, human metapneumovirus, parainfluenza 1-4, respiratory syncytial virus and rhinovirus) was spiked in with the cultivated isolates SARS-CoV Frankfurt-1 [1,13] and MERS-CoV EMC/2012 [14] with viral load per sample being 1.3 × 105 PFU and 2.4 × 105 PFU and Cq values of 23 and 22, respectively.

2.2. Metagenomic Next-Generation Sequencing (mNGS)

Library preparation and sequencing were performed using a previously validated protocol [15,16]. Briefly, 200 µl of patient samples were spiked with equine arteritis virus (EAV) and phocid herpesvirus-1 (PhHV-1) prior to NA extraction using the Magnapure 96 DNA and Viral NA Small volume extraction kit on the MagnaPure 96 system (Roche, Basel, Switzerland) resulting in 100 µL nucleic acid-containing eluate. Of this eluate, 50 µl per sample was used as input for the library prep, utilizing the NEBNext Ultra II Directional RNA Library prep kit for Illumina (New England Biolabs, Ipswich, MA, USA), dual indexed NEBNext Multiplex Oligos for Illumina (1.5µM), and a protocol optimized for processing RNA and DNA simultaneously in a single tube [15].

Library preps of the samples where processed both with and without enrichment for viruses using sequence capture probes (see below). Subsequent sequence analysis was performed using a NovaSeq6000 sequencing system (Illumina, San Diego, CA, USA) at GenomeScan BV to obtain approximately 10 million 150bp reads per sample.

2.3. Viral capture probe enrichment

Enrichment of viral sequences from the sample library pools was performed using the SeqCap EZ HyperCap kit according to the manufacturer's instructions (Roche, Basel, Switzerland). This kit uses a vertebrate virus SeqCap EZ probe pool designed to target a set of sequences from vertebrate viruses that were available in 2015 [10], including the following: *Coronaviridae* (NCBI:txid11118), *Coronavirinae* (NCBI:txid 693995), *Alphacoronavirus* (NCBI:txid693996), *Betacoronavirus* (NCBI:txid694002), *Gammacoronavirus* (NCBI:txid694013), and *Deltacoronavirus* (NCBI:txid1159901). Amplified DNA libraries from two SARS-CoV-2 samples and one negative control, with a combined mass of 1 µg, were pooled in equal amounts in a single enrichment experiment. Some adaptions were made: human Cot DNA and blocking oligos (Integrated DNA Technologies, Coralville, IA, USA) were added to each enrichment pool to prevent nonspecific binding and binding of human DNA to the probes. Subsequently, hybridization to the probe pool was performed for 40 hours. Next, the Hyber Cap Bead kit was used for washing the captured DNA, followed by post capture PCR amplification using the KAPA HiFi HotStart ReadyMix (2×) (Roche, Basel, Switzerland) and Illumina NGS primers (5 µM). The final washing step was performed using AMPure XP beads (Beckman Coulter, Inc., Brea, CA, USA) after which quality and quantity of the enriched libraries were assessed by Qubit analysis (Thermo Fisher, Waltham, MA, USA) and Bioanalyzer (Agilent, Santa Clara, CA, USA).

2.4. Sequence read classification: Centrifuge

After quality pre-processing using an in-house QC pipeline, Biopet version 0.9.0 [17] and removal of human reads after mapping them to human reference genome GRCh38 [18] with Bowtie2 version 2.3.4 [19], the remaining sequencing reads were taxonomically classified using Centrifuge 1.0.2-beta [20] with the databases prepared by taking all 12,302 Refseq viral genomes (as of Juny 16th, 2020) and extracting the GenBank records annotated before the dates of the existence of the MERS-CoV and SARS-CoV index patients in 2012 and 2002, respectively. Reads with multiple matches were assigned to the lowest common ancestor (k = 1). Taxonomic assignments of reads by Centrifuge were visualized with Krona version 2.0 [21].

2.5. In-house virus discovery protocol

Pre-processed short reads were *de novo* assembled into contigs using SPAdes version 3.10.1 [22]. All contigs were analyzed using the NCBI Basic Local Alignment Search Tool (BLAST 2.8.1) [23] using the BLAST NCBI's nucleotide (nt) database (accessed April 2018). Only viral hits for contigs with a length of ≥500bp were selected to identify the best shared homology to viruses. A length of 500bp was taken to ensure coverage of the built contigs by at least 3 reads, to rule out any possible contamination. Only hits dated prior to the date of emergence of the viruses were considered to mimic the virus discovery setting for SARS-CoV, MERS-CoV and SARS-CoV-2.

2.6. Genome Detective: commercial classification and discovery tool

After extraction of human reads, FASTQ files generated for SARS-CoV-2 samples (with and without viral enrichment) were uploaded for classification and *de novo* assembly by the commercial web-based tool Genome Detective v1.120 (www.genomedetective.com, accessed 2020-05-11) [9], using a reference database (generated 2019-09-21). In brief, after removal of low-quality reads and trimming by Trimmomatic [24], candidate viral reads were identified using the protein-based alignment method DIAMOND [25] in combination with the Swissprot UniRef90 protein database followed by *de novo* assembly using metaSPAdes [26]. Blastx and Blastn [23] were used to search for candidate reference sequences using the NCBI RefSeq virus database (accessed 2019-09-21). Consensus sequences were produced by joining *de novo* contigs using Advanced Genome Aligner [27].

3. Results

3.1. Classification of SARS-CoV-2, SARS-CoV, and MERS-CoV using databases created before the emergence of these viruses

To mimic the classification conditions present in the setting of virus discovery, viral metagenomic reference genome databases created before the emergence of SARS-CoV-2, SARS-CoV and MERS-CoV were used for the classification of sequence reads (December 2019 for the two SARS-CoV-2 positive samples, November 2002 for the SARS-CoV and June 2012 for the MERS-CoV positive samples). Classification results of viral reads are shown in Fig. 1 and Table 1. Sequence reads obtained for SARS-CoV-2 samples were classified as belonging to SARS coronavirus and Bat coronavirus BM48-31/BGR/2008. Sequence reads of the SARS-CoV sample were classified as belonging to Porcine epidemic diarrhoea virus and bovine coronavirus, and reads of the MERS-CoV sample as Bat coronavirus BM48-31/BGR/2008, belonging to the *Betacoronavirus* genus (Table 1).

3.2. Virus discovery: *de novo* assembly

Results of *de novo* assembly of all samples for contigs longer than 500bp are shown in Table 2. BLASTn was used to search for hits with sequence homology. Only viral hits with the lowest E-value of all matches identified that were submitted before the publication of SARS-CoV-2 genomes were considered. BLASTn search results of the contigs with *Coronaviridae* hits are listed in Table 2 including the length of the longest contig for each sample. Identity data of the hits with the lowest E-value are listed in Supplementary Table 1. Additional BLAST alignment figures of the longest contigs of both the SARS-CoV and MERS-CoV samples can be found in Supplementary Fig. 1 and 2, respectively.

3.3. Virus discovery of SARS-CoV-2 by GenomeDetective

GenomeDetective results of identification of SARS-CoV-2 sequences using a database created before the emergence of SARS-CoV-2 are shown in Fig. 2. SARS-CoV-2 sequences were identified as SARS-CoV, with nucleotide and amino acid identity of 80-81% and 83-85% respectively in combination with up to 98% genome coverage, being indicative for a novel finding.

3.4. Virus discovery using capture probes

The efficacy of a metagenomic sequencing protocol using capture probes targeting vertebrate virus sequences designed before the emergence of SARS-CoV-2, was studied in the context of virus discovery. We analyzed metagenomic data from the two SARS-CoV-2 positive samples prepared both with and without viral enrichment. The total amount of contigs and the number of contigs matching genomes of viruses from *Coronaviridae* are shown in Table 2 and Fig. 2. For the clinical sample with higher SARS-CoV-2 load (Cq 20), genome coverage was comparable (98% vs. 97% genome coverage), and for the sample with lower load (Cq 30), genome coverage was markedly higher (74% vs. 91% genome coverage) when the metagenomic protocol with viral capture probes was used.

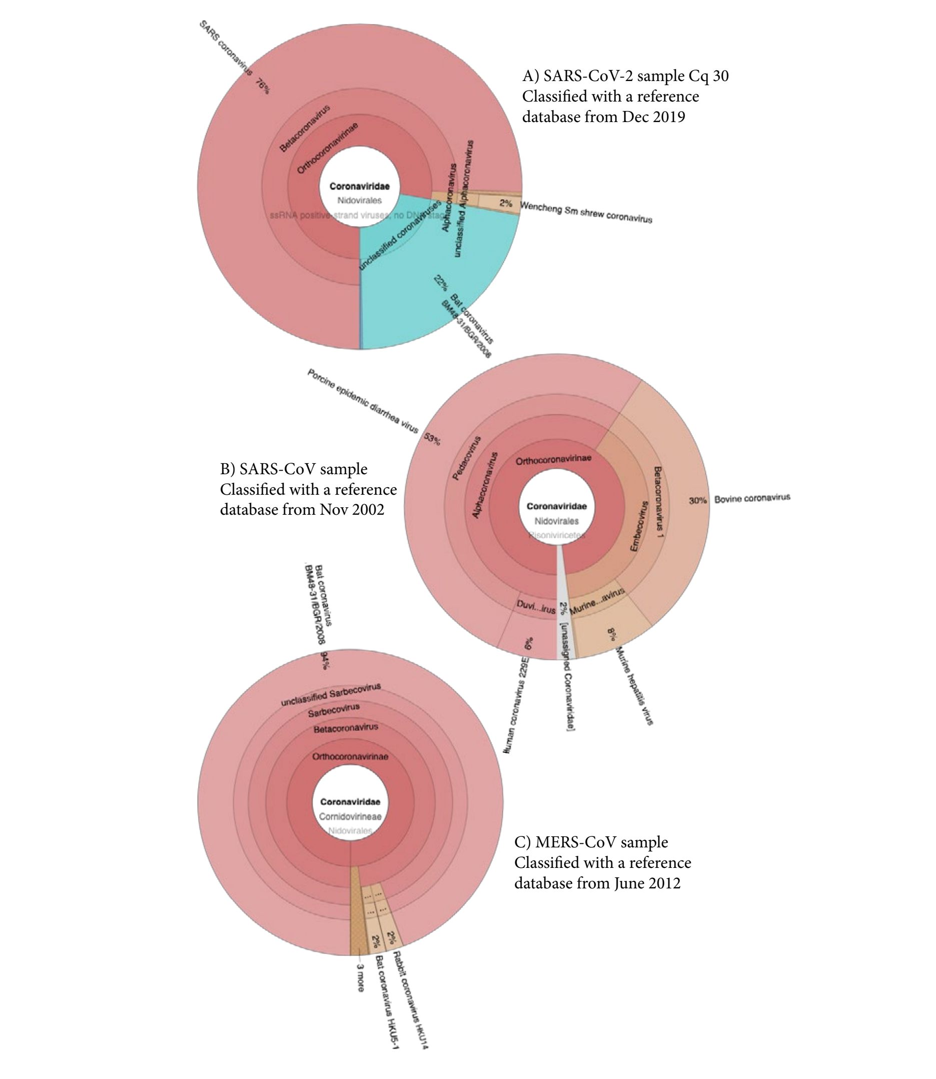


Figure 1. **Centrifuge classification results of viral reads of SARS-CoV-2, SARS-CoV, and MERS-positive samples, using viral metagenomic databases created before the emergence of these viruses. A) SARS-CoV-2, B) SARS-CoV, C) MERS.**

Table 1. Classification of SARS-CoV-2, SARS-CoV, and MERS sequence reads using reference databases created before their emergence, using metagenomic classifier Centrifuge.

Sample	Untargeted mNGS, or viral enrichment by capture probes	Total number of non-human reads	Number of reads classified as *Coronaviridae* (% of total non-human)	*Coronaviridae* assignment of >10% classified *Coronaviridae* reads
SARS-CoV-2 Patient A (Cq 20)	Untargeted	3,488,842	2,166 (0.06)	SARS-CoV Bat coronavirus BM48-31/ BGR/2008
	Viral capture [a]	9,582,942	3,518,798 (36.72)	SARS-CoV Bat coronavirus BM48-31/ BGR/2008
SARS-CoV-2 Patient B (Cq 30)	Untargeted	919,930	604 (0.07)	SARS-CoV Bat coronavirus BM48-31/ BGR/2008
	Viral capture [a]	9,894,246	572,061 (5.78)	SARS-CoV Bat coronavirus BM48-31/ BGR/2008
SARS-CoV Frankfurt-1 (Cq 23)	Untargeted	6,936,399	436 (0.006)	Bovine coronavirus Porcine epidemic diarrhea virus
MERS-CoV EMC/2012 (Cq 22)	Untargeted	8,201,535	8,748 (0.1)	Bat coronavirus BM48-31/ BGR/2008

[a] Enrichment by capture probes targeting vertebrate viruses designed in 2015

Table 2. **Classification of SARS-CoV-2, SARS-CoV, and MERS *de novo* assembled contigs using BLAST.**

Sample	Untargeted mNGS, or viral enrichment by capture probes	Total contigs ≥ 500bp	Viral contigs ≥ 500bp	Corona-viridae-contig ≥ 500bp	Length of the longest Corona-viridaecontig, bp	BLAST alignment length, bp	BLAST identity match, %	Subject taxonomy name	Release year of sequence of the species	Release year of sequence of the subject found
SARS-CoV-2 Patient A (Cq 20)	Untargeted	8,606	15	3	19,654	12,069	87.141	Bat SARS SL CoVZC45	2003	2018
	Viral capture [a]	8,232	51	31	5,811	5,820	90.567	Bat SARS SL CoVZC45	2003	2018
SARS-CoV-2 Patient B (Cq 30)	Untargeted	2,815	31	16	2,503	2,456	91.450	Bat SARS SL CoVZXC21	2003	2018
	Viral capture [a]	2,110	39	13	4,866	4,856	92.360	Bat SARS SL CoVZC45	2003	2018
SARS-CoV Frankfurt-1 (Cq 23)	Untargeted	3,836	10	1	29,692	1,236	72.411	Bovine coronavirus isolate 4-17-03	2001	2018
MERS-CoV EMC/2012 (Cq 22)	Untargeted	4,074	9	1	30,097	14,856	77.248	Bat coronavirus HKU4-1	2006	2006

Table showing the total number of built contigs with a length > = 500bp, the number of these contigs where the hit with the lowest E-value would be a hit to viruses, the number of contigs where the hit with the lowest E-value would be a hit to *Coronaviridae* and of this last group the length of the longest contig, the alignment length, identity match, taxonomic name of BLAST result and the release years of sequences belonging to the species and subjects found by BLAST.

[a] Enrichment by capture probes targeting vertebrate viruses designed in 2015

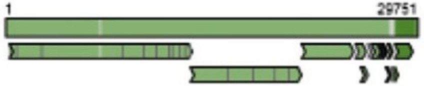
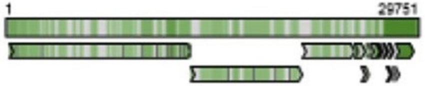
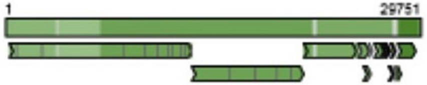
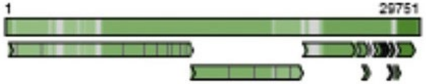
	Number of Configs	Number of Reads	Coverage, %	Depth of Covergae	idetify, %		Genome Coverage Assignment to Severe acute respiratory syndrome-related coronavirus
					NT	AA	
A) Untargeted Patient A (Cq 20)	3	10,426	94.4	43.7	79.6	83.2	
Patient B (Cq 30)	36	3,126	74.2	17	80.7	84.5	
B) Captured Patient A (Cq 20)	5	10,601,614	97.1	46,956,.	80.2	83.9	
Patient B (Cq 30)	12	1,942,472	91.3	9,041.4	80.9	84.9	

Figure 2. Discovery performance using metagenomic sequencing (A) without and (B) with enrichment by capture probes targeting vertebrate viruses, designed in 2015. Genome Detective classification of reads, coverage and aligment against the genome of Severe acute respiratory syndrome-related coronavirus are shown.

Reads mapping to the SARS-CoV-2 reference genome were used to visualize the difference in using capture probes as depicted in Fig. 3, where the SARS-CoV-2 genome is almost completely covered. The two largest contigs built by SPAdes that had a hit with the lowest E-value when BLASTed against genomes from *Coronaviridae*, were 4,866bp and 5,811bp in length for the two SARS-CoV-2 samples enriched using probes.

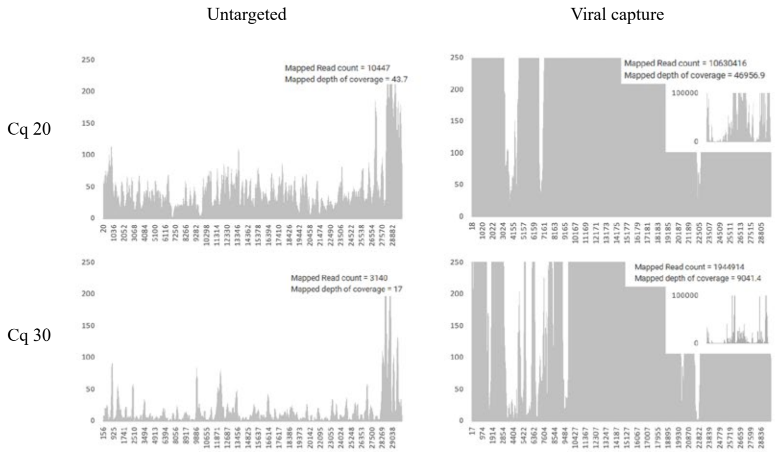


Figure 3. Coverage map of alignment against SARS-CoV-2 reference sequence NC_004718.2, without (left) and with (right) viral capture probes designed in 2015 after metagenomic sequencing of patient samples with respectively Cq 20 (upper graphs) and Cq 30 (lower graphs).

4. Discussion

In this study, we evaluated the performance of a metagenomic sequencing protocol for the identification of emerging viruses using clinical samples in combination with a simulated reference database. High and low loads of SARS-CoV-2, SARS-CoV, and MERS-CoV in clinical samples could be detected as ‘novel’ viruses, using only reference sequences created before these viruses emerged. Sequence reads were assigned to the closest relatives of these viruses available at that time and assembled with heterologous sequences to ‘novel’ consensus genomes. Low identity of these consensus genomes with genomes of closely related ones indicated a novel virus. Additionally, probes targeting sequences of vertebrate viruses, available prior to the coronavirus pandemic of 2020, succeeded in the capture of nearly the full genome of SARS-CoV-2. It must be noted that the validation was performed using emerging viruses with nucleotide identity of over 76% to their closest known relatives and conclusions cannot be extended to novel viruses which are less closely related. Nucleotide (and amino acid) identities reported in literature with regard to novel human pathogenic viruses vary, for example 50% for older viruses like SARS-CoV [1], 80% for MERS-CoV [14], 88% for parts of the Human Metapneumovirus [28] and up to 97.2% for parts of SARS-CoV-2 [29].

Several reports have shown an increase of 100-10,000 fold in sensitivity for detection of known viruses when using capture probes [10,30] and here we report the potential of using capture probes in the detection of novel viruses. Sequence variation was addressed in the probe design by retaining mutant or variant sequences if sequences diverged by more than 90% [10]. Lipkin and colleagues describe the capture of conserved regions of a rodent hepacivirus isolate with 75% identity using VirSeqCap VERT, and even 40% for detection rather than whole genome sequencing is suggested [10]. The capture probes used in this study targeted sequences of several isolates of alpha-, beta-, gamma-, and deltacoronaviruses. In this study the whole genome of SARS-CoV-2, with 76-100% overall nucleotide identity to the probe targets, was detected using these probes.

Metagenomic sequencing is increasingly being used in diagnostic laboratories as a hypothesis-free approach for suspected infectious diseases in undiagnosed cases. Metagenomic sequencing in diagnostic laboratories has resulted in the detection of pathogens present in the reference database but either not tested for by routine methods due to rare or unknown associations with a specific disease, or for which routine testing failed (e.g., due to primer mismatches). Additionally, mNGS enables the detection of novel pathogens not (yet) present in the databases. Common bioinformatic classifiers are usually not designed for discovery purposes, so additional algorithms including a separate validation to assess the performance in a discovery setting are needed. Reports on specific bioinformatic discovery tools typically describe the algorithm and an *in silico* analysis and here we present validation studies on the performance of virus discovery tools using clinical samples.

Implementation of virus discovery protocols in diagnostic laboratories may contribute to increased vigilance for emerging viruses and therefore aids in surveillance and pandemic preparedness.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

We would like to thank Joost van Harinxma thoe Slooten and Alhena Reyes for the library preparations and viral probe enrichments. Additionally, we would like to thank Lopje Höcker, Margriet Kraakman and Tom Vreeswijk for all their technical assistance in the lab, and Leon Mei and Michel Villerius for their bioinformatic support.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jcv.2020.104594.

References

[1] C. Drosten et al., ‘Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome’, N. Engl. J. Med., vol. 348, no. 20, pp. 1967–1976, May 2003, doi: 10.1056/NEJMoa030747.

[2] T. G. Ksiazek et al., ‘A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome’, N. Engl. J. Med., vol. 348, no. 20, pp. 1953–1966, May 2003, doi: 10.1056/NEJMoa030781.

[3] http://www.who.int/csr/don/02-jul-2020-mers-saudi-arabia/en/ (Accessed Jul. 04, 2020).

[4] http://www.euro.who.int/en/health-topics/communicable-diseases/influenza/pandemic-influenza/pandemic-preparedness/national-preparedness-plans (Accessed Jun. 30, 2020).

[5]. http://www.who.int/csr/disease/swineflu/global_pandemic_influenza_surveilance_apr09.pdf. Accessed: Jun. 30, 2020. [Online].

[6] S. Miller et al., ‘Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid’, Genome Res., vol. 29, no. 5, pp. 831–842, 2019, doi: 10.1101/gr.238170.118.

[7] Y. Li et al., ‘VIP: an integrated pipeline for metagenomics of virus identification and discovery’, Sci. Rep., vol. 6, no. 1, p. 23774, Apr. 2016, doi: 10.1038/srep23774.

[8] R. R. Miller, V. Montoya, J. L. Gardy, D. M. Patrick, and P. Tang, ‘Metagenomics for pathogen detection in public health’, Genome Med., vol. 5, no. 9, p. 81, 2013, doi: 10.1186/gm485.

[9] M. Vilsker et al., ‘Genome Detective: an automated system for virus identification from high-throughput sequencing data’, Bioinformatics, vol. 35, no. 5, pp. 871–873, Mar. 2019, doi: 10.1093/bioinformatics/bty695.

[10] T. Briese et al., ‘Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis’, mBio, vol. 6, no. 5, pp. e01491-15, Sep. 2015, doi: 10.1128/mBio.01491-15.

[11] T. N. Wylie, K. M. Wylie, B. N. Herter, and G. A. Storch, ‘Enhanced virome sequencing using targeted sequence capture’, Genome Res., vol. 25, no. 12, pp. 1910–1920, Dec. 2015, doi: 10.1101/gr.191049.115.

[12] V. M. Corman et al., ‘Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR’, Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull., vol. 25, no. 3, 2020, doi: 10.2807/1560-7917.ES.2020.25.3.2000045.

[13] V. Thiel et al., ‘Mechanisms and enzymes involved in SARS coronavirus genome expression’, J. Gen. Virol., vol. 84, no. 9, pp. 2305–2315, Sep. 2003, doi: 10.1099/vir.0.19424-0.

[14] A. M. Zaki, S. van Boheemen, T. M. Bestebroer, A. D. M. E. Osterhaus, and R. A. M. Fouchier, ‘Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia’, N. Engl. J. Med., vol. 367, no. 19, pp. 1814–1820, Nov. 2012, doi: 10.1056/NEJMoa1211721.

[15] S. van Boheemen et al., ‘Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients’, J. Mol. Diagn., vol. 22, no. 2, pp. 196–207, Feb. 2020, doi: 10.1016/j.jmoldx.2019.10.007.

[16] A. L. van Rijn et al., ‘The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease’, PLOS ONE, vol. 14, no. 10, p. e0223952, Oct. 2019, doi: 10.1371/journal.pone.0223952.

[17] http://biopet-docs.readthedocs.io/en/stable/ (Accessed Jul. 03, 2020).

[18] https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/ (Accessed Jul. 10, 2020).

[19] B. Langmead and S. L. Salzberg, ‘Fast gapped-read alignment with Bowtie 2’, Nat. Methods, vol. 9, no. 4, pp. 357–359, Apr. 2012, doi: 10.1038/nmeth.1923.

[20] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg, ‘Centrifuge: rapid and sensitive classification of metagenomic sequences’, Genome Res., vol. 26, no. 12, pp. 1721–1729, Dec. 2016, doi: 10.1101/gr.210641.116.

[21] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, ‘Interactive metagenomic visualization in a Web browser’, BMC Bioinformatics, vol. 12, no. 1, p. 385, Dec. 2011, doi: 10.1186/1471-2105-12-385.

[22] A. Bankevich et al., ‘SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing’, J. Comput. Biol., vol. 19, no. 5, pp. 455–477, May 2012, doi: 10.1089/cmb.2012.0021.

[23] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, ‘Basic local alignment search tool’, J. Mol. Biol., vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[24] A. M. Bolger, M. Lohse, and B. Usadel, ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, Bioinformatics, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[25] B. Buchfink, C. Xie, and D. H. Huson, ‘Fast and sensitive protein alignment using DIAMOND’, Nat. Methods, vol. 12, no. 1, pp. 59–60, Jan. 2015, doi: 10.1038/nmeth.3176.

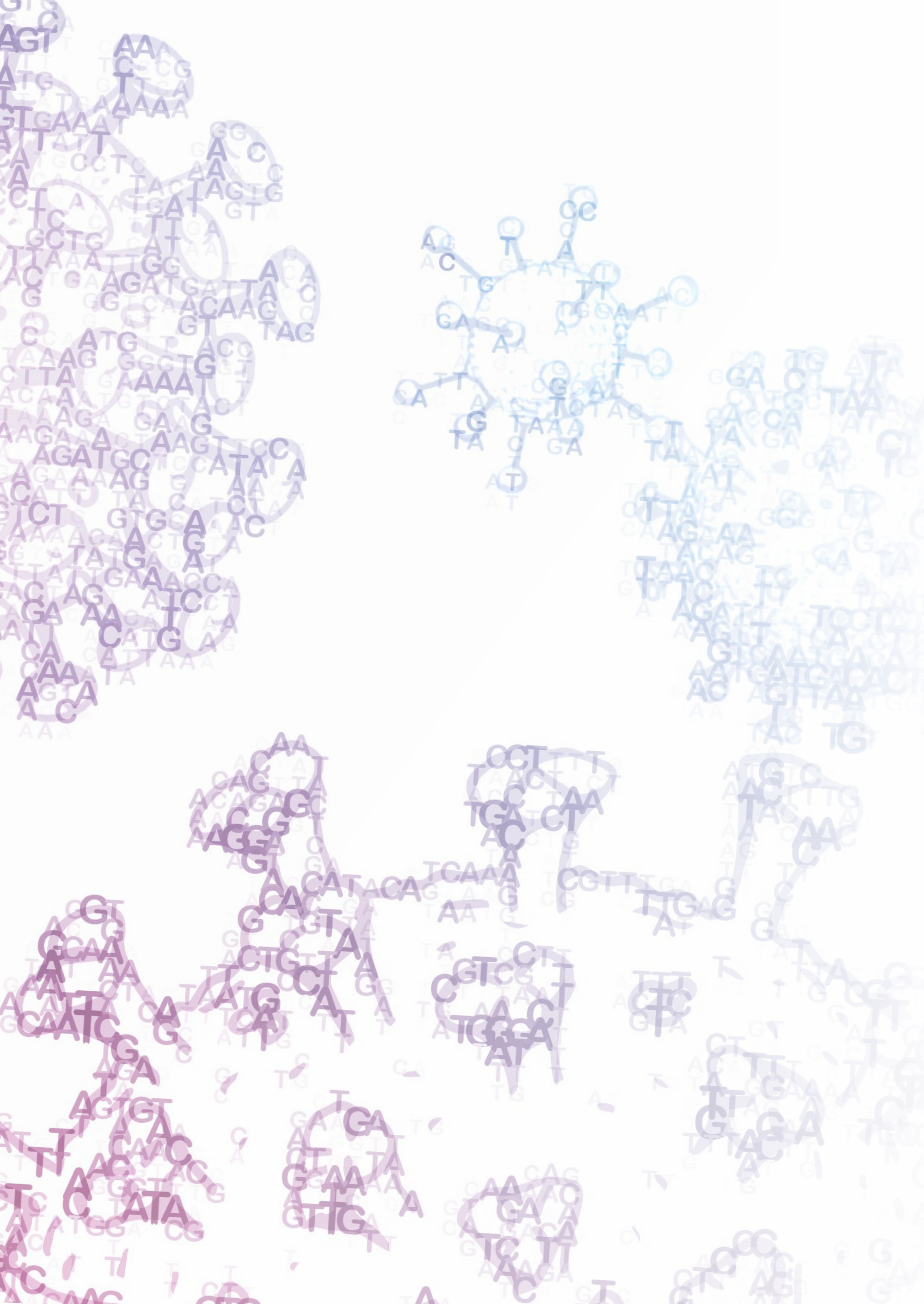
[26] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, ‘metaSPAdes: a new versatile metagenomic assembler’, Genome Res., vol. 27, no. 5, pp. 824–834, May 2017, doi: 10.1101/gr.213959.116.

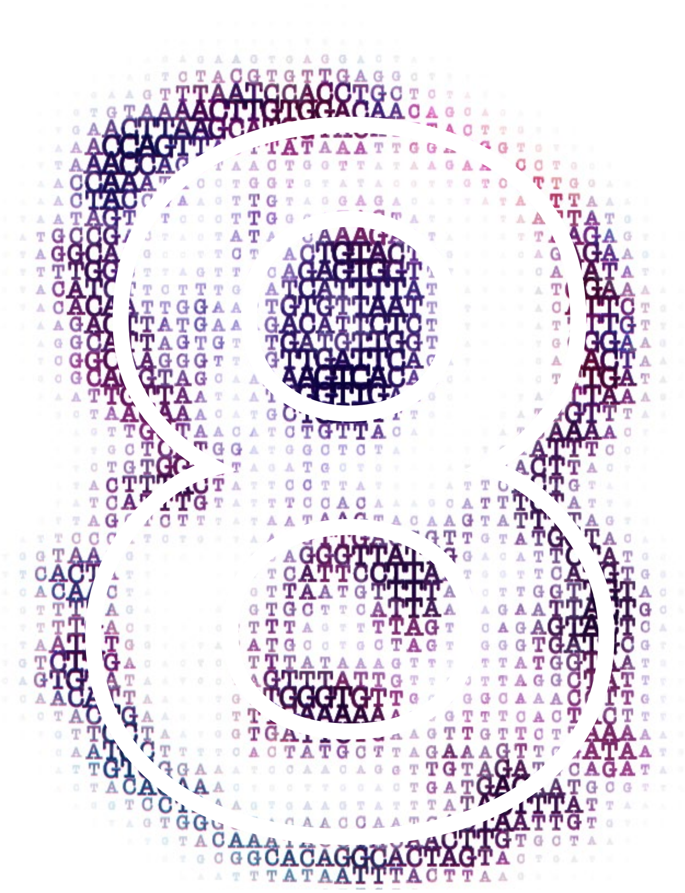
[27] K. Deforche, ‘An alignment method for nucleic acid sequences against annotated genomes’, Bioinformatics, preprint, Oct. 2017. doi: 10.1101/200394.

[28] B. G. van den Hoogen, T. M. Bestebroer, A. D. M. E. Osterhaus, and R. A. M. Fouchier, ‘Analysis of the Genomic Sequence of a Human Metapneumovirus’, Virology, vol. 295, no. 1, pp. 119–132, Mar. 2002, doi: 10.1006/viro.2001.1355.

[29] H. Zhou et al., ‘A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein’, Curr. Biol., vol. 30, no. 11, pp. 2196-2203.e3, Jun. 2020, doi: 10.1016/j.cub.2020.05.023.

[30] E. C. Carbo et al., ‘Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics’, J. Clin. Virol., p. 104566, Jul. 2020, doi: 10.1016/j.jcv.2020.104566.





Chapter 8 A comparison of five Illumina, Ion torrent, and nanopore sequencing technology-based approaches for whole genome sequencing of SARS-CoV-2

Ellen C. Carbo [1], Kees Mourik [1], Stefan A. Boers [1], Bas Oude Munnink [2], David Nieuwenhuijse [2], Marcel Jonges [3], Matthijs R.A. Welkers [3], Sebastien Matamoros [3], Joost van Harinxma thoe Slooten [1], Margriet Kraakman [1], Evita Karelioti [4], David van der Meer [4], Karin Ellen Veldkamp [1], Aloys C.M. Kroes [1], Igor Sidorov [1], Jutte J.C. de Vries [1]

1 Clinical Microbiological Laboratory, Department of Medical Microbiology, Leiden University Medical Center; Leiden, the Netherlands
2 Department of Viroscience, Erasmus Medical Centre, Rotterdam, the Netherlands
3 Department of Medical Microbiology and Infection Prevention, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands
4 GenomeScan B.V., Leiden, the Netherlands

Adjusted version of article published in European Journal of Clinical Microbiology & Infectious Diseases

Abstract

Rapid identification of the rise and spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants of concern currently remains critical for monitoring of the efficacy of diagnostics, therapeutics, vaccines, and control strategies. A wide range of SARS-CoV-2 next-generation sequencing (NGS) methods have been developed over the last years, but cross-sequence technology benchmarking studies are scarce. In the current study, 26 clinical samples were sequenced using five protocols: AmpliSeq SARS-CoV-2 (Illumina), EasySeq RC-PCR SARS-CoV-2 (Illumina/NimaGen), Ion AmpliSeq SARS-CoV-2 (Thermo Fisher), custom primer sets (Oxford Nanopore), and capture probe-based viral metagenomics (Roche/Illumina). Studied parameters included genome coverage, depth of coverage, amplicon distribution, and variant calling.

The median SARS-CoV-2 genome coverage of samples with cycle threshold (Ct) values of 30 and lower ranged from 81.6 to 99.8 for, respectively, the Oxford Nanopore protocol and Illumina Ampliseq protocol. Correlation of coverage with PCR Ct-values varied and was dependent on the protocol. Amplicon distribution signatures differed across the methods, with peak differences of up to 4 log10 at disbalanced positions in samples with high viral loads (Ct-values ≤ 23). Phylogenetic analyses of consensus sequences showed clustering independent of the workflow used. The proportion of SARS-CoV-2 reads in relation to background sequences, as a (cost-)efficiency metric, was highest for the EasySeq protocol. The hands-on time was lowest when using EasySeq and ONT protocols, with the latter additionally having the shortest sequence runtime.

In conclusion, the studied protocols differed on a variety of the studied metrics. This study provides data that can assist laboratories when selecting protocols for their specific setting.

Keywords:

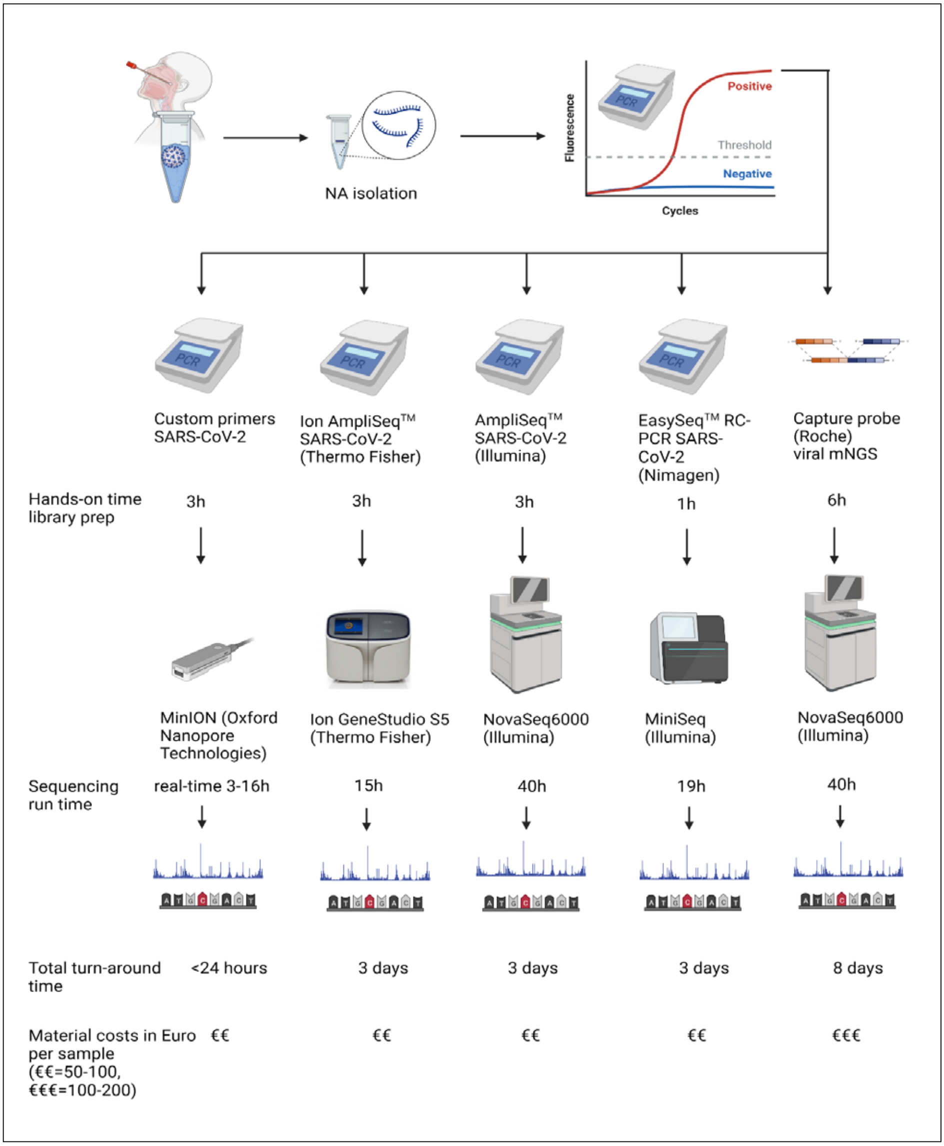
Whole genome sequencing; SARS-CoV-2; benchmark

Introduction

Genomic surveillance of severe acute respiratory syndrome coronavirus (SARS-CoV-2) has proven critical for early detection of the rise and spread of SARS-CoV-2 variants of concern, for monitoring and developing effective diagnostic, therapeutic, and preventive strategies [1-3]. In addition, genomic surveillance assists in contact tracing, transmission tracking at population level, and public-health decision making [4]. The widespread application of genomics for pandemic surveillance is exemplified by the more than 10 million SARS-CoV-2 sequences deposited in the GISAID repository as of April 2022 [5].

A wide range of SARS-CoV-2 next-generation sequencing (NGS) technologies and protocols have been developed and adapted since the first genome sequence was generated using a metagenomic approach [6-8]. SARS-CoV-2 whole genome sequencing (WGS) protocols have been improved to increase the technical performance, including sensitivity and genome coverage, and logistical aspects have also been addressed, such as scalability and hands-on time [9-12]. Studies have been published on SARS-CoV-2 WGS with innovative protocol adaptations in order to decrease the error rate and the turn-around-time by combining PCR and tagging steps [12]. However, these studies are typically focused on the technology developed by the authors, whereas comparison of a novel protocol with other methods is limited. Benchmark studies of SARS-CoV-2 genome sequencing technologies are scarce and generally restricted to comparison of protocols for the single type of sequencing technology available at the study site of the authors [13-15]. In contrast, cross-platform studies are still relatively scarce [16,17]. A recent external quality assessment (EQA) report assessed the outcome of complete workflows from nucleic acid extraction to the reported consensus sequence by testing SARS-CoV-2 cultured isolates; however, no detailed distinction between the different workflow components could be made [16].

Here, we describe a cross-platform benchmark study that includes Illumina, Ion torrent, and nanopore-based SARS-CoV-2 sequencing technologies in one study. Five protocols (Figure 1), employing a diversity of sequencers with a wide range of throughput, accuracy and runtime were compared using clinical samples. The performance was studied by comparing genome coverage, read depth, amplicon distribution, variant calling, and the proportion of on-target reads.



	Custom primers SARS-CoV-2	Ion AmpliSeq™ SARS-CoV-2 (Thermo Fisher)	AmpliSeq™ SARS-CoV-2 (Illumina)	EasySeq™ RC-PCR SARS-CoV-2 (Nimagen)	Capture probe (Roche) viral mNGS
Hands-on time library prep	3h	3h	3h	1h	6h
	MinION (Oxford Nanopore Technologies)	Ion GeneStudio S5 (Thermo Fisher)	NovaSeq6000 (Illumina)	MiniSeq (Illumina)	NovaSeq6000 (Illumina)
Sequencing run time	real-time 3-16h	15h	40h	19h	40h
Total turn-around time	<24 hours	3 days	3 days	3 days	8 days
Material costs in Euro per sample (€€=50-100, €€€=100-200)	€€	€€	€€	€€	€€€

Figure 1. Schematic overview of the design, workflow, and technologies adopted in this study.

Twenty-six respiratory samples, mainly nasopharyngeal swabs and tracheal aspirates, were tested by five SARS-CoV-2 WGS protocols. PCR Ct-values ranged from 13.9-33.6. To exclude potential variability resulting from different nucleic acid extraction methodologies, the extraction method used was identical for all five protocols. Four protocols were tiled amplicon based, one protocol was capture probe based, targeting all viruses known to infect vertebrates. In order to minimize potential differences resulting from variation in bioinformatic analyses tools and settings, a uniform pipeline for sequence data from Illumina and Ion platforms, for ONT data, platform-specific tools handling higher error rates were used to gain optimal results from this type of dataset (Suppl. Figure 1). Created using Biorender.com.

Methods

Sample selection

In total, 26 SARS-CoV-2 PCR positive samples of 24 patients were selected: nine tracheal aspirates, 16 nasopharyngeal/throat swabs, and one lung lymph node biopsy. Fifteen of these samples were obtained for cluster identification. Samples were retrospectively included to be tested with five WGS protocols. Samples were previously sent to the Clinical Microbiological Laboratory of the Leiden University Medical Center (LUMC, the Netherlands) for SARS-CoV-2 PCR testing in the period March - October 2020 (Wuhan-like viruses circulating). As previously described [18], and stored at -80 °C until WGS analysis. In total 26 samples with a wide range of Ct-values (13.9-33.6, confirmed by re-testing) were included to assess the performance of each of the five WGS protocols. The range and distribution of PCR Ct-values was chosen based on relevance for routine clinical practice.

Ethical approval

Approval was obtained from the ethical committee of the LUMC (B20.002, Biobank Infectious Diseases 2020-03), and the Institutional Review Board of the LUMC for observational Covid-19 studies (CoCo 2021-006).

Extraction of nucleic acids

To exclude potential variability resulting from different nucleic acid extraction methodologies, the extraction method used was identical for all five protocols. Nucleic acids were extracted from 200 ul plasma using the MagNApure96 DNA and Viral NA small volume extraction kit on the MagNAPure 96 System (Roche Diagnostics, Almere, the Netherlands) with 100 ul output eluate.

SARS-CoV-2 sequencing protocols (see also Figure 1)

Ampliseq SARS-CoV-2 sequencing (Illumina)

Libraries were prepared using the AmpliSeqTM SARS-CoV-2 Research Panel for Illumina®, which is a targeted RNA/cDNA amplicon assay for epidemiological research of the SARS-CoV-2 virus. This panel contains a two pool design of 247 amplicons/primer pairs (pool 1: 125 amplicons, pool 2: 122 amplicons). In total, 237 amplicons were SARS-CoV-2 targets while the remaining amplicons mapped to five different regions of the human genome and were used as control. The amplicons' lengths ranged from 125 to 275 bp. From each sample, 15 ul of eluate was concentrated using the Speedvac vacuum concentrator (Eppendorf, Hamburg, Germany). Samples were then dissolved in 10 µl AmpliSeq cDNA synthesis master mix. Next, the AmpliSeq cDNA Synthesis for Illumina Kit (Illumina) was used to reverse transcribe RNA to cDNA. Amplicon primer pools of the AmpliSeqTM SARS-CoV-2 Research Panel for Illumina® were subsequently added to each sample. cDNA target amplification reaction was performed according to manufacturer's instructions, followed by partial digestion of primer dimers. AmpliSeq CD indexes were then ligated and further library PCR amplification was performed. The libraries were purified with the AgencourtTM AMPureTM XP Reagent (Beckman Coulter). The final quality and quantity of each barcoded cDNA library was determined using the Fragment Analyzer (Agilent). From all amplified libraries, 2 µl was pooled and loaded for a short sequencing run to indicate the size of the intact libraries. Based on the indicative read counts, equimolar amounts of each sample were pooled (1.1 nM) and submitted for DNA sequencing using the NovaSeq6000 system (Illumina, San Diego, CA, USA) according to manufacturer's protocols. Approximately 10 million 150 bp paired-end reads were obtained per sample. Data processing was performed in real-time by the NovaSeq Control Software v1.7.

EasySeq RC-PCR SARS-CoV-2 sequencing (NimaGen/Illumina)

Libraries were prepared using the EasySeq RC-PCR SARS-CoV-2 kit version 4.02 (NimaGen) for Illumina as described by Coolen et al [12]. cDNA synthesis was performed using the iScriptTM Advanced cDNA Synthesis Kit (Bio-Rad) according to manufacturer's instructions using 10 ul of eluate. This version of the EasySeq RC-PCR SARS-CoV-2 kit uses 154 designed primer pairs (pool A and B) with a tiling strategy, resulting in approximately 435 bp size amplicons. The EasySeq protocol enables a one-step procedure for adding SARS-CoV-2 target specific PCR primers, sequence adapters and Unique Dual Indices (UDI's) by hybridization of the SARS-CoV-2 primers with universal primers that include adapters and UDI's. After the PCR with 5 µl cDNA as input, samples were pooled based on Ct value into pool A and B, which were individually cleaned using AmpliCleanTM Magnetic Bead PCR Clean-up Kit (NimaGen, Nijmegen, The Netherlands). Subsequently, quantification was performed using the Qubit double strand DNA (dsDNA) High Sensitivity assay kit on a Qubit 4.0 instrument (Life Technologies) and pool A and B were combined. Sequencing was performed on Illumina MiniSeq® using a Mid Output Kit (2 × 149 or 2 × 151-cycles) (Illumina, San Diego, CA, USA) by loading 0.8 pM on the flowcell, obtaining approximately 50,000 paired-end reads per sample. The sequence runs were conducted using a balanced library pooling strategy based on estimated cDNA input according to the manufacturer's protocol.

Ion AmpliSeq SARS-CoV-2 sequencing (Thermo Fisher)

The Ion AmpliSeq SARS-CoV-2 research panel supplied by Thermo Fisher Scientific contained 247 primer pairs designed to cover the SARS-CoV-2 genome with 125 to 275 bp overlapping amplicons. For cDNA synthesis, the SuperScipt VILO cDNA Synthesis Kit (11754050, ThermoFisher Scientific, The Netherlands) was used according to manufacturer's instructions using 7 µl of diluted nucleic acid solution to an estimated input of 100 copies/reaction using nuclease free water (AM9939, Ambion, Thermo Fisher Scientific, The Netherlands). SARS-CoV-2 whole genome amplification, adapter ligation and purification were performed using the Ion AmpliSeq SARS-CoV-2 Insight Research Assay (A51305, Thermo Fisher Scientific, The Netherlands) according to manufacturer's instruction. Libraries were quantified using the Ion Library TaqMan Quantitation Kit (4468802, Thermo Fisher Scientific, The Netherlands) according to manufacturer's instructions. Samples were then sequenced on an Ion GeneStudio S5 system (ThermoFisher Scientific, The Netherlands) using an Ion 540 chip (ThermoFisher Scientific, The Netherlands), obtaining approximately up to 1 million paired-end reads per sample.

Custom primers with MinION sequencing (ONT)

A SARS-CoV-2 specific multiplexed PCR for nanopore sequencing was performed using custom-made primers as previously described [4]. In short, primers for 89 overlapping amplicons spanning the whole SARS-CoV-2 genome were designed using primal [19]. The amplicon length was approximately 500 bp with a 75 bp overlap between the different amplicons. cDNA was transcribed using SuperScript III Reverse Transcriptase (Invitrogen, Darmstadt, Germany) [20]. Libraries were generated using the native barcode kits from Oxford Nanopore Technologies (EXP-NBD104, EXP-NBD114 and SQK-LSK109) using 5µl cDNA as input, and sequenced on a R9.4 flow cell multiplexing 96 samples per sequence run (Oude Munnink et al). On average, 68k reads with an average size of 423 bp were obtained per sample.

Capture probe (Roche) with viral metagenomic NGS (Illumina)

The viral metagenomic NGS protocol has previously been described [21-23]. After nucleic acid extraction, 50 µL of eluate was concentrated with the SpeedVac vacuum concentrator (Eppendorf, Hamburg, Germany) and dissolved in 10 µl fragmentation master mix (NEBNext). The NEBNext Ultra II Directional RNA Library prep kit (New England Biolabs, Ipswich, MA, USA) for Illumina was used for RNA library preparation, incorporating several alterations to the manufacturer's protocol to be able to detect both DNA and RNA in the sample. Specifically, poly-A mRNA capture isolation, rRNA depletion and DNase treatment steps were omitted and dual indexed adaptors were used. The SeqCap EZ Hypercap probes (Roche, Basel, Switzerland) were designed in 2015 to cover 207 taxa genomes of viruses known to infect vertebrates including humans [24]. Recently, it has been shown that the probes cover >99% of the SARS-CoV-2 genome [25] due to similarity with bat coronaviruses and the variability incorporated in the probe design. Viral DNA enrichment was performed using the SeqCap EZ HyperCap Workflow User's Guide in pools of four amplified DNA libraries with overnight probe incubation. Washing and recovering captured DNA was performed using the HyperCap Target Enrichment kit and HyberCap Bead kit. Lastly, post-capture PCR amplification was performed with KAPA HiFi HotStart ReadyMix (2X) and Illumina NGS primers following manufacturers' instructions, followed by AMPure bead purification. The quality and quantity of the post-capture multiplexed libraries were assessed by Fragment Analyzer (Agilent) or Bioanalyzer (Agilent, Santa Clara, CA, USA). Sequencing was performed on the NovaSeq6000 system (Illumina, San Diego, CA, USA) obtaining approximately 10 million 150 bp paired-end reads per samples.

Data analyses

In order to minimize potential differences resulting from variation in analysis tools and settings, a uniform pipeline for QC, trimming, mapping, and variant calling was used for sequence data from Illumina and Ion platforms (Supplementary Figure 1). For ONT data, platform-specific mapping and variant calling tools handling higher error rates were used to gain optimal results from this type of dataset.

Illumina data from AmpliSeq, EasySeq and viral metagenomic protocols

Demultiplexing was performed according to Illumina manufacturer protocol using bcl2fastq v2.20 (Illumina). Removal of duplicate reads was not performed since unique molecular identifiers (UMI's) in principle are not compatible with the non-random, tiled amplicon based WGS protocols in the current study, and were thus not incorporated in any of the wet lab procedures described here. Quality control and trimmings per read was performed utilizing Trimmomattic v0.36 [26]. To remove and count the number of sequence reads mapping to the human genome, reads were mapped to GRCh38 using Bowtie2 v2.1.0 [27]. Unmapped reads were subsequently mapped to the SARS-CoV2 genome NC_045512.2 [28]. Mapped reads were indexed in a genome sorted bam file by Samtools v1.7 [29,30]. Variant calling was done using Bcftools v.1.7 [31].

Ion AmpliSeq data

Primer-removed fastq-files were exported for further analysis using the Torrent Suite Software (ThermoFisher Scientific, The Netherlands). Per read quality control was performed using Trimmomatic v0.36 [26]. The resulting quality checked reads were first mapped to the human reference genome HG19 using BWA v0.7.17 [32] with default settings ("bwa bwasw") to remove all reads of potential human origin. Unmapped reads were subsequently mapped to the SARS-CoV-2 refence genome Wuhan-Hu-1 [33]. The resulting sequence alignment map (SAM) files were converted to BAM, sorted and indexed using SAMtools v1.14 [29,30]. Variant calling was performed using Bcftools v.1.7 [31].

ONT custom primers data

Demultiplexing was performed using Porechop v0.2.4 [34]. Primers were trimmed using Cutadapt v3.0 [35]. Reference-based alignment was carried out using Minimap2 v2.17-r941 [36] against both the human genome GRCH38 and SARS-CoV-2 genome NC_045512.2 [28]. Variant calling was performed by filtering of variants using the Python module Pysam v 0.16.0.1 [37].

Performance and statistical analyses

Mapping coverage was analysed using a threshold of 10x depth per base for all platform data except for ONT data, where a 20x depth per base was considered as threshold to ensure reliable variant calling [38]. Coverages per base were calculated using Samtools v1.7 [29,30] with the corresponding depth option. Correlation between genome coverage percentage and Ct-values was calculated using Spearmans' rho [39]. Read mapping quality and base quality (phred) were computed using Samtools v.11 [29,30] with the coverage option. High mapping quality represents a more unique alignment and low mapping quality represents a marginal difference between the alignment and the best secondary alignment option within the reference. High phred scores represent accurate base calling.

Phylogenetic trees

Maximum likelihood trees of the consensus genomes from all methods was generated using the Samtools consensus option [29], Clustal Omega v1.2.4 [39], FastTree v2.1.11 [40,41], and IQTree [42]. Consensus genomes with ≥98% genome coverage were included, genome coverages based on minimal 10x read depth for all methods, and 20x read depth for ONT sequencing. Variant frequencies of >50% were implemented in the consensus genome, though error profiles, like those of ONT, and short insertions/deletions (indels) not consistently called by Samtools can lead to an inaccuracy of the consensus.

Results

In total 26 clinical samples from 24 patients were sequenced using the five SARS-CoV-2 sequencing protocols included in the current comparison: AmpliSeq SARS-CoV-2 (Illumina), EasySeq RC-PCR SARS-CoV-2 (Nimagen/Illumina), Ion AmpliSeq SARS-CoV-2 (Thermo Fisher), custom SARS-CoV-2 primers-based (Oxford Nanopore), and capture probe (Roche) viral mNGS (Figure 1). Additional protocol characteristics, such as hands-on time and sequence runtime are listed in Suppl. Table 1. The breadth of genome coverage, depth of genome coverage, proportion of SARS-CoV-2 reads, and performance of variant calling were compared.

Genome coverage

SARS-CoV-2 genome coverages were generated using a 10x read depth threshold per base for Illumina and Ion Torrent data, and 20x for ONT sequence data (Figure 2, and Suppl. Table 2, incl. normalised read depth per 100,000 total reads.) (Baker et al). As anticipated, amplicon-based protocols generally resulted in higher genome coverage rates compared to the probe hybridization-based metagenomics protocol, though median genome coverages using the custom primer ONT protocol were within the same range for samples with Ct-values of ≤30 (81.2% for ONT and 86.7% for mNGS, Suppl. Table 2). The median genome coverage across the other three amplicon-based protocols was comparable for samples with Ct-values of ≤30: respectively 99.7% and 99.8% when using the Ion AmpliSeq and the Illumina AmpliSeq protocol, followed by the EasySeq protocol for Illumina (98.05%). An increase in Ct-values resulted in only limited reduction of genome coverage when using the Ion AmpliSeq (R = -0.327) and Illumina AmpliSeq (R = -0.523) protocols. When considering all samples, including high Ct values the genome coverage differed greatly between the amplicon-based protocols.

The median read depth of coverage per position ranged from 316 when using the Illumina EasySeq protocol to 860 when using ONT, and >2000 for the Ion AmpliSeq and the probe hybridization-based metagenomics protocol. This depended on the throughput of the platform and kit, the total number of reads requested, and the number of samples multiplexed.

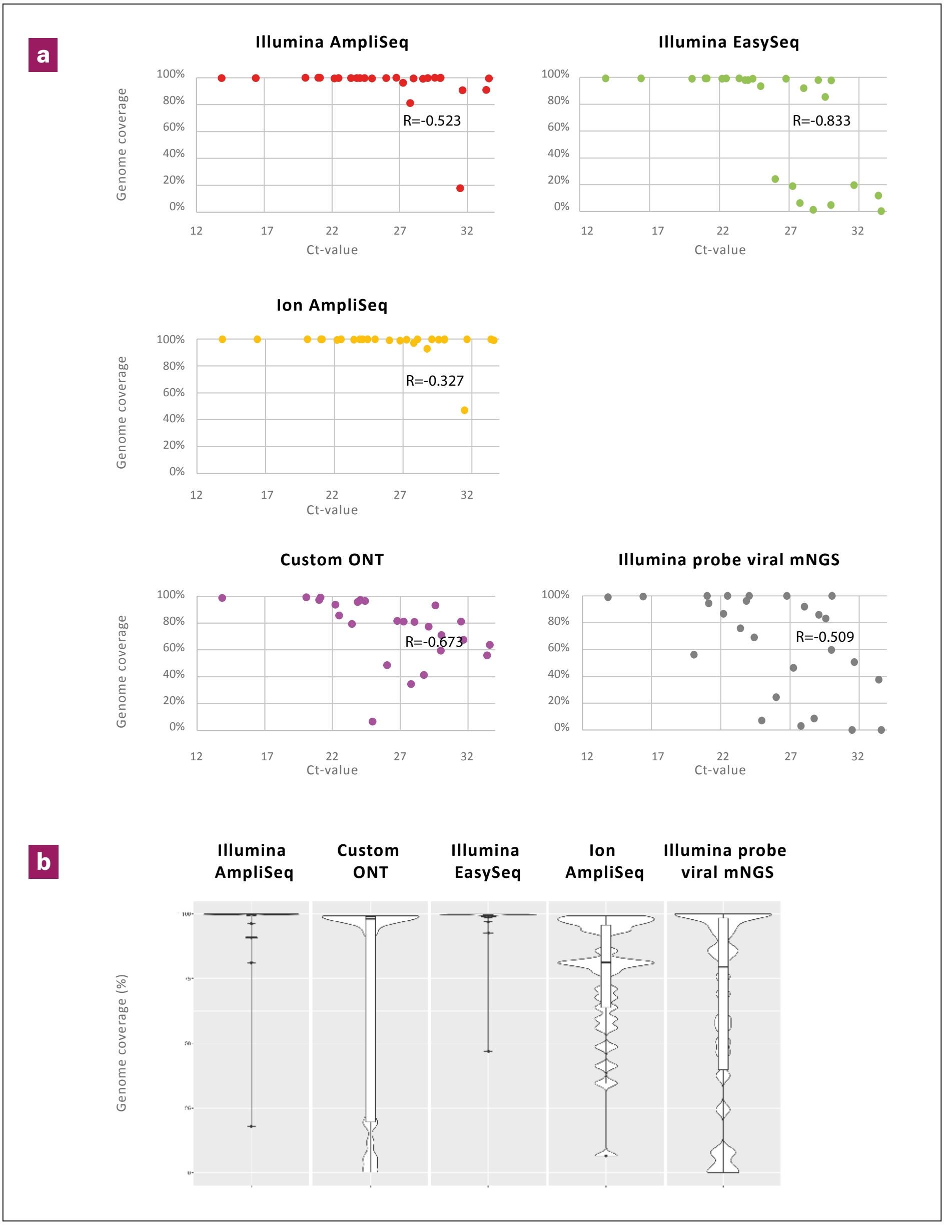


Figure 2. Proportion of SARS-CoV-2 genome coverage of sequencing reads using the five protocols compared. The scatter plots (a) indicate the SARS-CoV-2 genome (NC_045512.2) coverage per PCR Ct-values, each dot represents a single sample. A threshold of 10x depth per base was considered for all platform data except for ONT data, were a 20x depth per base was considered as threshold ensuring reliable variant calling. R values represent Spearmans' correlation coefficient (rho). The violin plots (b) indicate the distribution of the proportion covered per protocol, horizontal markers indicate the median, and the interquartile range.

SARS-CoV-2 amplicon balance

The SARS-CoV-2 amplicon balance was assessed by evaluating the distribution of sequence reads across the SARS-CoV-2 genome. The average read depth per genome position was computed for a selection of nine samples with the highest viral loads (Ct-values ranging from 13-23) (Figure 3). When comparing the genome coverage profiles across the five protocols, distinct signatures were observed for each method. The read depth was most even when using the Illumina AmpliSeq protocol, in contrast to the uneven depth obtained using the probe hybridization-based protocol. The difference in depth between depth of coverage peaks and dips varied generally 2 log10-fold when using the Illumina AmpliSeq protocol, up to 4 log10-fold for the probe-based viral metagenomics protocol. When examining the differences in read depths in more detail, certain positions had protocol dependent, structural lower read depth for multiple samples. An example of a protocol with a structural drop of depth (to 0-11X read depth per sample) was observed at genome position 4,117- 4,149 (ORF1a) when using the Illumina AmpliSeq and Ion Ampliseq protocols. These findings were indicative of a primer failure caused by a specific SNV. The custom ONT protocol resulted in several samples with a low read depth in the amplicons spanning the regions 2,690-2,715 and 6,260-6,490 (ORF1a). Hybridisation probe viral mNGS resulted in the largest regions with low coverage, especially regions 1,000-10,000 (ORF1a) and 22,250-23,000 (Spike), with the last one at risk for missing mutations in the spike protein.

Variant calling and phylogenetic analysis

To assess the performance of variant calling across the protocols, consensus sequences were aligned to the SARS-CoV-2 reference NC_045512.2; SNVs detected per protocol are depicted in Suppl. Table 3. Consensus sequences used to build a phylogenetic tree for samples in which ≥4 protocols had a genome coverage of 98% and higher (n=14 samples). In the phylogenetic tree where gaps in the sequence (uncovered positions and indels) were considered a match with the reference sequence (Figure 4a), consensus genomes of specific samples clustered independent of the used protocol and analysis pipeline. However, when gaps were simply masked in the pairwise comparison (affecting solely the denominator, the total number of positions counted), for highly identical sequences (lower part of the tree) some per protocol clustering was also observed across Illumina, Ion, ONT and probe-based technologies, up to 0.005 substitutions/site distances between methods (Figure 4b). These findings indicate the effect of gaps in sequences in relation to the type of cluster analyses in case of highly identical sequences.

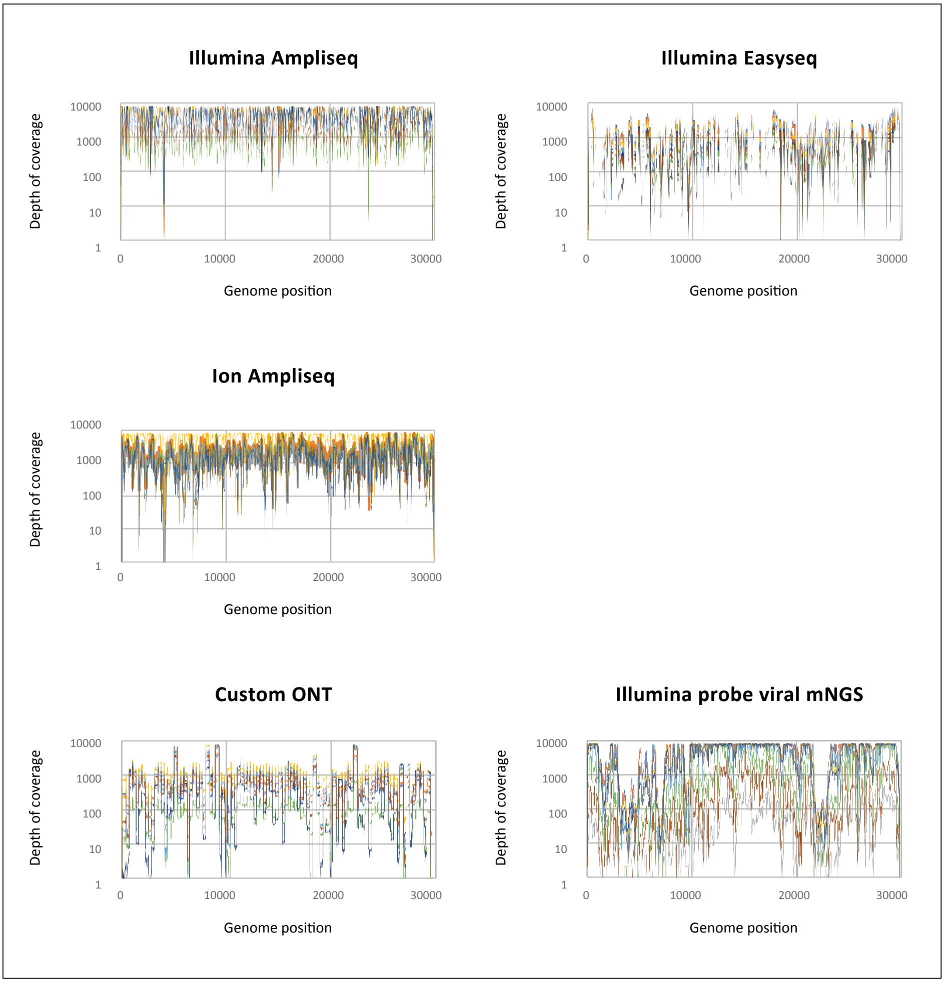


Figure 3. Distribution of sequence read depth over the SARS-CoV-2 genome using the five protocols compared.

The number of sequence reads (logarithmic scale) per SARS-CoV-2 genome (NC_045512.2) position, using the five protocols compared. A selection of nine samples with higher viral loads (Ct-values ranging from 13-23) is visualized. Each color represents an individual sample.

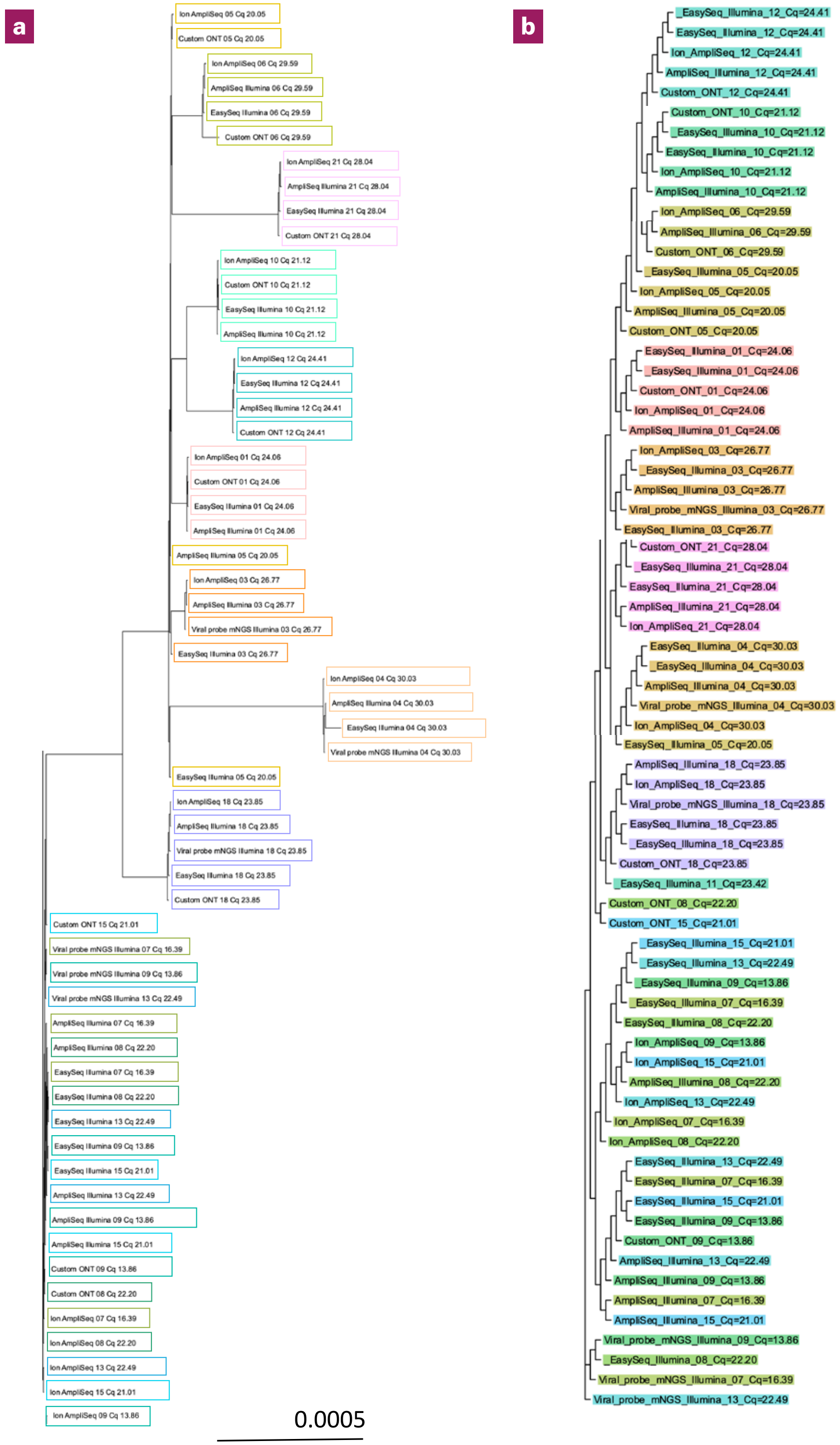

a
Ion AmpliSeq 05 Cq 20.05
Custom ONT 05 Cq 20.05
Ion AmpliSeq 06 Cq 29.59
AmpliSeq Illumina 06 Cq 29.59
EasySeq Illumina 06 Cq 29.59
Custom ONT 06 Cq 29.59
Ion AmpliSeq 21 Cq 28.04
AmpliSeq Illumina 21 Cq 28.04
EasySeq Illumina 21 Cq 28.04
Custom ONT 21 Cq 28.04
Ion AmpliSeq 10 Cq 21.12
Custom ONT 10 Cq 21.12
EasySeq Illumina 10 Cq 21.12
AmpliSeq Illumina 10 Cq 21.12
Ion AmpliSeq 12 Cq 24.41
EasySeq Illumina 12 Cq 24.41
AmpliSeq Illumina 12 Cq 24.41
Custom ONT 12 Cq 24.41
Ion AmpliSeq 01 Cq 24.06
Custom ONT 01 Cq 24.06
EasySeq Illumina 01 Cq 24.06
AmpliSeq Illumina 01 Cq 24.06
AmpliSeq Illumina 05 Cq 20.05
Ion AmpliSeq 03 Cq 26.77
AmpliSeq Illumina 03 Cq 26.77
Viral probe mNGS Illumina 03 Cq 26.77
EasySeq Illumina 03 Cq 26.77
Ion AmpliSeq 04 Cq 30.03
AmpliSeq Illumina 04 Cq 30.03
EasySeq Illumina 04 Cq 30.03
Viral probe mNGS Illumina 04 Cq 30.03
EasySeq Illumina 05 Cq 20.05
Ion AmpliSeq 18 Cq 23.85
AmpliSeq Illumina 18 Cq 23.85
Viral probe mNGS Illumina 18 Cq 23.85
EasySeq Illumina 18 Cq 23.85
Custom ONT 18 Cq 23.85
Custom ONT 15 Cq 21.01
Viral probe mNGS Illumina 07 Cq 16.39
Viral probe mNGS Illumina 09 Cq 13.86
Viral probe mNGS Illumina 13 Cq 22.49
AmpliSeq Illumina 07 Cq 16.39
AmpliSeq Illumina 08 Cq 22.20
EasySeq Illumina 07 Cq 16.39
EasySeq Illumina 08 Cq 22.20
EasySeq Illumina 13 Cq 22.49
EasySeq Illumina 09 Cq 13.86
EasySeq Illumina 15 Cq 21.01
AmpliSeq Illumina 13 Cq 22.49
AmpliSeq Illumina 09 Cq 13.86
AmpliSeq Illumina 15 Cq 21.01
Custom ONT 09 Cq 13.86
Custom ONT 08 Cq 22.20
Ion AmpliSeq 07 Cq 16.39
Ion AmpliSeq 08 Cq 22.20
Ion AmpliSeq 13 Cq 22.49
Ion AmpliSeq 15 Cq 21.01
Ion AmpliSeq 09 Cq 13.86
0.0005
b
_EasySeq_Illumina_12_Cq=24.41
EasySeq_Illumina_12_Cq=24.41
Ion_AmpliSeq_12_Cq=24.41
AmpliSeq_Illumina_12_Cq=24.41
Custom_ONT_12_Cq=24.41
Custom_ONT_10_Cq=21.12
_EasySeq_Illumina_10_Cq=21.12
EasySeq_Illumina_10_Cq=21.12
Ion_AmpliSeq_10_Cq=21.12
AmpliSeq_Illumina_10_Cq=21.12
Ion_AmpliSeq_06_Cq=29.59
AmpliSeq_Illumina_06_Cq=29.59
Custom_ONT_06_Cq=29.59
_EasySeq_Illumina_05_Cq=20.05
Ion_AmpliSeq_05_Cq=20.05
AmpliSeq_Illumina_05_Cq=20.05
Custom_ONT_05_Cq=20.05
EasySeq_Illumina_01_Cq=24.06
_EasySeq_Illumina_01_Cq=24.06
Custom_ONT_01_Cq=24.06
Ion_AmpliSeq_01_Cq=24.06
AmpliSeq_Illumina_01_Cq=24.06
Ion_AmpliSeq_03_Cq=26.77
_EasySeq_Illumina_03_Cq=26.77
AmpliSeq_Illumina_03_Cq=26.77
Viral_probe_mNGS_Illumina_03_Cq=26.77
EasySeq_Illumina_03_Cq=26.77
Custom_ONT_21_Cq=28.04
_EasySeq_Illumina_21_Cq=28.04
EasySeq_Illumina_21_Cq=28.04
AmpliSeq_Illumina_21_Cq=28.04
Ion_AmpliSeq_21_Cq=28.04
EasySeq_Illumina_04_Cq=30.03
_EasySeq_Illumina_04_Cq=30.03
AmpliSeq_Illumina_04_Cq=30.03
Viral_probe_mNGS_Illumina_04_Cq=30.03
Ion_AmpliSeq_04_Cq=30.03
EasySeq_Illumina_05_Cq=20.05
AmpliSeq_Illumina_18_Cq=23.85
Ion_AmpliSeq_18_Cq=23.85
Viral_probe_mNGS_Illumina_18_Cq=23.85
EasySeq_Illumina_18_Cq=23.85
_EasySeq_Illumina_18_Cq=23.85
Custom_ONT_18_Cq=23.85
_EasySeq_Illumina_11_Cq=23.42
Custom_ONT_08_Cq=22.20
Custom_ONT_15_Cq=21.01
_EasySeq_Illumina_15_Cq=21.01
_EasySeq_Illumina_13_Cq=22.49
_EasySeq_Illumina_09_Cq=13.86
_EasySeq_Illumina_07_Cq=16.39
EasySeq_Illumina_08_Cq=22.20
Ion_AmpliSeq_09_Cq=13.86
Ion_AmpliSeq_15_Cq=21.01
AmpliSeq_Illumina_08_Cq=22.20
Ion_AmpliSeq_13_Cq=22.49
Ion_AmpliSeq_07_Cq=16.39
Ion_AmpliSeq_08_Cq=22.20
EasySeq_Illumina_13_Cq=22.49
EasySeq_Illumina_07_Cq=16.39
EasySeq_Illumina_15_Cq=21.01
EasySeq_Illumina_09_Cq=13.86
Custom_ONT_09_Cq=13.86
AmpliSeq_Illumina_13_Cq=22.49
AmpliSeq_Illumina_09_Cq=13.86
AmpliSeq_Illumina_07_Cq=16.39
AmpliSeq_Illumina_15_Cq=21.01
Viral_probe_mNGS_Illumina_09_Cq=13.86
_EasySeq_Illumina_08_Cq=22.20
Viral_probe_mNGS_Illumina_07_Cq=16.39
Viral_probe_mNGS_Illumina_13_Cq=22.49
0.005

Figure 4. Tree of likelihood ratios based on consensus sequences of samples with genome coverages of ≥98% for each of the protocols.

Phylogenetic trees were build base on consensus sequences resulting from each of the protocols (FastTree [41,41] and IQTree [42]). For readability, a magnification is shown that includes samples with ≥98% genome coverage for four or more of the protocols (14 samples). A threshold of 10x depth per base was considered for all platform data except for ONT data, were a 20x depth per base was considered. Each color represents an individual sample. Clustering was independent of the protocol (a) IQTree, gtr [42], (b), however when gaps in the sequences (deletions and uncovered positions) were masked instead of considered as matches, in cases of closely related sequences (lower part of the tree) also clustering per protocol was detected.

SARS-CoV-2 sequencing efficiency: proportion of SARS-CoV-2 reads

To assess the efficiency of the protocols for sequencing SARS-CoV-2 genome in relation to background sequences, the proportion of SARS-CoV-2 read counts per sample, as opposed to human and other (bacterial) read counts, were computed (Figure 5). As anticipated, the proportion of SARS-CoV-2 sequences was higher for amplicon-based protocols in comparison to the hybrid capture-based protocol, but differed considerably among the last. The proportion of SARS-CoV-2 specific reads varied from 73.72% on average when using the Illumina EasySeq protocol, down to 8.19% on average when using the Illumina probe viral mNGS protocol. Mapping percentages of human reads ranged from 0.03%-99.87% for Illumina and Ion torrent amplicon-based protocols up to 69.98% on average for the Illumina probe viral mNGS protocol, with the long read ONT workflow resulting in the lowest number of human reads. Samples with an inefficient amplification, resulting in a low percentage of SARS-CoV-2 reads, showed a reverse pattern in the percentage of human reads (Figure 5). As can be deduced from these findings combined with Figure 2, some protocols with lower SARS-CoV-2 sequence efficiency compensated for these results by deeper sequencing.

Quality performance

To assess the mapping quality scores, representing the probability that a read is misaligned, median mapping quality scores were assessed (Suppl. Table 2). The mapping quality for all protocols was higher than 40, which equals a mapping accuracy of 99.99%. The median base quality (Phred) scores reflecting the estimates of errors emitted by the sequencing machines ranged from Q23.8 (ONT, P_{error} 0.004%) and Q26.6 (Ion, P_{error} 0.002%) to Q36 for Illumina protocols (P_{error} 0.0003%).

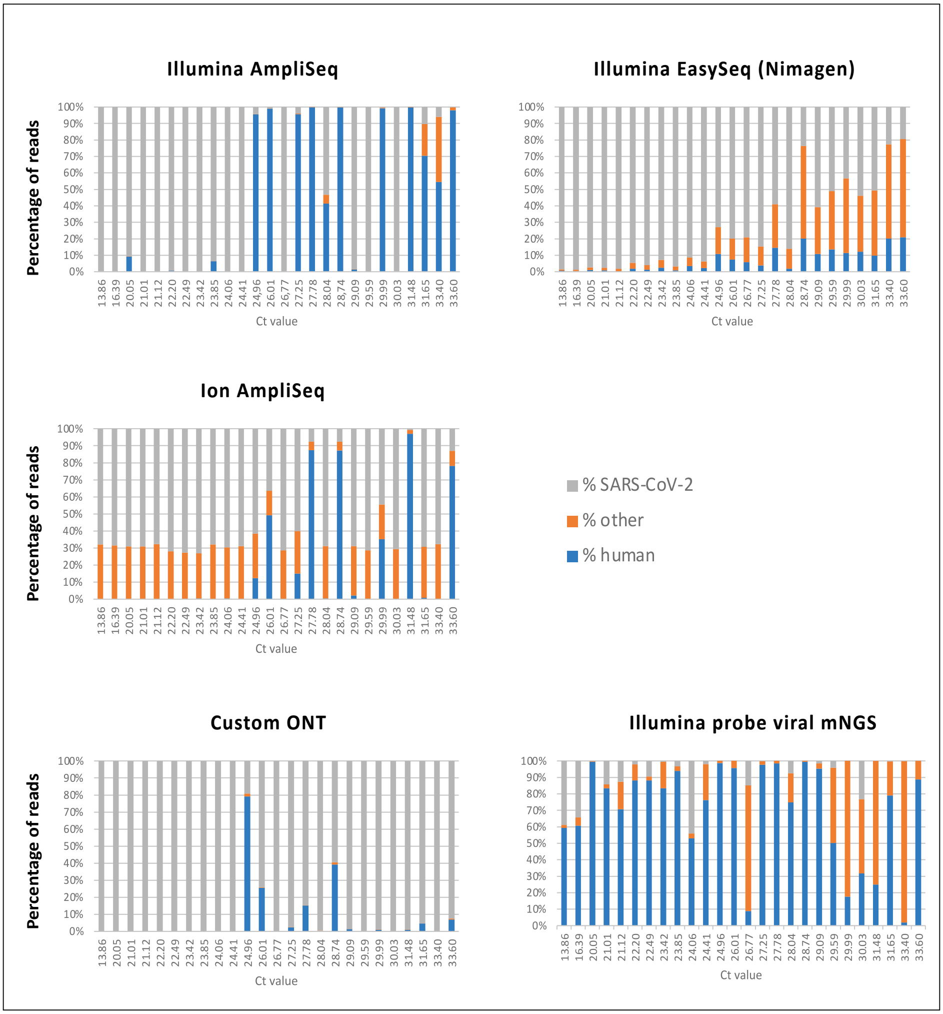


Figure 5. Proportion of SARS-CoV-2 read counts, compared to human and other (bacterial) read counts.

The proportion of SARS-CoV-2, human, and other read counts is shown for each of the five protocols. Each bar (PCR-Ct value) represents an individual sample.

Discussion

In this cross-platform benchmarking using clinical samples, the protocols differed with regard to the varying metrics studied. Each protocol had their own characteristics, advantages and disadvantages. When considering genome coverage, the Illumina and Ion Torrent amplicon-based protocols were in favor. However, amplicon balance was not always even and showed protocol specific drops. Protocols with uneven distribution of sequencing depth among amplicons may benefit from primer redesign or rebalancing of the primer pool to obtain a more even coverage threshold in difficult regions of the genome [37]. Phylogenetic analysis indicated the effect of gaps in sequences in relation to the type of cluster analyses in case of highly identical sequences, possibly resulting from platform-associated effects such as deletion artefacts. This is in contrast to the setting of cluster analyses using sequences obtained using a single platform, since the likelihood of technology-associated characteristics in the sequences may be approximately evenly distributed over the samples. The SARS-CoV-2 sequence efficiency in relation to background sequences was highest for the Illumina EasySeq protocol, comparable with the Ion Ampliseq protocol while the ONT protocol proportionally had the lowest number of human reads. Illumina EasySeq and the ONT protocol had the shortest hands-on time, with the latter additionally having the shortest sequence runtime and real-time data analysis.

As the pandemic continues worldwide and novel variants of interest and variants of concern continue to emerge [43,44], genomic surveillance remains a critical component of the sustained management approach adhered to by the WHO [45]. Accordingly, the need for rapid SARS-CoV-2 genome sequencing protocols that can be easily adopted, automated and that are flexible and scalable remains crucial. Innovative protocol adaptations aiming at high quality sequencing of low viral load samples (Ct-values >30) [11], inherent part of the diagnostic practice, have recently been reported, and such contributions may benefit the worldwide sequence community dedicated to surveillance. Implementation and compatibility of sequence regimes are influenced by characteristics of the local laboratory settings such as the availability of local resources and sequencing platforms with high or low-throughput nature. Reduction of the hands-on time needed for library preparation and overall turnaround time, scalability, and increased cost-efficiency of protocols would be beneficial in broader settings. Here, we aimed to provide data that can assist laboratories when selecting protocols for their local setting by comparing five platforms.

Drops in read depth of certain amplicons were detected in this study using different protocols. Regions with low read depth can result from i) low amplicon coverage by design. High coverage regions have been correlated by coverage of multiple amplicons, whereas genome regions with coverage by only one amplicon resulted in low coverage [13]. Low read depth can also result from ii) a SARS-CoV-2 variant resulting in primer mismatch in that particular amplicon, iii) low efficiency of matching primers in multiplex reactions, or iiii) an imbalance of the primer concentrations present in the multiplex. In our study, the length in bp of the drop in read depth assisted the distinction between single nucleotide variants resulting in a primer mismatch and low coverage by design as underlying cause. Besides low coverage, another factor that can compromise SNV detection are primer-originated "contaminated" sequences that are PCR-amplified [13]. Wet lab methods, and similarly bioinformatic tools can influence the performance of variant detection. Inaccurate trimming of primer sequences can mask or introduce SNVs located in the primer binding site, however our study was not designed to detect such a phenomenon. Also, for example, Minimap2 [35], designed for analyses of sequences from relatively high error-rate platforms, allows considerable mismatches in the alignment with the reference sequence, whereas more stringent mapping tools can result in an absence of coverage in the mutated region. Differentiation of these type of effects resulting from analyses would require a design with cross-comparison of bioinformatic tools, which was not part of the current study. Finally, the current study was restricted by our sample collection time frame (2020), thus our analyses did not contain the later emerged mutants.

Viral (DNA/RNA) metagenomic sequencing has increasingly been adopted for pathogen diagnostics, microbiome analyses, and transcriptome analyses. The focus of the current study specifically was based on SARS-CoV-2 sequencing and specific protocols to enrich for SARS-CoV-2. Metagenomic methods work well for high-throughput sequencing of samples with high viral loads but did not perform the most stable and accurate for low viral load samples, however they were the original clinical request at a time where commercial kits had not been developed yet. This exemplifies the benefit of the approach in earlier stages of pandemics. In later stages of the pandemic it appeared beneficial to have protocols available which also work for lower viral load samples.

Importantly, with the above described pursuing emergence of variants, there is a vital need for sequencing-based approaches that tolerate mutations [46]. Probe capture-based approaches can tolerate large target sequence differences of ~10%

or more from probe sequences [47,48] in comparison with primer-based approaches . These characteristics have resulted in FDA emergency-use-authorization for hybridization-based SARS-CoV-2 genome sequencing in September 2021, in order to improve genomic surveillance of SARS-CoV-2 variants, for tracking viral evolution and guiding vaccine updates [49].

In summary, in this study five cross-platform protocols for SARS-CoV-2 genome sequencing were benchmarked and evaluated on both technical performance and practicality. The results of our study build upon previous reports by providing additional comparison data testing Illumina, Ion Torrent and ONT sequencing in parallel, incorporating technically innovative protocol steps including several analysis workflows. These data will be specifically of assistance for the sequence laboratories dedicated to ongoing surveillance efforts.

Funding

None

Conflicts of interest

The authors to have no conflicts of interest to declare.

Supplementary File Information

Suppl. Figure 1. Overview of bioinformatic analyses tools used in the current study (created using Biorender.com).

Suppl. Table 1. Protocol characteristics of the five SARS-CoV-2 sequence methods compared in the current study. NA; not applicable

Suppl. Table 2. Overview of SARS-CoV-2 PCR Ct-values per sample, genome coverage, mean depth, normalised depth, mean base quality, and mean mapping quality, per sequencing protocol. Normalised depth was calculated per 100,000 total reads.

Suppl. Table 3. Overview of SNPs and indels called by the different protocols (Q13 threshold). A threshold of 10x depth per base was considered for all platform data except for ONT data, were a 20x depth per base was considered. Per genome position, read depths are shown for respectively reference and alternate calls (DP4; ref forward, ref backward, alternate forward, alternate backward counts). *; no variant called or no coverage of position

References

[1] W. T. Harvey et al., 'SARS-CoV-2 variants, spike mutations and immune escape', Nat Rev Microbiol, vol. 19, no. 7, pp. 409–424, Jul. 2021, doi: 10.1038/s41579-021-00573-0.

[2] K. Tao et al., 'The biological and clinical significance of emerging SARS-CoV-2 variants', Nat Rev Genet, vol. 22, no. 12, pp. 757–773, Dec. 2021, doi: 10.1038/s41576-021-00408-x.

[3] Z. Chen et al., 'Global landscape of SARS-CoV-2 genomic surveillance and data sharing', Nat Genet, vol. 54, no. 4, pp. 499–507, Apr. 2022, doi: 10.1038/s41588-022-01033-y.

[4] B. B. Oude Munnink et al., 'Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands', Nat Med, vol. 26, no. 9, pp. 1405–1410, Sep. 2020, doi: 10.1038/s41591-020-0997-y.

[5] 'https://www.gisaid.org/', Apr. 2022.

[6] F. Wu et al., 'A new coronavirus associated with human respiratory disease in China', Nature, vol. 579, no. 7798, pp. 265–269, Mar. 2020, doi: 10.1038/s41586-020-2008-3.

[7] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, 'Shotgun metagenomics, from sampling to analysis', Nat Biotechnol, vol. 35, no. 9, pp. 833–844, Sep. 2017, doi: 10.1038/nbt.3935.

[8] M. Chiara et al., 'Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities', Briefings in Bioinformatics, vol. 22, no. 2, pp. 616–630, Mar. 2021, doi: 10.1093/bib/bbaa297.

[9] M. Simonetti et al., 'COVseq is a cost-effective workflow for mass-scale SARS-CoV-2 genomic surveillance', Nat Commun, vol. 12, no. 1, p. 3903, Dec. 2021, doi: 10.1038/s41467-021-24078-9.

[10] S. H. Rosenthal et al., 'Development and validation of a high throughput SARS-CoV-2 whole genome sequencing workflow in a clinical laboratory', Sci Rep, vol. 12, no. 1, p. 2054, Dec. 2022, doi: 10.1038/s41598-022-06091-0.

[11] H. Choi, M. Hwang, D. H. Navarathna, J. Xu, J. Lukey, and C. Jinadatha, 'Performance of COVIDSeq and Swift Normalase Amplicon SARS-CoV-2 Panels for SARS-CoV-2 Genome Sequencing: Practical Guide and Combining FASTQ Strategy', J Clin Microbiol, vol. 60, no. 4, pp. e00025-22, Apr. 2022, doi: 10.1128/jcm.00025-22.

[12] J. P. M. Coolen et al., 'SARS-CoV-2 whole-genome sequencing using reverse complement PCR: For easy, fast and accurate outbreak and variant analysis.', Journal of Clinical Virology, vol. 144, p. 104993, Nov. 2021, doi: 10.1016/j.jcv.2021.104993.

[13] T. Liu et al., 'A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples', iScience, vol. 24, no. 8, p. 102892, Aug. 2021, doi: 10.1016/j.isci.2021.102892.

[14] J. A. Nasir et al., 'A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture', Viruses, vol. 12, no. 8, p. 895, Aug. 2020, doi: 10.3390/v12080895.

[15] M. Xiao et al., 'Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples', Genome Med, vol. 12, no. 1, p. 57, Dec. 2020, doi: 10.1186/s13073-020-00751-4.

[16] F. Wegner et al., 'External Quality Assessment of SARS-CoV-2 Sequencing: an ESGMD-SSM Pilot Trial across 15 European Laboratories', J Clin Microbiol, vol. 60, no. 1, pp. e01698-21, Jan. 2022, doi: 10.1128/JCM.01698-21.

[17] J. Plitnick, S. Griesemer, E. Lasek-Nesselquist, N. Singh, D. M. Lamson, and K. St. George, 'Whole-Genome Sequencing of SARS-CoV-2: Assessment of the Ion Torrent AmpliSeq Panel and Comparison with the Illumina MiSeq ARTIC Protocol', J Clin Microbiol, vol. 59, no. 12, pp. e00649-21, Nov. 2021, doi: 10.1128/JCM.00649-21.

[18] M. Zlei et al., 'Absence of rapid T cell control corresponds with delayed viral clearance in hospitalised COVID-19 patients', In Review, preprint, Aug. 2021. doi: 10.21203/rs.3.rs-783703/v1.

[19] J. Quick et al., 'Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples', Nat Protoc, vol. 12, no. 6, pp. 1261–1276, Jun. 2017, doi: 10.1038/nprot.2017.066.

[20] V. M. Corman et al., 'Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR', Euro Surveill., vol. 25, no. 3, 2020, doi: 10.2807/1560-7917.ES.2020.25.3.2000045.

[21] E. C. Carbo et al., 'Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics', Journal of Clinical Virology, p. 104566, Jul. 2020, doi: 10.1016/j.jcv.2020.104566.

[22] S. van Boheemen et al., 'Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients', The Journal of Molecular Diagnostics, vol. 22, no. 2, pp. 196–207, Feb. 2020, doi: 10.1016/j.jmoldx.2019.10.007.

[23] A. L. van Rijn et al., 'The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease', PLoS ONE, vol. 14, no. 10, p. e0223952, Oct. 2019, doi: 10.1371/journal.pone.0223952.

[24] T. Briese et al., 'Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis', mBio, vol. 6, no. 5, pp. e01491-15, Sep. 2015, doi: 10.1128/mBio.01491-15.

[25] E. C. Carbo et al., 'Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness', Journal of Clinical Virology, vol. 131, p. 104594, Oct. 2020, doi: 10.1016/j.jcv.2020.104594.

[26] A. M. Bolger, M. Lohse, and B. Usadel, 'Trimmomatic: a flexible trimmer for Illumina sequence data', Bioinformatics, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[27] B. Langmead, 'Aligning Short Sequencing Reads with Bowtie', Curr. Protoc. Bioinform., vol. 32, no. 1, Dec. 2010, doi: 10.1002/0471250953.bi1107s32.

[28] 'https://www.ncbi.nlm.nih.gov/nuccore/1798174254', Apr. 2021.

[29] H. Li et al., 'The Sequence Alignment/Map format and SAMtools', Bioinformatics, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[30] H. Li, 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', Bioinformatics, vol. 27, no. 21, pp. 2987–2993, Nov. 2011, doi: 10.1093/bioinformatics/btr509.

[31] P. Danecek et al., 'Twelve years of SAMtools and BCFtools', GigaScience, vol. 10, no. 2, p. giab008, Jan. 2021, doi: 10.1093/gigascience/giab008.

[32] H. Li and R. Durbin, 'Fast and accurate long-read alignment with Burrows-Wheeler transform', Bioinformatics, vol. 26, no. 5, pp. 589–595, Mar. 2010, doi: 10.1093/bioinformatics/btp698.

[33] 'https://www.ncbi.nlm.nih.gov/nuccore/MN908947'.

[34] GitHub - rrwick/Porechop: adapter trimmer for Oxford Nanopore reads.

[35] M. Martin, 'Cutadapt removes adapter sequences from high-throughput sequencing reads', EMBnet j., vol. 17, no. 1, p. 10, May 2011, doi: 10.14806/ej.17.1.200.

[36] H. Li, 'Minimap2: pairwise alignment for nucleotide sequences', Bioinformatics, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191..

[37] 'https://github.com/pysam-developers/pysam'.

[38] D. J. Baker et al., 'CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes', Genome Med, vol. 13, no. 1, p. 21, Dec. 2021, doi: 10.1186/s13073-021-00839-5.

[39] C. Spearman, 'The Proof and Measurement of Association between Two Things', The American Journal of Psychology, vol. 15, no. 1, p. 72, Jan. 1904, doi: 10.2307/1412159.

[40] F. Sievers and D. G. Higgins, 'Clustal Omega', Current Protocols in Bioinformatics, vol. 48, no. 1, Dec. 2014, doi: 10.1002/0471250953.bi0313s48.

[41] M. N. Price, P. S. Dehal, and A. P. Arkin, 'FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix', Molecular Biology and Evolution, vol. 26, no. 7, pp. 1641–1650, Jul. 2009, doi: 10.1093/molbev/msp077.

[42] M. N. Price, P. S. Dehal, and A. P. Arkin, 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments', PLoS ONE, vol. 5, no. 3, p. e9490, Mar. 2010, doi: 10.1371/journal.pone.0009490.

[43] A. Maxmen, 'Are new Omicron subvariants a threat? Here's how scientists are keeping watch', Nature, vol. 604, no. 7907, pp. 605–606, Apr. 2022, doi: 10.1038/d41586-022-01069-4.

[44] E. Callaway, 'Are COVID surges becoming more predictable? New Omicron variants offer a hint', Nature, vol. 605, no. 7909, pp. 204–206, May 2022, doi: 10.1038/d41586-022-01240-x.

[45] 'https://www.who.int/westernpacific/news-room/feature-stories/item/moving-from-pandemic-response-to-sustained-management-of-covid-19-in-the-western-pacific-region'.

[46] N. M. Butler, P. A. Atkins, D. F. Voytas, and D. S. Douches, 'Generation and Inheritance of Targeted Mutations in Potato (Solanum tuberosum L.) Using the CRISPR/Cas System', PLoS ONE, vol. 10, no. 12, p. e0144591, Dec. 2015, doi: 10.1371/journal.pone.0144591.

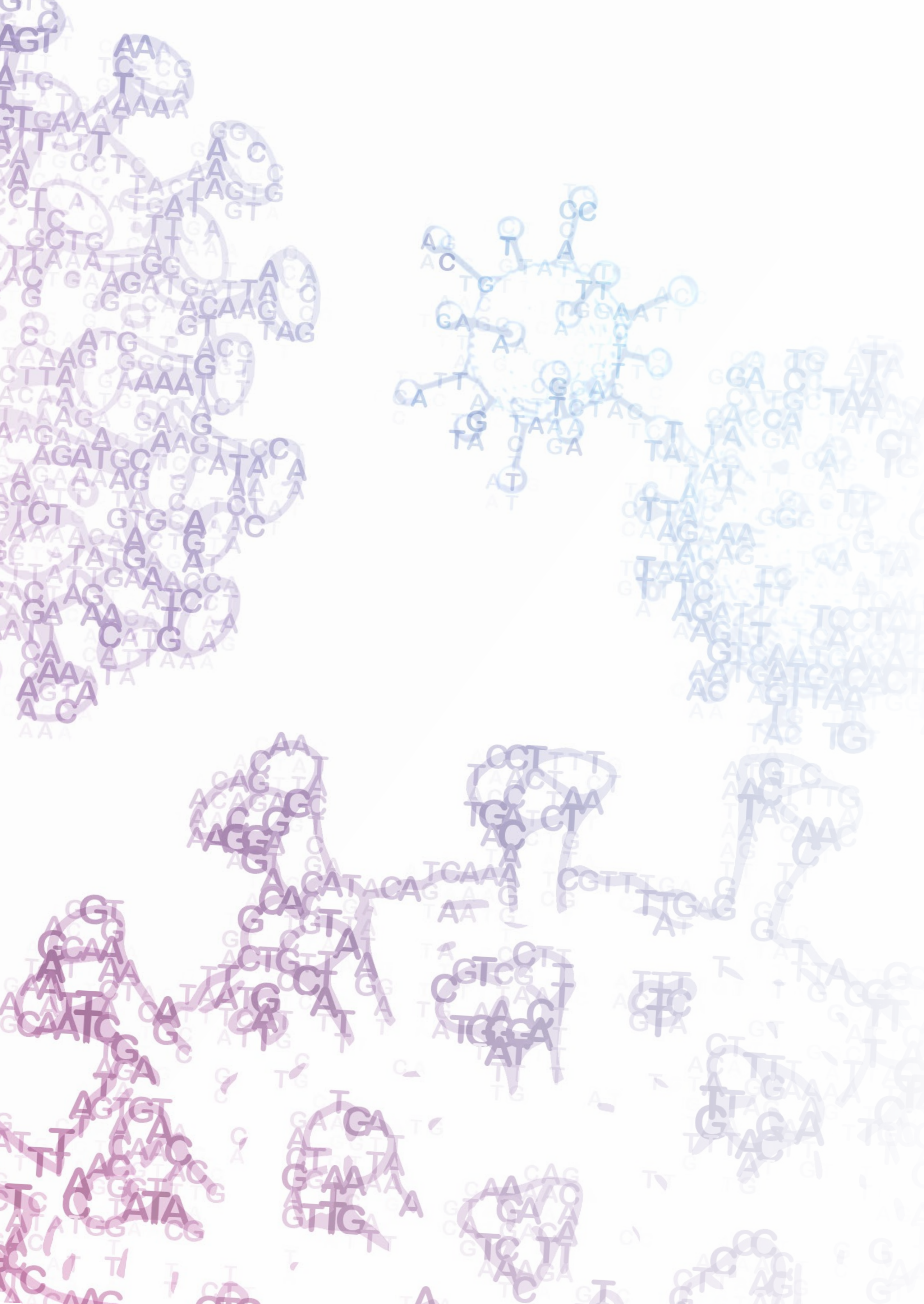
[47] 'https://www.twistbioscience.com/resources/white-paper/effects-mismatches-dna-capture-hybridization', Apr. 2022.

[48] 'https://apps.who.int/iris/handle/10665/338480', Apr. 2022.

[49] D. Nagy-Szakal et al., 'Targeted Hybridization Capture of SARS-CoV-2 and Metagenomics Enables Genetic Variant Discovery and Nasal Microbiome Insights', Microbiol Spectr, vol. 9, no. 2, pp. e00197-21, Oct. 2021, doi: 10.1128/Spectrum.00197-21.

The human mind treats a new idea
the same way the body treats
a strange protein; it rejects it.

Peter Medawar, in “The Art of the Soluble”, 1960



Chapter 9 **General Discussion**

Viral metagenomic next-generation sequencing (mNGS), an approach to potentially identify all viral genomes in a sample at once, is a promising contribution to the current virus diagnostic repertoire in modern health care. With already more than 1,000 virus species known to be able to infect humans [1], a densely populated civilization and a constant threat of zoonotic infections [2], it is a worthwhile addition to the current methods in which either one virus is tested (traditional PCR test), or a limited number of viruses when combined PCR tests are used. This discussion will initially focus on the applications, diagnostic yield and potential of viral metagenomic sequencing. Further, mNGS diagnostic test accuracy advancement, both within the wet laboratory and using bioinformatics, will then be discussed. Additionally, in-depth advances in the genetics analysis of whole genome sequencing of a single-virus genome will be explained. An in-depth view on the limitations of metagenomic sequencing and an outlook on the future of molecular diagnostics will be presented in the last two sections.

Implementing viral metagenomic sequencing

Viral metagenomics improves diagnostic yield

With various viruses that can infect humans and many undiagnosed cases [3-7], implementing metagenomic sequencing in a clinical setting will potentially lead to the identification of more viruses and an increased number of patients diagnosed with a viral infection. One of the aims of the research of this thesis was to assess the improved diagnostic yield using metagenomics: the proportion of additional potential pathogenic viruses that can be found after initial testing remained negative. In **chapter 2**, a systematic review and meta-analysis was conducted and an additional 10.88% (95% CI 4.6-17.15%) of viruses were detected that were not identified by traditional diagnostic testing in patients suffering from meningoencephalitis [8-17]. A selection of reports on patients from (sub)tropical climate regions revealed an additional diagnostic yield of 21.61% (95% CI 12.16-31.07%) partially since the initial test spectrum was more limited, the decreased vaccine administration in this region, and an increased risk of mosquito born viral diseases that are more frequent in (sub) tropical climates. In **chapter 3**, a cohort of hematologic patients suffering from encephalitis was tested and a corresponding additional diagnostic yield of 12.2% (95% CI 2.2-22.2%) was observed.

In **chapter 4**, patient sera were tested from a cohort of international travellers returning with febrile illness, resulting in 6.3% (95% CI -2.4-17.2%) of cases where additional pathogenic viruses were detected. This number seems comparable to the result of a similar study on travellers with febrile illness where in three out of 40 patients (7.5%) extra pathogenic viruses were detected based on mNGS results [18].

Longitudinal testing of transplantation patients by means of metagenomic sequencing is present in **chapter 6**. In this study, BKV, CMV, and HHV6B were additionally detected by mNGS in three out of six patients (50%), and all additional findings were confirmed either by qPCR or supported by auxiliary bioinformatic analysis.

A systematic review and meta-analysis presented in **chapter 1** showed a relatively high number of additional viral findings - 28.73% (CI [19.80, 37.63]) - when assessing studies of diverse patient types that were negative during initial testing and mNGS was used as a second step approach. In the research of this thesis, additional findings were found in 6.3% (95% CI [–2.4, 17.2]) of returning travellers with febrile illness. Two prospective papers describing metagenomics in a clinical setting identified 13 out of 58 central nervous system infections by means of metagenomics that were not found by PCR (22%) [10], and an additional 24 (23%) pathogenic virus infections in 105 patients in a tertiary diagnostic unit [11]. The research of this thesis and available literature show that the use of metagenomics as a second step approach when initial testing is negative improves the diagnostic yield. This accounts for patients suffering from encephalitis where metagenomics can detect a neuroinvasive pathogen [10], for travellers returning with febrile illness, and for immunocompromised patients where unexpected viruses can be detected.

Viral capture probes increase diagnostic test sensitivity

Most previously published studies focused on metagenomics and the test accuracy for the detection of bacteria, with a significant knowledge gap concerning viruses. Two systematic review studies have been published on the overall test performance, with one focusing on metagenomic sequencing for all pathogens including studies prior to August 2020 (note: including papers published before this date) [19], and one focusing on lower respiratory tract infections [20]. A combined overview of two papers focusing (partly) on viruses is shown in Table 1. Wilson et al. [10] showed a relatively low sensitivity of 0.55; however, when looking more into detail in the virus diagnoses missed by mNGS, most of these were found positive in IgM by serology testing while when followed up by qPCR testing these also remained negative. Only two out of 204 results were positive by means of qPCR due to low pathogen titers [10]. In the manuscript by Parize et al., a single viral pathogen was undetected by means of mNGS, attributable to the different sample type that was used: a sample positive human cytomegalovirus (CMV) was identified in whole blood, and for mNGS only plasma was used. When testing the plasma by means of qPCR, CMV was not detected, as CMV probably was residing in leukocytes and not accessible for amplification. Amending this finding would lead to a sensitivity of 100% in this particular study when incorporating only virus data [21]. In the study by Hong et al., a sensitivity of 0.74 was found; however, the portion of mNGS samples resulting negative were found positive only by serological testing [22].

Viral pathogens originally detected by means of PCR were confirmed by viral metagenomics as described in **chapter 3 and 4**, due to the usage of a more sensitive, capture probe-based enrichment, instead of solely performing shotgun metagenomics. In **chapter 6,** a 100% sensitivity was indicated; all initial positive qPCR results were positive by mNGS. Collectively, the majority of published studies and our findings illustrate the high sensitivity of mNGS to identify viruses in samples.

The results of an extensive comparison of shotgun metagenomics with metagenomics using viral capture probes are described in **chapter 3**. Data showed that with shotgun metagenomics several pathogens were marked as false negative, after having a positive diagnostic PCR result. In contrast, metagenomics with viral capture probes performed in a much more sensitive manner, with 1,283-38,749,926 sequence reads per pathogenic virus was found positive by means of PCR. The viral capture probe metagenomic method not only resulted in a sensitivity of 100%, but yielded 100-10,000-fold more sequence reads compared to shotgun metagenomics. An overview of technical aspects of protocols of the few European centres that offer viral metagenomics in a clinical setting is presented in **chapter 2.** It shows that these diagnostic laboratories offering viral metagenomics services are aiming at increased sensitivity by either using viral metagenomic probes, or by performing shotgun metagenomics in parallel for both DNA-based and RNA-based organisms. In **chapter 4,** travellers returning with febrile illness were tested by viral capture probe metagenomics, and all earlier positive PCR test results were confirmed resulting in a sensitivity of 100% of the mNGS method. Transplantation patients that were longitudinally sampled and sequenced using mNGS had positive mNGS results for the viruses that initially tested positive by means of qPCR (cytomegalovirus (CMV), Epstein-Barr virus (EBV), BK polyomavirus (BKV), adenovirus (ADV), parvovirus B19 (B19V), and torque teno-virus (TTV)), resulting in a sensitivity of 100%, as it is shown in **chapter 6**.

Amino acid-based taxonomic classifying tools perform the most accurate

Taxonomic classifiers for virus identification are widely available and use different underlying algorithms [34,35]. A ring trial in Switzerland [36] reported that the chosen algorithms influenced the overall performance of mNGS, more than the chosen reference databases. Only a limited number of studies report benchmarking ‘dry lab’ protocols, despite bioinformatic protocol validation being equally important to wet laboratory validation for accurate performance. Many tools were specifically designed for bacterial detection – such as Kraken [37] and CLARK [38] – and it is especially important to validate these tools for virus identification prior to use for that aim. The limited amount of benchmark publications have focused more on bacterial analysis [39-45], mostly only performing in silico analysis of artificial sequence data [39,46,47], or NGS data for mock samples that are typically less diverse compared to real clinical samples [39,48].

Bioinformatic taxonomic classifiers were benchmarked, as described in **chapter 5**. Up to a billion sequence reads of 88 respiratory samples were used for benchmarking of five classifiers for performance based on results of 1,144 PCR tests used as the gold standard. Sensitivity and specificity of the classifiers tested ranged from 83% to 100% and 90% to 99%, respectively, and was dependent on the classification level and data pre-processing. The bioinformatic tool reaching the highest sensitivity was the Kaiju tool [40] with k-mer classification based on amino acids. Exclusion of human reads generally resulted in increased specificity. Normalization of read counts for genome length resulted in a minor effect on overall performance, however it negatively affected the detection of targets with read counts around detection level.

In a benchmark of the European network of next-generation sequencing [49], datasets from real clinical metagenomic samples (tested positive for viral pathogens) were distributed to thirteen collaborating centres. The optimal performing tool, both for sensitivity and specificity was the MetaMix classification tool [49,50]. This tool, like Kaiju [40] performing the most optimal in **chapter 5**, is based on amino acid identification which, due to lower mutation rates of amino acid compared to DNA/RNA, results in a higher sensitivity, mainly for highly divergent viruses [39,40]. To distinguish contamination from real clinical findings and to further enhance specificity, respectively, tools for removal of sequences detected in negative control samples can be used [51,52], and extra mapping/alignment steps can be added to assess the distribution of sequence reads over the viral genome.

Table 1. **Overview of sensitivity and specificity from reports on (viral) metagenomics.**

Study	Type of sample	Sequencing technique	Gold standard	Sensitivity	Specificity
Hong et al. [22]	Cerebrospinal fluid	Illumina MiSeq	PCR	0.74	0.66
Miller et al. [23]	Cerebrospinal fluid	Illumina HiSeq	Conventional laboratory results and additional molecular testing	0.89	0.99
Wilson et al. [10]	Cerebrospinal fluid	Illumina HiSeq	Conventional laboratory results and additional molecular testing	0.55 Higher for only viruses	0.98
Blauwkamp et al. [24]	Plasma (cfDNA)	Illumina NextSeq 500	Conventional laboratory results and additional molecular testing	0.93	0.63
Parize et al. [21]	Plasma	Ion Proton	Culture, serological diagnosis and PCR	0.63 1.0 (virus only)	0.71
Somasekar et al. [25]	Serum	Illumina HiSeq	PCR	0.96	1
Rossoff et al. [26]	Plasma	Illumina NextSeq 500	Clinical review	0.92	0.64
Schlaberg et al. [27]	Respiratory	Illumina HiSeq 2500	Culture, serological diagnosis and PCR	0.90	0.64
Doan et al. [28]	Intraocular fluid	Illumina HiSeq 4000	PCR	0.87	0.78
Langelier et al. [29]	TA	Illumina HiSeq 4000	Clinical microbiologic testing	1.00	0.88
Wang et al. [30]	Pulmonary biopsy and BALFs	NA	Conventional tests	0.97	0.63
Van Rijn et al. [31]	Nasopharyngeal samples	Illumina NextSeq 500	PCR	0.96	0.98
Huang et al. [32]	Lung tissue, BALF, and PSB	BGISEQ-100	Culture, microscopic examination	0.88	0.81
Van Boheemen et al. [33]	Nasopharyngeal washings, sputa, BALF, bronchial washing and throat swab	Illumina HiSeq 4000 and NextSeq 500	PCR	0.83	0.94

Adapted table of data of two papers focusing on test accuracy and only including papers focusing on viruses or when more than >1 virus found. [19,20]

Abbreviations; BALF, broncho-alveolar lavage fluid; NA, not applicable; PSB, protected specimen brushes; TA, tracheal aspirate. cfDNA, cell free DNA. Clinical review indicates that an organism was classified as clinically relevant by a treating physician, and if unclear was determined by a 2nd paediatric infectious disease (ID) physician, finally relying on the opinion of a third physician in case of discrepant opinions.

Further advantages of metagenomics

A characteristic advantage of metagenomics is that it is a pathogen-agnostic test (Figure 1). No specific pathogen needs to be expected in contrast to a PCR test, or a multiplex of PCR tests. Additionally, mutations occurring in evolving viruses in the primer target regions lead to a false-negative PCR test result whereas viral metagenomic diagnostics would potentially pick up viruses with mutations. In addition, the host transcriptome can be interpreted straight from sequence data after certain shotgun metagenomic protocols.

Metagenomic sequencing results in information about the nucleotide sequences of virus species presented in a given sample, and these sequences can be used for typing and for phylogenetic analyses for these viruses. In **chapter 4,** subsequent typing of viruses detected in the serum samples of travellers resulted in characterization of serotypes and genotypes of the detected viruses, and enabled phylogenetic analysis of the Dengue viruses detected in the serum samples of travellers directly from the metagenomic test results. These results illustrate that viral metagenomic analysis is not only suitable in the detection of extra viruses, but additionally, viruses can be correctly typed, further aiding phylogenetic analysis. Once the nucleotide sequences are established, this information can be used for finding resistance mutations as well.

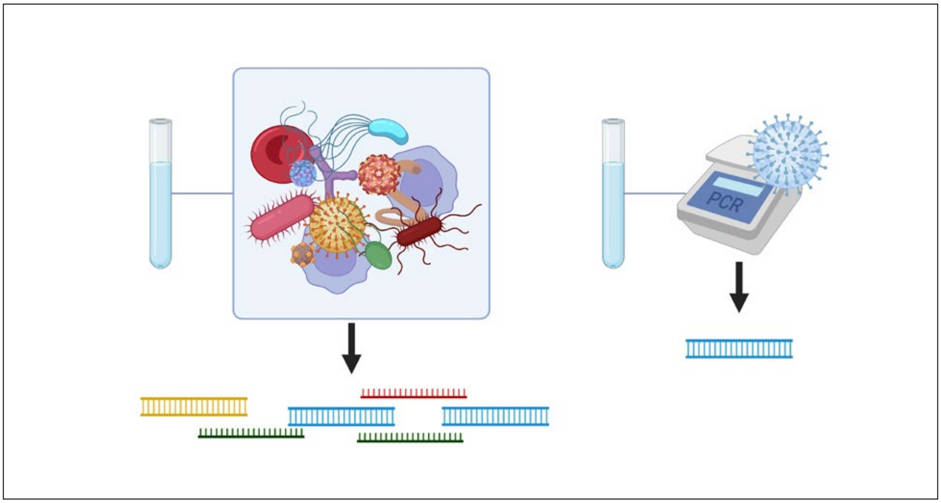


Figure 1. Pathogen-agnostic and unbiased testing using metagenomics versus PCR, testing only known pathogens.

Metagenomic sequencing giving information about all species present in a sample, versus PCR where one or a handful known viruses are tested on being present in a sample. Created using Biorender.com.

Quantification by means of metagenomic sequencing

Quantification of viral load is possible based on metagenomics data. The precision of this greatly depends upon correct classification of the viral pathogen. After sequencing of clinical samples positive for various viruses, normalizing the sequence reads for total read count and genome length, a quantitative correlation between qPCR and metagenomic target reads was found of 62.7% (**chapter 5**). The coefficient of determination varied per bioinformatic tool, data pre-processing, and per virus, and R^2 ranged 15.1-63.4%, with 63.4% scored by amino acid-based classifier. Divergent viruses such as rhinoviruses were the most challenging in assessing correlation of sequence reads with Ct-values. Only a limited number of rhinoviruses are present in the underlying RefSeq database and it could be that precision was decreased as a result, as previously observed in the study of Menzel et al. [40]. In **chapter 6**, longitudinal plasma samples from six patients and qPCR positive for transplantation-related DNA viruses were tested using mNGS in combination with calibration samples. Viral loads as determined based on mNGS results correlated with the qPCR results, with inter-method differences in viral loads per virus ranging from 0.19 log10 IU/mL for EBV to 0.90 log10 copies/mL for ADV. The patterns of viral loads of patients tracked over time based on the metagenomic classifying results resembled that of the loads established by means of qPCR. This was in line with a mNGS report using calibration samples, where identical challenges with torque teno virus (TTV) quantification are discussed as in our study since there was no calibration material available [53]. The results that this paper of Shah et al. describes further imply that viral metagenomic sequencing can be used in a quantitative manner where viral loads are identified straight from metagenomic sequence data [53].

Discovery of viruses directly from clinical samples

Viral metagenomics played a major role in the discovery of SARS-CoV-2 and the characterization of the viral genome when there were several patients in Wuhan presenting with fever and respiratory failure, and screening routine respiratory pathogens for these patients gave negative results [54-56]. **Chapter 7** illustrates that metagenomic sequencing in a clinical setting can be successfully used for virus discovery directly from patient samples. Mimicking virus discovery, using only viruses present in databases from before the discovery of SARS-CoV, MERS-CoV and SARS-CoV-2, revealed that these viruses could be labelled as indicative for a novel coronavirus. Bioinformatic tools Centrifuge and Genome Detective [57] showed classification of reads to the closest relative of the emerging coronavirus. Contig genome assemblies with lengths ranging from 2,503 to 30,097 nucleotides, created out of the patient sequence data, could be linked with low nucleotide identity to coronaviruses present before the emerging virus by means of BLAST [58]. These results validate discovery of these novel viruses direct from clinical respiratory samples. Capture probes designed before the emergence of a virus can aid positive discovery findings, supported by the mismatches that are allowed during capture enrichment, or the presence of many homologic regions in a known closely related virus, resulting in effective virus discovery as long as a virus from the same genus or family is present in the probe kit.

Whole genome sequencing (WGS) of SARS-CoV-2 for surveillance

The genomic surveillance of SARS-CoV-2 is of great importance for monitoring and detection of variants of concern, and for developing diagnostic, therapeutic and preventative strategies [59-61]. The most sequenced pathogen for surveillance currently is SARS-CoV-2 and worldwide consensus sequences can be uploaded to GISAID [62] guiding phylogenetics of a given sample, not only in local test sets but additionally in relation to sequences from globally. This kind of surveillance is mostly performed via WGS of patient samples targeting one specific virus.

In **chapter 7**, sequencing of two SARS-CoV-2 genomes using both a shotgun and a viral metagenomic capture probe method is described, and an increase in genome coverage when using capture probes is demonstrated. Few comparisons have been published, though WGS comparisons are usually limited to a single type of sequencing principle [63-65] whereas only two benchmark studies dealt with cross-platform protocols [66-67]. However, these studies for the most part indicate that amplicon-based methods yield the highest genome coverage. A more extensive comparison including viral probe metagenomic sequencing and several amplicon-based WGS protocols designed for SARS-CoV-2 is shown in **chapter 8**. Amplicon-based WGS protocols gave an overall median genome coverage of 81.6-99.8% (samples with CT-values of 30 and lower), with custom primers for Oxford Nanopore Technology (ONT) performing the lowest, and Illumina Ampliseq protocol resulting in the highest coverage. Amplicon distribution signatures differed across methods, illustrating the need to acquire coverage statistics when interested in certain genes or domains. Phylogenetic clustering of consensus sequences were independent of the workflow used, though in some cases it resulted in clustering per method when using settings where gaps were masked.

The usage of viral metagenomic probes showed an 86.7% median genome coverage, demonstrating that this method can indeed be of aid for limited surveillance when no specific genome amplicon kits are yet available, for instance when concerning novel or emerging viruses.

Challenges in viral metagenomics

General limitations

One of the challenges of current viral metagenomics protocols is the required turnaround time and costs of the NGS technique. Even with the current decline in sequencing costs, metagenomic sequencing is still more expensive compared to testing with PCR. Additionally, whereas PCR can provide results in only less than an hour, metagenomic sequencing takes approximately 2-6 days, depending on the protocol. However, diagnostic departments are becoming less reluctant to use more expensive and time-consuming NGS methods since the pandemic presented them with a great necessity for WGS for surveillance, spending more money on a metagenomic test could save money in other health care departments [68], due to an increased diagnostic yield.

For implementation in clinical settings, standardization of protocol validation is limited, although first attempts for establishing standardized guidelines have been reported [34,69]. Another limitation of metagenomic sequencing is the impairment of the data due to the high abundance of host cell material, and the potential threat of contamination, although contamination can partially be controlled for by sequencing an environmental control.

The research described in this thesis does not include bacterial, fungal or any other pathogenic classification of microorganisms, therefore it presents an overview of viruses and lacks a broader perspective that yields a higher diagnostic potential when looking at all organisms at once. Another general characteristic to take into account is that metagenomic sequencing can lead to incidental findings, such as the hepatitis C virus finding in the cohort of travellers described in **chapter 4** of this thesis, and the HIV findings in a Swiss study [11]. Even though these findings may not always be clinically expected based on the patient's syndrome, when using metagenomic sequencing, clinicians should be aware that there is always a possibility of finding unexpected viral pathogens as bystander infections.

Platform-specific sequence errors

The metagenomic sequencing in this thesis was performed using Illumina sequencing, a platform in which index hopping – the swapping of sample indexes leading to incorrect assignment of reads to a neighbouring sample – can occur. Other sequence platforms not impaired by index swapping were not evaluated in this thesis.

Though this effect can be limited by using dual indexing, adding unique barcodes at both ends of the sequencing reads [70]. Illumina platforms are also known to have a median error rate of 0.109% for the NovaSeq 6000, 0.429% for the NextSeq 500 and 0.613% for the MiniSeq, of which Novaseq6000 and MiniSeq were included in the WGS comparison in **chapter 8**. [71] These error rates might impair a correct establishment of nucleotides, potentially leading to incorrect typing and mutation calling. One of the platforms currently most suited to determine minor variants would be PacBio, resulting in reliable sequenced long reads enabling detection and phasing of variants that are only present in low percentages in a sample [72]. Additional platforms can be used as well, ideally in combination with unique molecular identifiers applied during the library preparation, resulting in a unique label per every single molecule and allowing for amplification error filtering in subsequent bioinformatic analyses [73]. Alternatively, other sequence protocols for labelling unique molecules can be used, for instance single molecule molecular inversion probes [74,75].

Bioinformatics: always a challenge

The performance of metagenomic sequencing is greatly dependent on accurate data analyses after the sequence reads are obtained from the sequencer. Various tools and pipelines exist, though standardized validation formats are lacking. As mentioned above, the majority of tools for classification and assembly are initially built for other organisms than viruses, rendering validation specifically for viruses of the utmost importance. Benchmarks are scarce or based on in silico data sets or mock samples with low abundance of the background sequences [39,46-48]. Misclassification of human genome sequence reads has been reported for several taxonomic classifiers [39], which is in line with our findings (**chapter 5**). This is most likely due to the presence of human genomic host reads in microbial assemblies uploaded to reference databases [76,77]. Other species can also lead to inaccurate uploads to GenBank, for instance the Illumina control phage PhiX174 that is present in many uploaded assemblies [78,79]. This viral phage is often used as a control for Illumina sequence runs and not always completely filtered out of sequence data [80]. Database curation should be improved when the database is used for metagenomic analyses, ideally by admitting only iterative assemblies based on long reads and by applying automated scripts to control for host material and contamination since research has shown that over 2,000,000 entries in Genbank contain cross-kingdom contamination [81]. Despite the high number of virus genome sequences available publicly, the list is incomplete and many virus genomes, especially those of bacteriophages, need to be sequenced and assembled to be added to public databases. Lower numbers of reference genomes available for specific targets lead to decreased sensitivity and specificity [40]. To expand databases, the viral dark matter needs to be identified. Viral dark matter is sequence data resembling viruses though currently not immediately identified by regular classifiers [82,83]. Further bioinformatic issues may arise from the fact that many microbial laboratories lack bioinformaticians or lack the access to a high-performance computing cluster. With several cloud- or web-based user-friendly software tools for viral metagenomic analysis [57,84-86], local removal of human host reads is required before uploading the data, as even with viral metagenomic target probes as with amplicon WGS protocols there are usually human reads present in a sample after sequencing (**chapter 8**).

The future of viral metagenomic sequencing

Viral metagenomic as an add-on test for difficult to diagnose cases

Applying viral metagenomic sequencing, as described in this thesis, resulted in additional viral pathogenic findings, from 6.3 and 10.88% in two of our own cohorts of patients to 28.73% (95% CI [19.80-37.63]) in a systematic review as described in the introduction. To identify causes of infections in, for instance, the 20-62% [87-89] of patients suspected for acute respiratory infection where no microbial agent is detected, or in the up to 63% of encephalitis patients that remain without a causal pathogen [3], viral metagenomics can aid as an add-on test to the current diagnostic repertoire of clinical testing. In the formal diagnostic algorithm of the Dutch Society of Medical Microbiology (NVMM) for the paediatric patients suffering from acute hepatitis in 2022, viral metagenomics is officially advised on biopsies (and plasma or feces) in cases where other results are inconclusive [90]. The use of viral metagenomics may be additionally justified in severely affected infectious patients where no causal pathogenic viral pathogen is found by traditional testing methods. Metagenomic sequencing currently is more expensive compared to traditional tests when including lab costs only; however, recurrent or sequential negative test results in the microbiology department lead to extra costs elsewhere in the health care system. A cost analysis performed on the detection of infectious diseases by mNGS in cases with pyrexia of unknown origin justified implementation of metagenomic sequencing minimally as a second line investigation [68].

Sensitivity rates of pathogen detection have been published of >83% and often >90% (Table 1), and the 100% sensitivity in our research (**chapter 3, 4 and 6**) using viral capture probes indicates that this technique is becoming a trustworthy method to begin to implement in diagnostics. With limits of detection between 10-1,000 copies/ml, viral pathogens do not need to be highly abundant in a patient sample to be detected. With the additional information that can be retrieved from the metagenomic sequencing data for typing, resistance, phylogenetic information and virus discovery, it provides extra information for antiviral treatment, outbreak monitoring and surveillance.

Overcoming technical challenges

Sequencing costs constantly decline, and as of this year the sequencing of a full human genome is possible for $100 [91]. Workflows for WGS library preparation have been made faster by adding sequence adapters in a two-step amplification protocols. However, for shotgun libraries such advances still have to be developed, and these protocols currently take six hours hands-on time (**chapter 8**). Sequencing instruments are becoming much faster in sequencing: whereas the first NGS machines were running for days, Illumina NextSeq 500 now has a minimum runtime of 12 hours, MiSeq minimally four hours, and a recent paper shows that pathogens can be detected from the ONT platform in combination with real-time analysis 30-38 minutes after the start of the Minion sequencer for highly abundant pathogens [92,93]. Besides being fast, ONT sequencers are handheld devices that are relatively cheap for laboratories compared to the investment needed for other sequence platforms. The size is also compact making them more even suitable for remote locations and – in the future – for patient bedside sequencing.

A challenge in metagenomic sequencing is the background level of host sequence reads, and with proper enrichment for viruses or depletion of host material it is easier to detect a potential viral pathogen. Centrifugation, filtration, and DNase treatment have not proven to be effective in every case [32,34,35,95]. Ribosomal RNA depletion and poly-A tail enrichment is sometimes used, though the latter may lead to false negative results in detection of viruses in a non-replicative state or those that translate without poly-A tail [34]. Another comparison of human genome depletion methods has been performed in a microbiome study where selective lysis of cells and endonuclease digestion worked well, and where benzonase increased metagenomic sequencing coverage [95]. However, sizeable benchmarks of host depletion methods are lacking specifically for viral metagenomics.

‘Virome in a bottle’ as a validation sample

Other aspects currently lacking with regard to the implementation of viral mNGS are uniform metagenomic validation samples. Only benchmark samples with limitations, such as cultured mock samples with limited sample sizes, or only in silico samples, are available. However, widely available and uniform benchmark samples resembling backgrounds reflecting real patient samples and containing several viral pathogens with different established viral loads are needed. Such benchmark material, like the “Genome in a Bottle” samples [96] is available for clinical genetics and used for validation in clinical genetic laboratories around the world, would be of great benefit to the metagenomics community [96-98]. The Genome in a Bottle materials are reference samples sold as vials containing human DNA. These samples contain an entire human genome, and even a combination of three human genomes can be bought, of which every known SNP and indel is additionally available in several different file formats. This allows a true reference so that every single mutation found in the lab’s diagnostic process can be accurately checked. An initiative like creating a uniform ‘microbiome/virome in a bottle’ (Figure 2) is greatly needed in the microbiology field, also making benchmarks more comparable within the field.

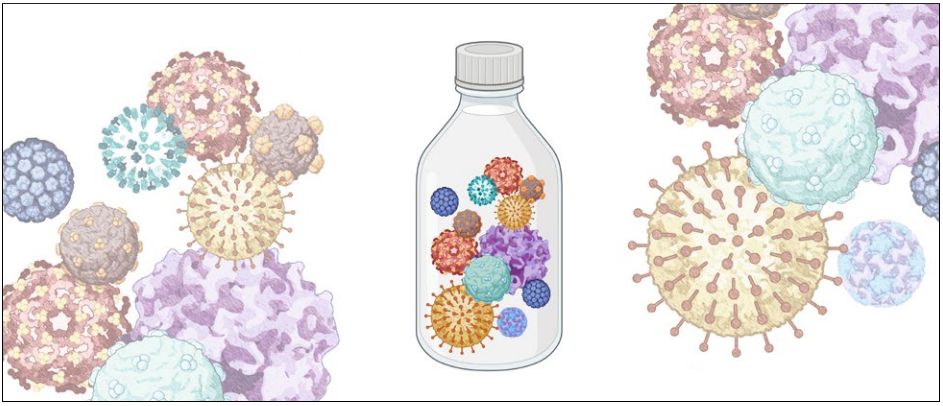


Figure 2. ‘Virome in a bottle’.

Example of a uniform mNGS benchmark sample containing different kinds of viruses. Created using Biorender.com.

Solving computational challenges and solving challenges computationally

Concerning bioinformatics, the microbiology community would benefit from training a higher number of more specialized bioinformaticians or data analysts, not just to only work on microbiology, but also to educate them on FAIR principles [99], and on privacy issues. Laboratory technicians could be trained in performing simple data analysis using a graphical user interface. Departments should not limit their focus on implementing the wet lab part of NGS, but they should additionally think about the hardware, and the costs of running and storing of analysis data. Long read sequencing will aid in distinguishing viral quasi-species, the same species of virus being present in a sample but with variating genomes due to high mutation rates. Bioinformatic tools like haplotype aware variant callers can aid this detection, though these are designed for human genomes and need to be benchmarked for detection of viral quasi-species. High quality quasi-species detection can additionally aid in tracking and surveillance of recombination in viruses [100].

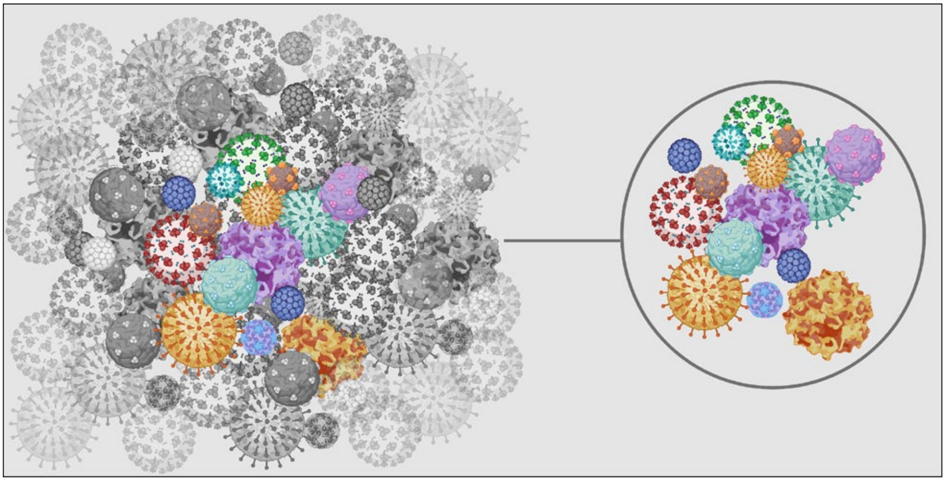


Figure 3. 'Viral dark matter' and relatively few viruses identified.

The coloured viruses represent the identified viruses, the dark grey coloured viruses represent the viruses that show resemblance to the viruses that we already are aware of and are present in our reference databases. The light grey faded viruses represent the viral dark matter, unidentified sequences that make op 40-90% of certain sample types [83]. Expanding databases and methods for detection are needed to further identify these. Created using Biorender.com.

Viral dark matter

Viral dark matter (Figure 3), the uncharacterized sequences, should be explored more to expand public databases, since there are still many sequences resembling viral genetic material though sequences are not directly identified as a virus [82,83]. There is a lot about the microbial world that remains unknown. Creation of databases and tools to identify the unknown are needed, however, some of this information could already be supplied from existing data. We now have unravelled almost all species and substances on land, and humans even explored deep oceans and even space, though a large part of the microorganisms right in front of us and in our bodies are still unidentified. In some samples, 40-90% of the sequences remain unidentified [83] and are thought to make up the viral dark matter, as these sequences resemble viral material though they do not match the reference sets. Detection of these viral dark matter sequences can be performing using a tool called VirSorter [101], detects viral signals based on viral protein resemblance in assembled contigs without sequence data directly matching a known virus family though not with a large resemblance percentage. Many sequences that resemble viruses that are already present in our reference databases are most likely undiscovered bacteriophages. The tools needed for deciphering this data are based on virus discovery tools, though an extensive benchmark for these discovery tools is currently lacking.

Expanding virome databases

Investments would be beneficial to create reference databases of the virome of healthy and affected individuals in various sample types, and this can aid the differentiation between pathogenic and non-pathogenic viruses. This can be partly done based on sequence material that is publicly available and of which the raw data are shared within the science community. Using such a methodology, novel coronaviruses were recently found [102], and for blood a DNA virome [103] and cell free DNA virome [104] is already assembled, still leaving a variety of sample types to be explored. Another opportunity is to classify viral sequences of available data at the cancer genome atlas (TCGA) database [105] or in other public databases with sequence data of cancer patients to find associations between certain types of cancers and viruses [106,107].

Artificial intelligence aiding viral health care

State-of-the-art artificial intelligence (AI) implementation in virology has greatly increased since the SARS-CoV-2 pandemic started. AI models are used for outbreak epidemiology: to provide information on the infection rate, number of cases, transmission dynamics and predicting the development and outcome of an outbreak [108,109]. Low-income countries lacking PCR data could even use mobile health technology [110] by applying AI on survey and sensor data from smart devices to predict the number of positive virus as was done for COVID-19 cases [111,112]. Extensive research has been performed on how AI can aid COVID-19 diagnosis, with a review reporting an accuracy as high as 70.00-99.92% in 46 studies on AI-assisted diagnosis. This included an accuracy of 74.4-95.20% on prognosis of critical COVID-19 patients [108]. AI can additionally be applied for developing therapeutics and vaccines strategies. In the recently published review, an overview of eight studies using AI on COVID-19 is provided, mainly focusing on drug discovery or drug redirecting [108]. One study utilized reverse vaccinology and machine learning to find a vaccine for COVID-19 [108,113], while another study used the data of the GISAID [62] database to find vaccine targets [114]. These AI-assisted methods can be translated to potentially other viruses and outbreaks in the future. Likely, AI techniques may additionally be used for mutation prediction, further exploring and identifying viruses and viral dark matter, and perhaps to predict clinical outcome of pathogens present in metagenomic samples.

Collaborations within the field of diagnostic microbiology

Within the field of microbiology, collaboration is needed with partners worldwide to bridge the gaps that currently exist due to insufficient virome databases, the lack of a suggested validation 'virome in a bottle' sample and to establish a standardized validation approach for NGS protocols. At the beginning of the SARS-CoV-2 pandemic, surveillance in the Netherlands was slightly behind when compared to the UK and Denmark, where a larger number of samples were sequenced compared to the proportion of cases. Currently, a relatively similar number of samples are sequenced in the Netherlands compared to the UK, Iceland, Denmark or Australia [61,115]. The organization of sequencing was scattered early in the pandemic: sequencing largely depended on local initiatives mainly organised by University Medical Centres. On one hand it was positive that these centres thrived in such a fast way, as it was needed to sequence extensively for surveillance. On the other hand, the efficiency of the implementation was questionable since centres were individually testing, optimising and validating a WGS SARS-CoV-2 lab protocol and creating a bioinformatic pipeline for analysis, while better collaboration could have saved time and effort.

The metagenome aiding personalized medicine

Within hospitals, interdisciplinary laboratory departments can combine standardized approaches for isolating nucleotides and sequencing. This type of collaboration can also be applied in bioinformatics, since a variety of tools used in microbiology originate from the human genetics field. There is a variety of clinical information present in patients with samples when looking at the metagenome (Figure 4). One of the potential collaborations based on this approach would be with the pharmacy department for utilizing the pharmacogenetic data to predict an individual's drug response on both a pharmacokinetic level, for instance predicting the metabolising enzyme capacities of an individual, and pharmacodynamic level [116]. Currently, there are already collaborations between microbiology and pharmaceutics departments/

companies for the development of vaccine and anti-viral treatments, though pharmacogenetics is not implemented in daily clinical decision making. With sequencing more samples of patients with infectious diseases, the pharmacogenetics [117] should not be forgotten to facilitate future prescribed drugs in a diagnostic setting: joint use of NGS data can aid in not only selecting drugs targeting specific pathogens, but can additionally be based on genetic variations in the drug-metabolising enzyme genes of the human host for personalized medicine [117].

A pathogen-agnostic test for both pathogens and pathogenic host gene variants

It would be useful to ascertain disease severity and investigate why certain people get more severely ill compared to others with joint insight from human genetics and metagenome sequencing. For instance, genetic variation in the ACE2 gene of the human genome is associated with disease severity in individuals with SARS-COV-2 infection [118-120], and most likely many other associations are yet to be found. Genome-wide sequencing revealed that immunity genes are under pathogen-imposed selection pressure [121], and some differences resulted from specific outbreaks like tuberculosis [122].

Since the course of infection can be influenced by (inherited) autoimmune or autoinflammatory disorders, it would be of great value to have the combined knowledge on both the infection of a patient and any hereditary disorder present interfering with the patient's immune system. Regarding metagenomics, it would be, for instance, beneficial when using a shotgun metagenomics approach for undiagnosed but suspected cases of encephalitis in young patients, to additionally diagnostically screen for genetic pathogenic variants for immunological disorders. Coexistence between autoimmune encephalitis and other systemic auto immune diseases have been previously described [123]. This type of dual application of metagenomics can be used in various kinds of infectious diseases and sample types. Research has already shown that using NGS in severely diseased infants for detection of hereditary mutations will lead to lower morbidity and mortality [124-126]. A more expanded approach, taking into account both the genome of the individual and the metagenomic sequence data, could in the future also lead to similar reduction of morbidity and mortality.

The field of clinical genetics is leaning to a genome-first approach, where genetic variants of interests are agnostically linked to the associated phenotype. Metagenomic sequencing as a combined agnostic test for both pathogens and pathogenic host variants could be the next step in molecular diagnostics (Figure 4). In the future, the wide availability of shotgun metagenomic sequence data of many different sample types and locations that are tested can be of help to the clinical genetics field as well. This is particularly relevant for mosaic mutations. Mosaic mutations are present in very minor fractions in whole blood, but fully penetrant in certain body parts. These mutations can perhaps be earlier detected when testing different sample types from different locations, instead of only testing DNA isolated from whole blood samples, the common utilized sample type in clinical genetics research. [74,75].

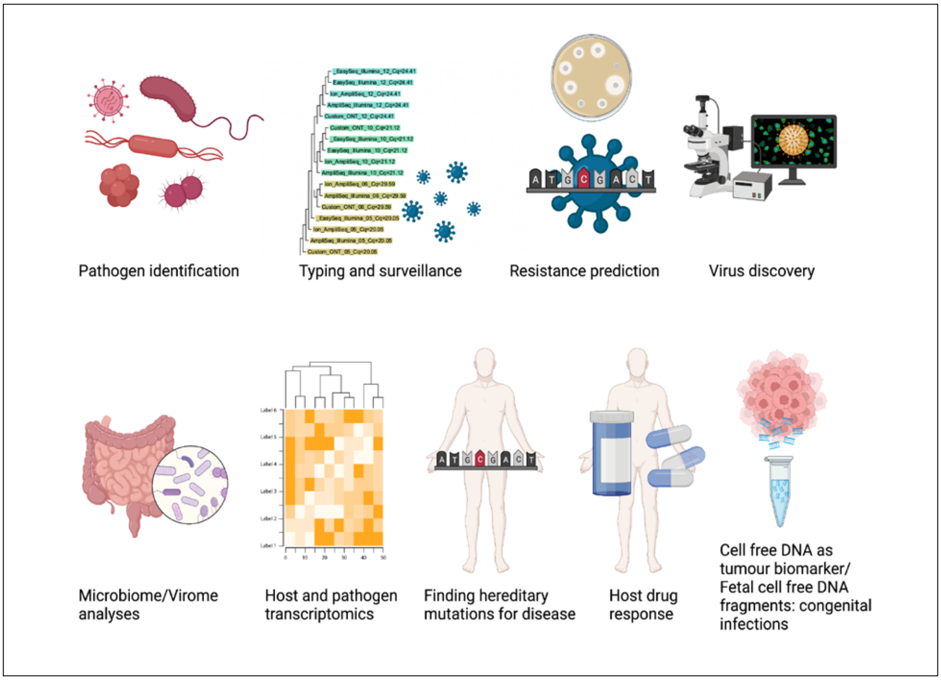

_EasySeq_Illumina_12_Cq=24.41
EasySeq_Illumina_12_Cq=24.41
Ion_AmpliSeq_12_Cq=24.41
AmpliSeq_Illumina_12_Cq=24.41
Custom_ONT_12_Cq=24.41
Custom_ONT_10_Cq=21.12
_EasySeq_Illumina_10_Cq=21.12
EasySeq_Illumina_10_Cq=21.12
Ion_AmpliSeq_10_Cq=21.12
AmpliSeq_Illumina_10_Cq=21.12
Ion_AmpliSeq_06_Cq=29.59
AmpliSeq_Illumina_06_Cq=29.59
Custom_ONT_06_Cq=29.59
_EasySeq_Illumina_05_Cq=20.05
Ion_AmpliSeq_05_Cq=20.05
AmpliSeq_Illumina_05_Cq=20.05
Custom_ONT_05_Cq=20.05
A T G C G A C T
Pathogen identification
Typing and surveillance
Resistance prediction
Virus discovery
Label 6
Label 5
Label 4
Label 3
Label 2
Label 1
0 10 20 30 40 50
A T G C G A C T
Microbiome/Virome analyses
Host and pathogen transcriptomics
Finding hereditary mutations for disease
Host drug response
Cell free DNA as tumour biomarker/ Fetal cell free DNA fragments: congenital infections

Figure 4. One sequence combination test for all departments: metagenomic sequencing as a potential combined clinical application.

Sequencing the complete metagenome, all the genetic material present in a sample, enables identification of pathogens and in addition provides detailed information used for the typing, surveillance and identification of resistance mutations. This metagenomic test can be used for virus discovery and microbiome/ virome analyses. The host-pathogen interaction can be interpreted from transcriptome data, providing information about what genes are activated or repressed. In collaboration with other health care departments, hereditary mutations can be identified in a combined agnostic test for both pathogens and pathogenic variants in the host genome as the next step in molecular diagnostics. Drug metabolizing enzyme information can be retrieved, and the cell-free DNA (cfDNA) can be used as a biomarker for tumour detection, and in addition for the identification of congenital infections. Created using Biorender.com.

Metagenomic sequencing aiding tumour detection

Cell-free DNA (cfDNA) sequenced in the metagenomic sequencing process is useful to detect congenital infection in the fetal cfDNA, especially CMV [127] and parvovirus [104,128]. Sequencing of cfDNA, also considered a liquid biopsy, is additionally a potential biomarker for detecting cancerous tumours within patients for the pathology field. Patients with fast dividing tumour cells have cfDNA in large proportions in serum or plasma as a result of cellular necrosis and apoptosis [129-132]. The detection of cancerous small DNA particles is based on finding specific mutations in the circulating tumour DNA that are not present in the DNA of white blood cells [133-137]. The viral target enrichment panel used in the research in this thesis (**chapter 3 and 4**) can be further enriched using probes of the Cancer Personalized Profiling (CAPP-Seq) sequencing kit, a method proven to be successful in finding and identifying the circulating tumour DNA particles in the cfDNA liquid biopsies [129,138].

Time to update Koch's postulates

With a growing number of viruses detected by means of viral metagenomics and other sequencing techniques, modern thoughts about causality have to be explored as viruses can be found that are not always known to be causal for disease. Around 1880, Robert Koch postulated criteria to establish a causal relationship between a microbe and a disease [139,140]. Over the years these criteria have been updated to four criteria (Figure 5) [141-144], as inoculating an organism with the potentially pathogenic microorganism was not included in the original postulates [139,140]. However, some of Koch's postulates are hard to bring into research practice. Some pathogens cannot be grown in pure culture and Koch's criteria to have "no abundance of the disease-causing organism in healthy individuals" is difficult to prove as per identified virus it is difficult to test many healthy patients efficiently, shortly after the moment a novel or unexpected virus is detected using viral metagenomics. Furthermore, it will be difficult to receive medical ethical approval to follow up in humans on criteria 3 by introducing the cultured disease-causing microorganism into a healthy individual. Additionally, multi-factorial causes, like host health circumstances and dose of infection, play an additional role, and make it harder to rule out any confounding factors.

To investigate microorganism prevalence in the sequence era, it is required to create virome databases that are made publicly available and that can be filtered on abundance of microorganism for clinical syndrome and sample types. In the data that is currently publicly available in sequencing databases, the virome information can also be retrieved [102-104]. In human genetics, mutations are checked for abundance in databases such as GNOMAD [145,146], 1000 genomes [147] and GoNL [148], as genetic disease can never be present in large percentages of healthy individuals. The microbiology community needs similar databases with both prevalence and disease information in order to enable a better differentiation between a healthy and unhealthy microorganism population, and it is recommended that the microbiology community get these in place to be ready for the future.

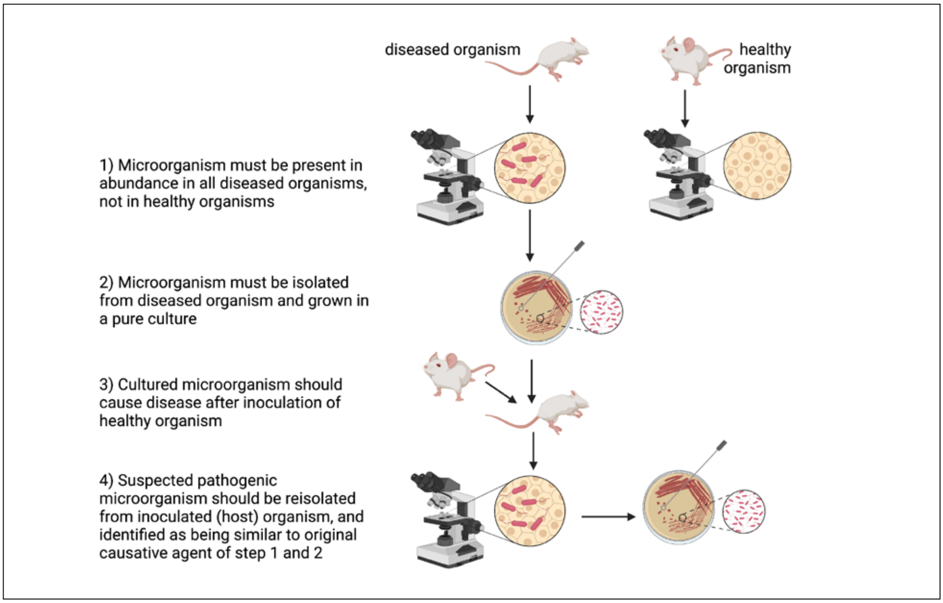

diseased organism
healthy organism
1) Microorganism must be present in abundance in all diseased organisms, not in healthy organisms
2) Microorganism must be isolated from diseased organism and grown in a pure culture
3) Cultured microorganism should cause disease after inoculation of healthy organism
4) Suspected pathogenic microorganism should be reisolated from inoculated (host) organism, and identified as being similar to original causative agent of step 1 and 2

Figure 5. Koch's postulates from the 19th century.

The expanded criteria of the outdated Koch's postulates. The four criteria were designed to establish a causal relationship between a microorganism and a disease. In diseased organisms the suspected microorganism must be found in abundance, and not in healthy organisms (1), and the suspected microorganism must be isolated and grown in pure culture (2). The cultured microorganism, when inoculated into a healthy organism, should cause disease (3) and after infection and disease the suspected pathogenic microorganism had to be reisolated and identified as the original causative agent (4). Created using Biorender.com.

When analysing NGS data from metagenomic sequencing, host-pathogen interactions and virus activity within a host can be investigated to find proof of causality and virulence directly from the available metagenomic sequence data. Some RNA library protocols can differentiate between plus and minus strands, thus providing information about viral activity [149]. Virus transcription is known to differ between stages of infection [150-152] and with this information virus activity can be taken into account in the diagnosis of a patient. Additionally, with transcriptomic sequence read analyses the differential and co-expression of the host immune transcriptome can be mapped [153]. In this way, transcriptomics will give information on viral gene activity and a host response [154-156]. Identifying pathogens using viral metagenomics, retrieving information on prevalence in sequence databases and investigating virulence will lead to novel sequence proof of Koch's postulates (Figure 6). The evaluation of the activity of a virus and corresponding host response will result in an evolution of viral molecular diagnostics. Most likely simply stating if a virus is present or absent and in what quantity will in the future be outdated by novel findings, when precise viral and host immunity activity can be predicted in an infection site based on metagenomic sequencing.

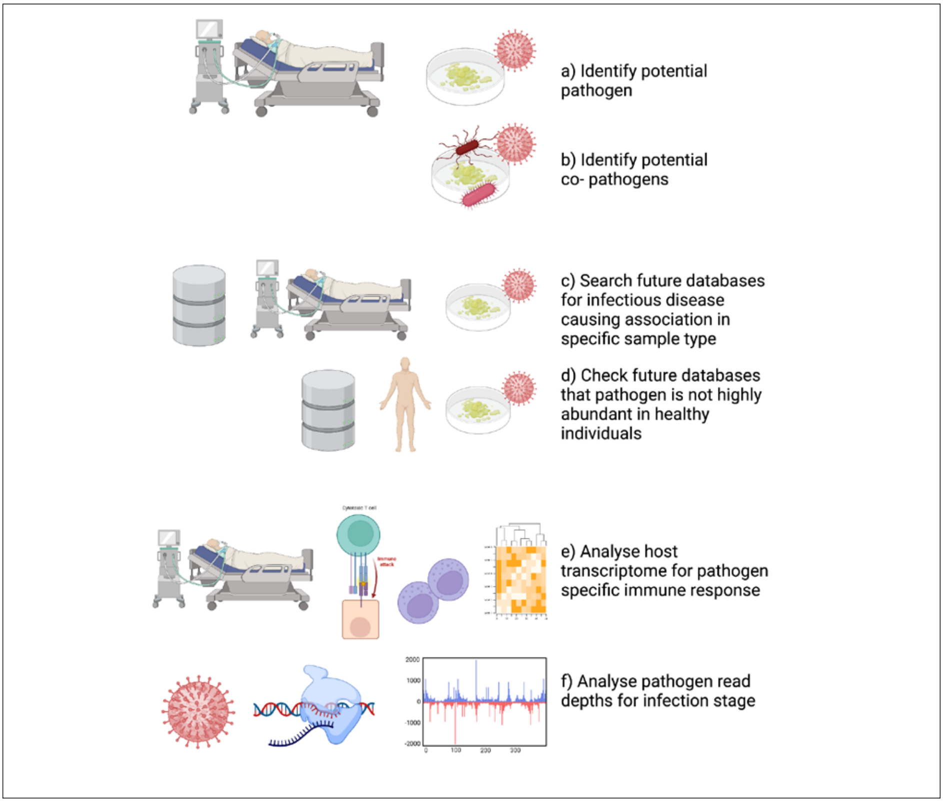

a) Identify potential pathogen
b) Identify potential co- pathogens
c) Search future databases for infectious disease causing association in specific sample type
d) Check future databases that pathogen is not highly abundant in healthy individuals
Cytotoxic T cell
Immune attack
e) Analyse host transcriptome for pathogen specific immune response
2000
1000
0
-1000
-2000
0
100
200
300
f) Analyse pathogen read depths for infection stage

Figure 6. The updated Koch's postulates for the 'next-generation sequencing era'.

When applying the updated Koch's postulates directly on metagenomic NGS data to establish a causal relationship between a microbe and a disease, the first step is to identify a pathogen (a) and to rule out any other potential pathogen within a sample (b). After these steps, future databases should be checked to see if this pathogen is linked to disease-causing symptoms (c), and whether the pathogen is present in the same sample type in healthy individuals (d). In the future, the host-pathogen interaction can be followed by looking at the immune response in a host by analysing the host transcriptome (e) and the pathogen activity and infection stage can be tracked (f). Created using Biorender.com.

Concluding remarks

In the near future of diagnosing infectious diseases, metagenomic sequencing will be implemented at larger scale as a secondary test to evaluate all organisms present at once, by sequencing all available genetic material in a sample. This is the pathogen-agnostic way of identifying a virus that might be pathogenic. Though currently an expensive and time-consuming test, metagenomic sequencing will ultimately improve the diagnostic yield and potentially lead to lower costs when other diagnostic and treatment areas are included in the consideration of costs. Sensitivity of mNGS can be increased by the use of capture probes and more optimal taxonomic classifications tools. With the information available on resistance mutations and typing, metagenomic sequencing provides useful data for virus surveillance. Viruses can be discovered directly from patient samples as exemplified in the beginning of the SARS-CoV-2 pandemic. In the future, sequencing protocols are expected to be faster and more applicable, and with improved filtering of genetic host sequences and addressing causation problems, disease-causing viruses can be differentiated from the regular virome.

A combined metagenomic sequencing test can be used to detect infecting organisms, but can be additionally useful for looking at pathogen activity, pathogen resistance, host transcriptome activity, host pharmacogenetics, genetic inherited pathogenic defaults of a host, and tumour surveillance. By combining all these data from metagenomic sequencing, this test has the promise to function as a multidimensional future diagnostic test aiding multiple clinical disciplines.

References

[1] G. Lasso et al., ‘A Structure-Informed Atlas of Human-Virus Interactions’, Cell, vol. 178, no. 6, pp. 1526-1541.e16, Sep. 2019, doi: 10.1016/j.cell.2019.08.005.

[2] L. H. Taylor, S. M. Latham, and M. E. J. woolhouse, ‘Risk factors for human disease emergence’, Phil. Trans. R. Soc. Lond. B, vol. 356, no. 1411, pp. 983–989, Jul. 2001, doi: 10.1098/rstb.2001.0888.

[3] J. Granerod and N. S. Crowcroft, ‘The epidemiology of acute encephalitis’, Neuropsychological Rehabilitation, vol. 17, no. 4–5, pp. 406–428, Aug. 2007, doi: 10.1080/09602010600989620.

[4] P. Kennedy, P.-L. Quan, and W. Lipkin, ‘Viral Encephalitis of Unknown Cause: Current Perspective and Recent Advances’, Viruses, vol. 9, no. 6, p. 138, Jun. 2017, doi: 10.3390/v9060138.

[5] S. Jain et al., ‘Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults’, N Engl J Med, vol. 373, no. 5, pp. 415–427, Jul. 2015, doi: 10.1056/NEJMoa1500245.

[6] T. Heikkinen and A. Järvinen, ‘The common cold’, The Lancet, vol. 361, no. 9351, pp. 51–59, Jan. 2003, doi: 10.1016/S0140-6736(03)12162-9.

[7] M. Ieven et al., ‘Aetiology of lower respiratory tract infection in adults in primary care: a prospective study in 11 European countries’, Clinical Microbiology and Infection, vol. 24, no. 11, pp. 1158–1163, Nov. 2018, doi: 10.1016/j.cmi.2018.02.004.

[8] E. C. Carbo et al., ‘Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics’, Journal of Clinical Virology, vol. 130, p. 104566, Sep. 2020, doi: 10.1016/j.jcv.2020.104566.

[9] J. C. Haston et al., ‘Prospective Cohort Study of Next-Generation Sequencing as a Diagnostic Modality for Unexplained Encephalitis in Children’, Journal of the Pediatric Infectious Diseases Society, vol. 9, no. 3, pp. 326–333, Jul. 2020, doi: 10.1093/jpids/piz032.

[10] M. R. Wilson et al., ‘Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis’, N Engl J Med, vol. 380, no. 24, pp. 2327–2340, Jun. 2019, doi: 10.1056/NEJMoa1803396.

[11] Kufner et al., ‘Two Years of Viral Metagenomics in a Tertiary Diagnostics Unit: Evaluation of the First 105 Cases’, Genes, vol. 10, no. 9, p. 661, Aug. 2019, doi: 10.3390/genes10090661.

[12] S. L. Salzberg et al., ‘Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system’, Neurol Neuroimmunol Neuroinflamm, vol. 3, no. 4, p. e251, Aug. 2016, doi: 10.1212/NXI.0000000000000251.

[13] H. E. Ambrose et al., ‘Diagnostic Strategy Used To Establish Etiologies of Encephalitis in a Prospective Cohort of Patients in England’, Journal of Clinical Microbiology, vol. 49, no. 10, pp. 3576–3583, Oct. 2011, doi: 10.1128/JCM.00862-11.

[14] S. Saha et al., ‘Unbiased Metagenomic Sequencing for Pediatric Meningitis in Bangladesh Reveals Neuroinvasive Chikungunya Virus Outbreak and Other Unrealized Pathogens’, mBio, vol. 10, no. 6, pp. e02877-19, /mbio/10/6/mBio.02877-19.atom, Dec. 2019, doi: 10.1128/mBio.02877-19.

[15] P. Turner et al., 'The aetiologies of central nervous system infections in hospitalised Cambodian children', BMC Infect Dis, vol. 17, no. 1, p. 806, Dec. 2017, doi: 10.1186/s12879-017-2915-6.

[16] J. Kawada et al., 'Next-Generation Sequencing for the Identification of Viruses in Pediatric Acute Encephalitis and Encephalopathy', Open Forum Infectious Diseases, vol. 3, no. suppl_1, p. 1172, Dec. 2016, doi: 10.1093/ofid/ofw172.875.

[17] S. L. Smits et al., 'Novel Cyclovirus in Human Cerebrospinal Fluid, Malawi, 2010–2011', Emerg. Infect. Dis., vol. 19, no. 9, Sep. 2013, doi: 10.3201/eid1909.130404.

[18] H. Jerome et al., 'Metagenomic next-generation sequencing aids the diagnosis of viral infections in febrile returning travellers', Journal of Infection, vol. 79, no. 4, pp. 383–388, Oct. 2019, doi: 10.1016/j.jinf.2019.08.003.

[19] K. N. Govender, T. L. Street, N. D. Sanderson, and D. W. Eyre, 'Metagenomic Sequencing as a Pathogen-Agnostic Clinical Diagnostic Tool for Infectious Diseases: a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies', J Clin Microbiol, vol. 59, no. 9, pp. e02916-20, Aug. 2021, doi: 10.1128/JCM.02916-20.

[20] Z. Diao, D. Han, R. Zhang, and J. Li, 'Metagenomics next-generation sequencing tests take the stage in the diagnosis of lower respiratory tract infections', Journal of Advanced Research, vol. 38, pp. 201–212, May 2022, doi: 10.1016/j.jare.2021.09.012.

[21] P. Parize et al., 'Untargeted next-generation sequencing-based first-line diagnosis of infection in immunocompromised adults: a multicentre, blinded, prospective study', Clinical Microbiology and Infection, vol. 23, no. 8, p. 574.e1-574.e6, Aug. 2017, doi: 10.1016/j.cmi.2017.02.006.

[22] N. T. T. Hong et al., 'Performance of Metagenomic Next-Generation Sequencing for the Diagnosis of Viral Meningoencephalitis in a Resource-Limited Setting', Open Forum Infectious Diseases, vol. 7, no. 3, p. ofaa046, Mar. 2020, doi: 10.1093/ofid/ofaa046.

[23] S. Miller et al., 'Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid', Genome Res., vol. 29, no. 5, pp. 831–842, 2019, doi: 10.1101/gr.238170.118.

[24] T. A. Blauwkamp et al., 'Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease', Nat Microbiol, vol. 4, no. 4, pp. 663–674, Apr. 2019, doi: 10.1038/s41564-018-0349-6.

[25] S. Somasekar et al., 'Viral Surveillance in Serum Samples From Patients With Acute Liver Failure By Metagenomic Next-Generation Sequencing', Clinical Infectious Diseases, vol. 65, no. 9, pp. 1477–1485, Oct. 2017, doi: 10.1093/cid/cix596.

[26] J. Rossoff et al., 'Noninvasive Diagnosis of Infection Using Plasma Next-Generation Sequencing: A Single-Center Experience', Open Forum Infectious Diseases, vol. 6, no. 8, p. ofz327, Aug. 2019, doi: 10.1093/ofid/ofz327.

[27] R. Schlaberg et al., 'Viral Pathogen Detection by Metagenomics and Pan-Viral Group Polymerase Chain Reaction in Children With Pneumonia Lacking Identifiable Etiology', The Journal of Infectious Diseases, vol. 215, no. 9, pp. 1407–1415, May 2017, doi: 10.1093/infdis/jix148.

[28] T. Doan et al., 'Metagenomic DNA Sequencing for the Diagnosis of Intraocular Infections', Ophthalmology, vol. 124, no. 8, pp. 1247–1248, Aug. 2017, doi: 10.1016/j.ophtha.2017.03.045.

[29] C. Langelier et al., 'Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults', Proc. Natl. Acad. Sci. U.S.A., vol. 115, no. 52, Dec. 2018, doi: 10.1073/pnas.1809700115.

[30] J. Wang, Y. Han, and J. Feng, 'Metagenomic next-generation sequencing for mixed pulmonary infection diagnosis', BMC Pulm Med, vol. 19, no. 1, p. 252, Dec. 2019, doi: 10.1186/s12890-019-1022-4.

[31] A. L. van Rijn et al., 'The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease', PLoS ONE, vol. 14, no. 10, p. e0223952, Oct. 2019, doi: 10.1371/journal.pone.0223952.

[32] J. Huang et al., 'Metagenomic Next-Generation Sequencing versus Traditional Pathogen Detection in the Diagnosis of Peripheral Pulmonary Infectious Lesions', IDR, vol. Volume 13, pp. 567–576, Feb. 2020, doi: 10.2147/IDR.S235182.

[33] S. van Boheemen et al., 'Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients', The Journal of Molecular Diagnostics, vol. 22, no. 2, pp. 196–207, Feb. 2020, doi: 10.1016/j.jmoldx.2019.10.007.

[34] J. J. C. de Vries et al., 'Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting', Journal of Clinical Virology, vol. 138, p. 104812, May 2021, doi: 10.1016/j.jcv.2021.104812.

[35] S. Nooij, D. Schmitz, H. Vennema, A. Kroneman, and M. P. G. Koopmans, 'Overview of Virus Metagenomic Classification Methods and Their Biological Applications', Front. Microbiol., vol. 9, p. 749, Apr. 2018, doi: 10.3389/fmicb.2018.00749.

[36] Junier et al., 'Viral Metagenomics in the Clinical Realm: Lessons Learned from a Swiss-Wide Ring Trial', Genes, vol. 10, no. 9, p. 655, Aug. 2019, doi: 10.3390/genes10090655.

[37] D. E. Wood and S. L. Salzberg, 'Kraken: ultrafast metagenomic sequence classification using exact alignments', Genome Biol, vol. 15, no. 3, p. R46, 2014, doi: 10.1186/gb-2014-15-3-r46.

[38] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, 'CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers', BMC Genomics, vol. 16, no. 1, p. 236, Dec. 2015, doi: 10.1186/s12864-015-1419-2.

[39] S. H. Ye, K. J. Siddle, D. J. Park, and P. C. Sabeti, 'Benchmarking Metagenomics Tools for Taxonomic Classification', Cell, vol. 178, no. 4, pp. 779–794, Aug. 2019, doi: 10.1016/j.cell.2019.07.010.

[40] P. Menzel, K. L. Ng, and A. Krogh, 'Fast and sensitive taxonomic classification for metagenomics with Kaiju', Nat Commun, vol. 7, no. 1, p. 11257, Sep. 2016, doi: 10.1038/ncomms11257.

[41] K. Mavromatis et al., 'Use of simulated data sets to evaluate the fidelity of metagenomic processing methods', Nat Methods, vol. 4, no. 6, pp. 495–500, Jun. 2007, doi: 10.1038/nmeth1043.

[42] F. Meyer, A. Bremges, P. Belmann, S. Janssen, A. C. McHardy, and D. Koslicki, 'Assessing taxonomic metagenome profilers with OPAL', Genome Biol, vol. 20, no. 1, p. 51, Dec. 2019, doi: 10.1186/s13059-019-1646-y.

[43] A. Sczyrba et al., 'Critical Assessment of Metagenome Interpretation — a benchmark of metagenomics software', Nat Methods, vol. 14, no. 11, pp. 1063–1071, Nov. 2017, doi: 10.1038/nmeth.4458.

[44] A. B. R. McIntyre et al., ‘Comprehensive benchmarking and ensemble approaches for metagenomic classifiers’, Genome Biol, vol. 18, no. 1, p. 182, Dec. 2017, doi: 10.1186/s13059-017-1299-7.

[45] Z. Sun et al., ‘Challenges in benchmarking metagenomic profilers’, Nat Methods, vol. 18, no. 6, pp. 618–626, Jun. 2021, doi: 10.1038/s41592-021-01141-3.

[46] A. Escobar-Zepeda et al., ‘Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics’, Sci Rep, vol. 8, no. 1, p. 12034, Dec. 2018, doi: 10.1038/s41598-018-30515-5.

[47] A. Brinkmann et al., ‘Proficiency Testing of Virus Diagnostics Based on Bioinformatics Analysis of Simulated In Silico High-Throughput Sequencing Data Sets’, J Clin Microbiol, vol. 57, no. 8, pp. e00466-19, /jcm/57/8/JCM.00466-19.atom, Jun. 2019, doi: 10.1128/JCM.00466-19.

[48] N. Couto et al., ‘Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens’, Sci Rep, vol. 8, no. 1, p. 13767, Dec. 2018, doi: 10.1038/s41598-018-31873-w.

[49] J. J. C. de Vries et al., ‘Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples’, Journal of Clinical Virology, p. 104908, Jul. 2021, doi: 10.1016/j.jcv.2021.104908.

[50] S. Morfopoulou and V. Plagnol, ‘Bayesian mixture analysis for metagenomic community profiling’, Bioinformatics, vol. 31, no. 18, pp. 2930–2938, Sep. 2015, doi: 10.1093/bioinformatics/btv317.

[51] J. M. Martí, ‘Recentrifuge: Robust comparative analysis and contamination removal for metagenomics’, PLoS Comput Biol, vol. 15, no. 4, p. e1006967, Apr. 2019, doi: 10.1371/journal.pcbi.1006967.

[52] V. C. Piro and B. Y. Renard, ‘Contamination detection and microbiome exploration with GRIMER’, Bioinformatics, preprint, Jun. 2021. doi: 10.1101/2021.06.22.449360.

[53] D. Shah, J. R. Brown, J. C. D. Lee, M. L. Carpenter, G. Wall, and J. Breuer, ‘Use of a sample-to-result shotgun metagenomics platform for the detection and quantification of viral pathogens in paediatric immunocompromised patients’, Journal of Clinical Virology Plus, vol. 2, no. 2, p. 100073, Jun. 2022, doi: 10.1016/j.jcvp.2022.100073.

[54] C. Huang et al., ‘Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China’, The Lancet, vol. 395, no. 10223, pp. 497–506, Feb. 2020, doi: 10.1016/S0140-6736(20)30183-5.

[55] P. Zhou et al., ‘Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin’, Microbiology, preprint, Jan. 2020. doi: 10.1101/2020.01.22.914952.

[56] N. Zhu et al., ‘A Novel Coronavirus from Patients with Pneumonia in China, 2019’, N Engl J Med, vol. 382, no. 8, pp. 727–733, Feb. 2020, doi: 10.1056/NEJMoa2001017.

[57] M. Vilsker et al., ‘Genome Detective: an automated system for virus identification from high-throughput sequencing data’, Bioinformatics, vol. 35, no. 5, pp. 871–873, Mar. 2019, doi: 10.1093/bioinformatics/bty695.

[58] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, ‘Basic local alignment search tool’, Journal of Molecular Biology, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[59] W. T. Harvey et al., ‘SARS-CoV-2 variants, spike mutations and immune escape’, Nat Rev Microbiol, vol. 19, no. 7, pp. 409–424, Jul. 2021, doi: 10.1038/s41579-021-00573-0.

[60] K. Tao et al., ‘The biological and clinical significance of emerging SARS-CoV-2 variants’, Nat Rev Genet, vol. 22, no. 12, pp. 757–773, Dec. 2021, doi: 10.1038/s41576-021-00408-x.

[61] Z. Chen et al., ‘Global landscape of SARS-CoV-2 genomic surveillance and data sharing’, Nat Genet, vol. 54, no. 4, pp. 499–507, Apr. 2022, doi: 10.1038/s41588-022-01033-y.

[62] ‘https://www.gisaid.org/’, Apr. 2022.

[63] D. Liu et al., ‘Development and Multicenter Assessment of a Reference Panel for Clinical Shotgun Metagenomics for Pathogen Detection’, In Review, preprint, Feb. 2021. doi: 10.21203/rs.3.rs-208796/v1.

[64] J. A. Nasir et al., ‘A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture’, Viruses, vol. 12, no. 8, p. 895, Aug. 2020, doi: 10.3390/v12080895.

[65] M. Xiao et al., ‘Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples’, Genome Med, vol. 12, no. 1, p. 57, Dec. 2020, doi: 10.1186/s13073-020-00751-4.

[66] F. Wegner et al., ‘External Quality Assessment of SARS-CoV-2 Sequencing: an ESGMD-SSM Pilot Trial across 15 European Laboratories’, J Clin Microbiol, vol. 60, no. 1, pp. e01698-21, Jan. 2022, doi: 10.1128/JCM.01698-21.

[67] J. Plitnick, S. Griesemer, E. Lasek-Nesselquist, N. Singh, D. M. Lamson, and K. St. George, ‘Whole-Genome Sequencing of SARS-CoV-2: Assessment of the Ion Torrent AmpliSeq Panel and Comparison with the Illumina MiSeq ARTIC Protocol’, J Clin Microbiol, vol. 59, no. 12, pp. e00649-21, Nov. 2021, doi: 10.1128/JCM.00649-21.

[68] J. H. Chai et al., ‘Cost-benefit analysis of introducing next-generation sequencing (metagenomic) pathogen testing in the setting of pyrexia of unknown origin’, PLoS ONE, vol. 13, no. 4, p. e0194648, Apr. 2018, doi: 10.1371/journal.pone.0194648.

[69] F. X. López-Labrador et al., ‘Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure’, Journal of Clinical Virology, vol. 134, p. 104691, Jan. 2021, doi: 10.1016/j.jcv.2020.104691.

[70] M. Costello et al., ‘Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms’, BMC Genomics, vol. 19, no. 1, p. 332, Dec. 2018, doi: 10.1186/s12864-018-4703-0.

[71] N. Stoler and A. Nekrutenko, ‘Sequencing error profiles of Illumina sequencing instruments’, NAR Genomics and Bioinformatics, vol. 3, no. 1, p. lqab019, Jan. 2021, doi: 10.1093/nargab/lqab019.

[72] Z. Feng, J. C. Clemente, B. Wong, and E. E. Schadt, ‘Detecting and phasing minor single-nucleotide variants from long-read sequencing data’, Nat Commun, vol. 12, no. 1, p. 3032, Dec. 2021, doi: 10.1038/s41467-021-23289-4.

[73] R. Kou et al., ‘Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations’, PLoS ONE, vol. 11, no. 1, p. e0146638, Jan. 2016, doi: 10.1371/journal.pone.0146638.

[74] J. B. Hiatt, C. C. Pritchard, S. J. Salipante, B. J. O’Roak, and J. Shendure, ‘Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation’, Genome Res., vol. 23, no. 5, pp. 843–854, May 2013, doi: 10.1101/gr.147686.112.

[75] B. Kant et al., ‘Gene Mosaicism Screening Using Single-Molecule Molecular Inversion Probes in Routine Diagnostics for Systemic Autoinflammatory Diseases’, The Journal of Molecular Diagnostics, vol. 21, no. 6, pp. 943–950, Nov. 2019, doi: 10.1016/j.jmoldx.2019.06.009.

[76] K. Kryukov and T. Imanishi, ‘Human Contamination in Public Genome Assemblies’, PLoS ONE, vol. 11, no. 9, p. e0162424, Sep. 2016, doi: 10.1371/journal.pone.0162424.

[77] M. S. Longo, M. J. O’Neill, and R. J. O’Neill, ‘Abundant Human DNA Contamination Identified in Non-Primate Genome Databases’, PLoS ONE, vol. 6, no. 2, p. e16410, Feb. 2011, doi: 10.1371/journal.pone.0016410.

[78] S. Mukherjee, M. Huntemann, N. Ivanova, N. C. Kyrpides, and A. Pati, ‘Large-scale contamination of microbial isolate genomes by Illumina PhiX control’, Stand in Genomic Sci, vol. 10, no. 1, p. 18, Dec. 2015, doi: 10.1186/1944-3277-10-18.

[79] M. Laurence, C. Hatzis, and D. E. Brash, ‘Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes’, PLoS ONE, vol. 9, no. 5, p. e97876, May 2014, doi: 10.1371/journal.pone.0097876.

[80] A. Rhie et al., ‘Towards complete and error-free genome assemblies of all vertebrate species’, Nature, vol. 592, no. 7856, pp. 737–746, Apr. 2021, doi: 10.1038/s41586-021-03451-0.

[81] M. Steinegger and S. L. Salzberg, ‘Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank’, Genome Biol, vol. 21, no. 1, p. 115, Dec. 2020, doi: 10.1186/s13059-020-02023-1.

[82] S. Roux, S. J. Hallam, T. Woyke, and M. B. Sullivan, ‘Viral dark matter and virus–host interactions resolved from publicly available microbial genomes’, eLife, vol. 4, p. e08490, Jul. 2015, doi: 10.7554/eLife.08490.

[83] S. R. Krishnamurthy and D. Wang, ‘Origins and challenges of viral dark matter’, Virus Research, vol. 239, pp. 136–142, Jul. 2017, doi: 10.1016/j.virusres.2017.02.002.

[84] T. G. Burland, ‘DNASTAR’s Lasergene Sequence Analysis Software’, in Bioinformatics Methods and Protocols, vol. 132, New Jersey: Humana Press, 1999, pp. 71–91. doi: 10.1385/1-59259-192-2:71.

[85] S. S. Minot, N. Krumm, and N. B. Greenfield, ‘One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification’, Bioinformatics, preprint, Sep. 2015. doi: 10.1101/027607.

[86] S. Flygare et al., ‘Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling’, Genome Biol, vol. 17, no. 1, p. 111, Dec. 2016, doi: 10.1186/s13059-016-0969-1.

[87] R. Lozano et al., ‘Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010’, The Lancet, vol. 380, no. 9859, pp. 2095–2128, Dec. 2012, doi: 10.1016/S0140-6736(12)61728-0.

[88] H. Nair et al., ‘Global and regional burden of hospital admissions for severe acute lower respiratory infections in young children in 2010: a systematic analysis’, The Lancet, vol. 381, no. 9875, pp. 1380–1390, Apr. 2013, doi: 10.1016/S0140-6736(12)61901-1.

[89] M. Bates, V. Mudenda, P. Mwaba, and A. Zumla, ‘Deaths due to respiratory tract infections in Africa: a review of autopsy studies’, Current Opinion in Pulmonary Medicine, vol. 19, no. 3, pp. 229–237, May 2013, doi: 10.1097/MCP.0b013e32835f4fe4.

[90] 'NVMM', Accessed: Aug. 03, 2022. [Online]. Available: www.nvmm.nl/media/4618/220422_diagnostisch-algoritme_nvmm-nwkv-nvk.pdf

[91] G. Almogy et al., 'Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform', Genomics, preprint, May 2022. doi: 10.1101/2022.05.29.493900.

[92] R. M. Leggett et al., 'Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens', Nat Microbiol, vol. 5, no. 3, pp. 430–442, Mar. 2020, doi: 10.1038/s41564-019-0626-z.

[93] M. D. Cao, D. Ganesamoorthy, A. G. Elliott, H. Zhang, M. A. Cooper, and L. J. M. Coin, 'Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinIONTM sequencing', GigaSci, vol. 5, no. 1, p. 32, Dec. 2016, doi: 10.1186/s13742-016-0137-2.

[94] C. P. Oechslin et al., 'Limited Correlation of Shotgun Metagenomics Following Host Depletion and Routine Diagnostics for Viruses and Bacteria in Low Concentrated Surrogate and Clinical Samples', Front. Cell. Infect. Microbiol., vol. 8, p. 375, Oct. 2018, doi: 10.3389/fcimb.2018.00375.

[95] M. T. Nelson et al., 'Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles', Cell Reports, vol. 26, no. 8, pp. 2227-2240.e5, Feb. 2019, doi: 10.1016/j.celrep.2019.01.091.

[96] 'GIAB', Accessed: Aug. 03, 2022. [Online]. Available: https://www.nist.gov/programs-projects/genome-bottle

[97] J. M. Zook et al., 'Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls', Nat Biotechnol, vol. 32, no. 3, pp. 246–251, Mar. 2014, doi: 10.1038/nbt.2835.

[98] J. M. Zook et al., 'An open resource for accurately benchmarking small variant and reference calls', Nat Biotechnol, vol. 37, no. 5, pp. 561–566, May 2019, doi: 10.1038/s41587-019-0074-6.

[99] 'GoFair', Accessed: Aug. 03, 2022. [Online]. Available: https://www.go-fair.org/fair-principles/

[100] E. Simon-Loriere and E. C. Holmes, 'Why do RNA viruses recombine?', Nat Rev Microbiol, vol. 9, no. 8, pp. 617–626, Aug. 2011, doi: 10.1038/nrmicro2614.

[101] S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan, 'VirSorter: mining viral signal from microbial genomic data', PeerJ, vol. 3, p. e985, May 2015, doi: 10.7717/peerj.985.

[102] R. C. Edgar et al., 'Petabase-scale sequence alignment catalyses viral discovery', Nature, vol. 602, no. 7895, pp. 142–147, Feb. 2022, doi: 10.1038/s41586-021-04332-2.

[103] A. Moustafa et al., 'The blood DNA virome in 8,000 humans', PLoS Pathog, vol. 13, no. 3, p. e1006292, Mar. 2017, doi: 10.1371/journal.ppat.1006292.

[104] J. Linthorst, M. M. M. Baksi, M. R. A. Welkers, and E. A. Sistermans, 'The cell▯free DNA virome of 108,349 Dutch pregnant women', Prenatal Diagnosis, p. pd.6143, Apr. 2022, doi: 10.1002/pd.6143.

[105] 'TCGA Research Network', Accessed: Sep. 04, 2022. [Online]. Available: https://www.cancer.gov/tcga

[106] D. M. Parkin, 'The global health burden of infection-associated cancers in the year 2002', Int. J. Cancer, vol. 118, no. 12, pp. 3030–3044, Jun. 2006, doi: 10.1002/ijc.21731.

[107] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, 'Global burden of cancers attributable to infections in 2012: a synthetic analysis', The Lancet Global Health, vol. 4, no. 9, pp. e609–e616, Sep. 2016, doi: 10.1016/S2214-109X(16)30143-7.

[108] L. Wang et al., 'Artificial Intelligence for COVID-19: A Systematic Review', Front. Med., vol. 8, p. 704256, Sep. 2021, doi: 10.3389/fmed.2021.704256.

[109] Q.-V. Pham, D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, 'Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts', IEEE Access, vol. 8, pp. 130820–130839, 2020, doi: 10.1109/ACCESS.2020.3009328.

[110] B. M. C. Silva, J. J. P. C. Rodrigues, I. de la Torre Díez, M. López-Coronado, and K. Saleem, 'Mobile-health: A review of current state in 2015', Journal of Biomedical Informatics, vol. 56, pp. 265–272, Aug. 2015, doi: 10.1016/j.jbi.2015.06.003.

[111] A. S. R. Srinivasa Rao and J. A. Vazquez, 'Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine', Infect Control Hosp Epidemiol, vol. 41, no. 7, pp. 826–830, Jul. 2020, doi: 10.1017/ice.2020.61.

[112] H. S. Maghdid, K. Z. Ghafoor, A. S. Sadiq, K. Curran, D. B. Rawat, and K. Rabie, 'A Novel AI-enabled Framework to Diagnose Coronavirus COVID 19 using Smartphone Embedded Sensors: Design Study'. arXiv, May 30, 2020. Accessed: Aug. 04, 2022. [Online]. Available: http://arxiv.org/abs/2003.07434

[113] E. Ong, M. U. Wong, A. Huffman, and Y. He, 'COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning', Front. Immunol., vol. 11, p. 1581, Jul. 2020, doi: 10.3389/fimmu.2020.01581.

[114] S. F. Ahmed, A. A. Quadeer, and M. R. McKay, 'Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies', Viruses, vol. 12, no. 3, p. E254, Feb. 2020, doi: 10.3390/v12030254.

[115] Y. Furuse, 'Genomic sequencing effort for SARS-CoV-2 by country during the pandemic', International Journal of Infectious Diseases, vol. 103, pp. 305–307, Feb. 2021, doi: 10.1016/j.ijid.2020.12.034.

[116] K. Krebs and L. Milani, 'Translating pharmacogenomics into clinical decisions: do not let the perfect be the enemy of the good', Hum Genomics, vol. 13, no. 1, p. 39, Dec. 2019, doi: 10.1186/s40246-019-0229-z.

[117] M. Pirmohamed, 'Pharmacogenetics and pharmacogenomics: Editorial', British Journal of Clinical Pharmacology, vol. 52, no. 4, pp. 345–347, Oct. 2001, doi: 10.1046/j.0306-5251.2001.01498.x.

[118] C. I. van der Made et al., 'Presence of Genetic Variants Among Young Men With Severe COVID-19', JAMA, vol. 324, no. 7, p. 663, Aug. 2020, doi: 10.1001/jama.2020.13719.

[119] D. Burgner, S. E. Jamieson, and J. M. Blackwell, 'Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better?', The Lancet Infectious Diseases, vol. 6, no. 10, pp. 653–663, Oct. 2006, doi: 10.1016/S1473-3099(06)70601-6.

[120] L. Kachuri et al., 'The landscape of host genetic factors involved in immune response to common viral infections', Genome Med, vol. 12, no. 1, p. 93, Dec. 2020, doi: 10.1186/s13073-020-00790-x.

[121] L. Quintana-Murci, 'Human Immunology through the Lens of Evolutionary Genetics', Cell, vol. 177, no. 1, pp. 184–199, Mar. 2019, doi: 10.1016/j.cell.2019.02.033.

[122] G. Kerner et al., 'Homozygosity for TYK2 P1104A underlies tuberculosis in about 1% of patients in a cohort of European ancestry', Proc. Natl. Acad. Sci. U.S.A., vol. 116, no. 21, pp. 10430–10434, May 2019, doi: 10.1073/pnas.1903561116.

[123] J. Zhao et al., ‘Coexistence of Autoimmune Encephalitis and Other Systemic Autoimmune Diseases’, Front. Neurol., vol. 10, p. 1142, Oct. 2019, doi: 10.3389/fneur.2019.01142.

[124] L. Farnaes et al., ‘Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization’, npj Genomic Med, vol. 3, no. 1, p. 10, Dec. 2018, doi: 10.1038/s41525-018-0049-4.

[125] H. Daoud et al., ‘Next-generation sequencing for diagnosis of rare diseases in the neonatal intensive care unit’, CMAJ, vol. 188, no. 11, pp. E254–E260, Aug. 2016, doi: 10.1503/cmaj.150823.

[126] C. J. Saunders et al., ‘Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units’, Sci. Transl. Med., vol. 4, no. 154, Oct. 2012, doi: 10.1126/scitranslmed.3004041.

[127] V. Chesnais et al., ‘Using massively parallel shotgun sequencing of maternal plasmatic cell-free DNA for cytomegalovirus DNA detection during pregnancy: a proof of concept study’, Sci Rep, vol. 8, no. 1, p. 4321, Dec. 2018, doi: 10.1038/s41598-018-22414-6.

[128] J. Linthorst, M. R. A. Welkers, and E. A. Sistermans, ‘Distinct fragmentation patterns of circulating viral cell-free DNA in 83,552 non-invasive prenatal testing samples’, EVCNA, 2021, doi: 10.20517/evcna.2021.13.

[129] A. M. Newman et al., ‘An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage’, Nat Med, vol. 20, no. 5, pp. 548–554, May 2014, doi: 10.1038/nm.3519.

[130] G. D. Sorenson, D. M. Pribish, F. H. Valone, V. A. Memoli, D. J. Bzik, and S. L. Yao, ‘Soluble normal and mutated DNA sequences from single-copy genes in human blood’, Cancer Epidemiol Biomarkers Prev, vol. 3, no. 1, pp. 67–71, Feb. 1994.

[131] V. Vasioukhin, P. Anker, P. Maurice, J. Lyautey, C. Lederrey, and M. Stroun, ‘Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia’, Br J Haematol, vol. 86, no. 4, pp. 774–779, Apr. 1994, doi: 10.1111/j.1365-2141.1994.tb04828.x.

[132] A. J. Bronkhorst, V. Ungerer, and S. Holdenrieder, ‘The emerging role of cell-free DNA as a molecular marker for cancer management’, Biomolecular Detection and Quantification, vol. 17, p. 100087, Mar. 2019, doi: 10.1016/j.bdq.2019.100087.

[133] J. D. Cohen et al., ‘Detection and localization of surgically resectable cancers with a multi-analyte blood test’, Science, vol. 359, no. 6378, pp. 926–930, Feb. 2018, doi: 10.1126/science.aar3247.

[134] J. Vendrell, F. Mau-Them, B. Béganton, S. Godreuil, P. Coopman, and J. Solassol, ‘Circulating Cell Free Tumor DNA Detection as a Routine Tool forLung Cancer Patient Management’, IJMS, vol. 18, no. 2, p. 264, Jan. 2017, doi: 10.3390/ijms18020264.

[135] E. Kidess and S. S. Jeffrey, ‘Circulating tumor cells versus tumor-derived cell-free DNA: rivals or partners in cancer care in the era of single-cell analysis?’, Genome Med, vol. 5, no. 8, p. 70, Aug. 2013, doi: 10.1186/gm474.

[136] J. Phallen et al., ‘Direct detection of early-stage cancers using circulating tumor DNA’, Sci Transl Med, vol. 9, no. 403, p. eaan2415, Aug. 2017, doi: 10.1126/scitranslmed.aan2415.

[137] S. Cristiano et al., ‘Genome-wide cell-free DNA fragmentation in patients with cancer’, Nature, vol. 570, no. 7761, pp. 385–389, Jun. 2019, doi: 10.1038/s41586-019-1272-6.

[138] F. Scherer et al., 'Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA', Sci. Transl. Med., vol. 8, no. 364, Nov. 2016, doi: 10.1126/scitranslmed.aai8545.

[139] Koch R. The etiology of anthrax, based on the life history of Bacillus anthracis. Beiträge zur Biologie der Pflanzen. 1876;2(2):277-310.

[140] Koch R. Die Atiologie der Tuberculose. Berl. Klin. Wochenschr. 1882;19:221-30.

[141] Loeffler F (1884) Mitt. Aus dem Kaiserl. Gesundheitsamte, 2, 421e499.

[142] T. M. Rivers, 'Viruses and Koch's Postulates', J Bacteriol, vol. 33, no. 1, pp. 1–12, Jan. 1937, doi: 10.1128/jb.33.1.1-12.1937.

[143] Evans AS. Causation and disease: the Henle-Koch postulates revisited. Yale J Biol Med. 1976 May;49(2):175-95.

[144] S. Falkow, 'Molecular Koch's Postulates Applied to Microbial Pathogenicity', Clinical Infectious Diseases, vol. 10, no. Supplement 2, pp. S274–S276, Aug. 1988, doi: 10.1093/cid/10.Supplement_2.S274.

[145] R. L. Collins et al., 'A structural variation reference for medical and population genetics', Nature, vol. 581, no. 7809, pp. 444–451, May 2020, doi: 10.1038/s41586-020-2287-8.

[146] K. J. Karczewski et al., 'The mutational constraint spectrum quantified from variation in 141,456 humans', Nature, vol. 581, no. 7809, pp. 434–443, May 2020, doi: 10.1038/s41586-020-2308-7.

[147] The 1000 Genomes Project Consortium et al., 'A global reference for human genetic variation', Nature, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.

[148] D. I. Boomsma et al., 'The Genome of the Netherlands: design, and project goals', Eur J Hum Genet, vol. 22, no. 2, pp. 221–227, Feb. 2014, doi: 10.1038/ejhg.2013.118.

[149] D. P. Depledge, I. Mohr, and A. C. Wilson, 'Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes', J Virol, vol. 93, no. 1, pp. e01342-18, Jan. 2019, doi: 10.1128/JVI.01342-18.

[150] S. Boersma et al., 'Translation and Replication Dynamics of Single RNA Viruses', Cell, vol. 183, no. 7, pp. 1930-1945.e23, Dec. 2020, doi: 10.1016/j.cell.2020.10.019.

[151] J. Sun et al., 'Comparative Transcriptome Analysis Reveals the Intensive Early Stage Responses of Host Cells to SARS-CoV-2 Infection', Front. Microbiol., vol. 11, p. 593857, Nov. 2020, doi: 10.3389/fmicb.2020.593857.

[152] Z. Yang, D. P. Bruno, C. A. Martens, S. F. Porcella, and B. Moss, 'Simultaneous high-resolution analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing', Proc. Natl. Acad. Sci. U.S.A., vol. 107, no. 25, pp. 11513–11518, Jun. 2010, doi: 10.1073/pnas.1006594107.

[153] P. Khatri, M. Sirota, and A. J. Butte, 'Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges', PLoS Comput Biol, vol. 8, no. 2, p. e1002375, Feb. 2012, doi: 10.1371/journal.pcbi.1002375.

[154] A. J. Westermann, S. A. Gorski, and J. Vogel, 'Dual RNA-seq of pathogen and host', Nat Rev Microbiol, vol. 10, no. 9, pp. 618–630, Sep. 2012, doi: 10.1038/nrmicro2852.

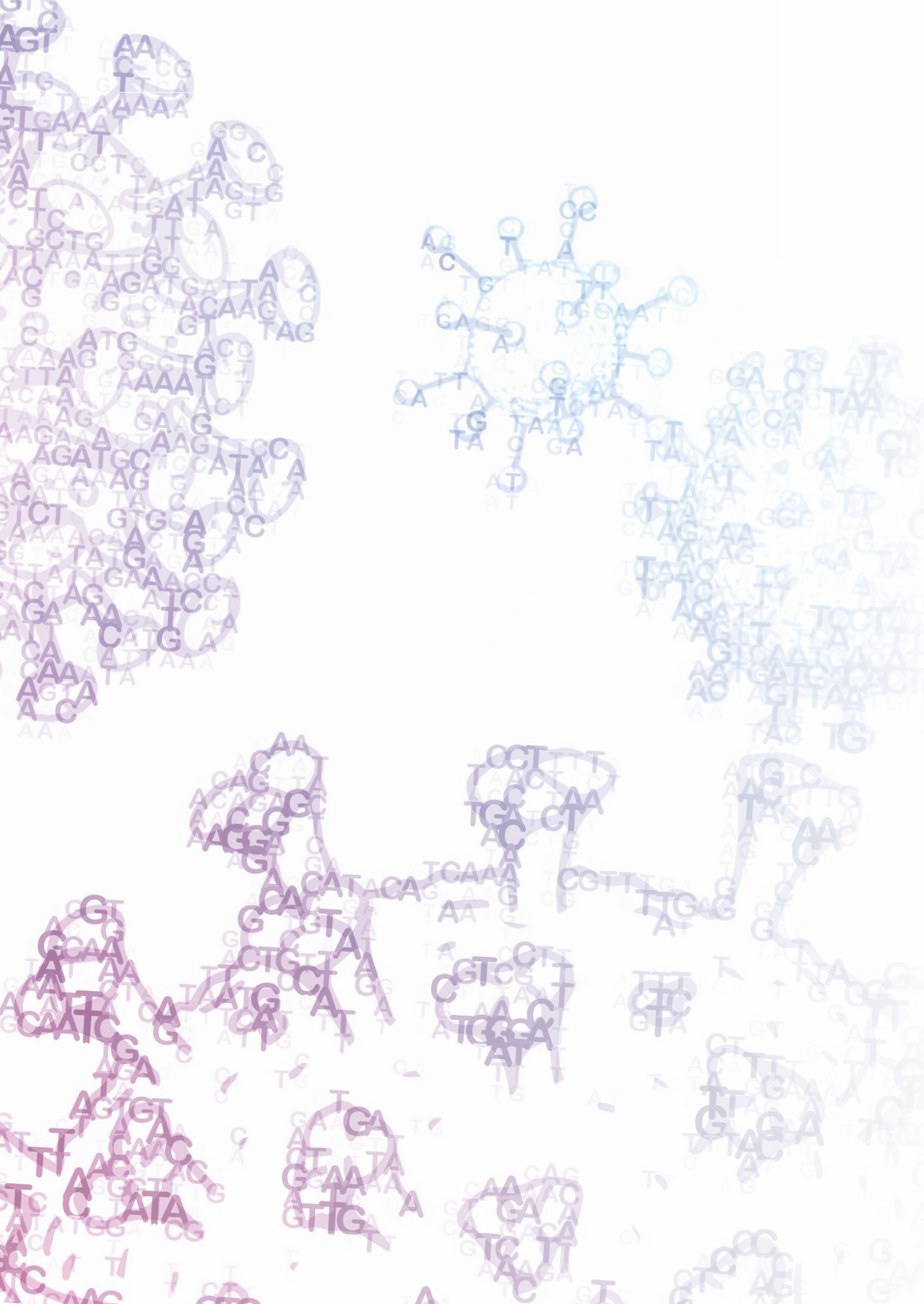
[155] A. J. Westermann et al., 'Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions', Nature, vol. 529, no. 7587, pp. 496–501, Jan. 2016, doi: 10.1038/nature16547.

[156] A. J. Westermann, L. Barquist, and J. Vogel, 'Resolving host–pathogen interactions by dual RNA-seq', PLoS Pathog, vol. 13, no. 2, p. e1006033, Feb. 2017, doi: 10.1371/journal.ppat.1006033.

Science and everyday life cannot

and should not be separated

Rosalind Franklin, in a letter to her father



Addendum

Dutch Summary / Nederlandse Samenvatting

Metagenomische sequentie analyse binnen de klinische virologie: Ontwikkelingen in pathogeendetectie en toekomstperspectieven

Om virussen te detecteren zoekt men in de hedendaagse diagnostiek gericht naar een of meerdere specifieke virussen. Zodra vooraf bekend is welk virus men verwacht te vinden bij een patiënt, is het mogelijk om een gerichte *polymerase chain reaction* (PCR) in te zetten. Een andere benadering is om het metagenoom in kaart te brengen, dan is het mogelijk om alle virussen in één keer tegelijk te analyseren. Hierbij wordt al het genetisch materiaal rechtstreeks uit een patiëntenmonster opgewerkt en geanalyseerd op de aanwezigheid van virale micro-organismen. Met deze methode kan ook andere genetische informatie opgespoord worden, zoals aanwezige bacteriën of andere organismen, maar ook bepaalde genetische eigenschappen van de gastheer zelf. Deze mogelijkheid heeft daarom belangrijke toekomstperspectieven. De focus van dit proefschrift ligt op de metagenomische sequentie analyse, die het mogelijk maakt om alle voor de mens pathogene virussen rechtstreeks uit patiëntmateriaal met eenzelfde test te detecteren.

De WHO rapporteert cijfers over de wereldwijd voorkomende doodsoorzaken, en waar vaak wordt aangenomen dat dit hart- en vaatziekten zijn, blijkt als de cijfers van alle verschillende infectieziektes bij elkaar worden opgeteld inclusief lagere luchtweginfecties, dat infectieziekten op nummer een te staan. Sinds de SARS-CoV-2 pandemie is meer dan ooit duidelijk geworden dat infectieziekten ook in onze westerse wereld tot ontwrichting van de samenleving zorgen. In de hedendaagse maatschappij leven wij dicht op elkaar en dicht op andere dierlijke organismen. De overdracht van virusinfecties tussen mensen en dieren vormt een constant en reëel risico. Virussen kunnen evolueren en soms ook overgedragen worden op andere organismen dan de oorspronkelijke gastheer. Dat maakt de ontwikkeling van een test waarmee in één keer naar alle virussen tegelijk kan worden gekeken rechtstreeks in patiëntmateriaal zo belangrijk. Met de bestaande diagnostische testen kunnen patiënten die besmet raakten met hetzij een nieuw, hetzij een dierlijk, hetzij met een niet verwacht virus niet of slechts moeilijk gediagnostiseerd worden.

Bij een metagenomische analyse wordt al het erfelijke materiaal, dus zowel het DNA-als het RNA-materiaal, rechtstreeks uit een patiëntenmonster geïsoleerd. Dit wordt vervolgens opgewerkt met een *library preparation* om zo het genetische materiaal geschikt te maken voor sequentie analyse. Dit gebeurt door het metagenoom te fragmenteren en door RNA om te zetten in *copy DNA* (cDNA). DNA-fragmenten worden verder behandeld door *adapter linking*, barcodering en amplificatie van het DNA-materiaal. Eventueel kan er voor het daadwerkelijke *sequencen* nog een verrijking plaatsvinden met gerichte *probe* sequenties, waardoor alle virussen worden geselecteerd door *probe* extractie. Vervolgens worden, met of zonder virale verrijking, de DNA-fragmenten op een *sequencer* geladen. Na amplificatie leest deze de nucleotiden af van de verschillende fragmenten die aanwezig waren in de patiëntenmonsters. Om deze, soms miljoenen, *reads* verder te verwerken ten einde een virus te detecteren, is speciale bioinformatische classificatiesoftware nodig.

Bij aanvang van dit promotietraject waren er nauwelijks *systematic reviews* en meta-analyses bekend die onderzoek hebben gedaan naar de hoeveelheid additionele diagnoses die zouden plaatsvinden bij gebruik van de metagenomische *sequence* test. Er waren tevens vrijwel geen onafhankelijke vergelijkingen van bioinformatische *tools*. En er waren weinig publicaties voor het verbeteren van een diagnostisch toepasbaar protocol. In deze thesis is er verder gegaan met deze openstaande vraagstellingen.

Encefalitis is een voorbeeld van een ziektebeeld waarbij een PCR-test niet in alle gevallen uitsluitsel geeft. **Hoofdstuk 2** geeft een overzicht van verschillende wetenschappelijke artikelen over onderzoek naar encefalitis met metagenomische sequentie analyse. De focus lag op het aantal additionele virussen die werden gevonden in vergelijking met de initiële traditionele diagnostische testen die waren uitgevoerd. Met een meta-analyse is er aangetoond dat er in de verschillende onderzoeken 10,9 procent extra virale diagnoses konden worden gesteld indien men naar alle virussen tegelijk zou kijken i.p.v. gericht naar enkele vooraf verdachte virussen. Omdat de sensitiviteit van deze test lager was dan een reguliere diagnostische test, is er in **hoofdstuk 3** verder gekeken naar hoe de techniek verbeterd kon worden wat betreft sensitiviteit en specificiteit door het vergelijken van patiëntenmonsters waarbij zowel metagenomische sequentie analyse werd gebruikt alsook de reguliere diagnostische tests. Daarnaast is er een vergelijking gemaakt tussen *shotgun* metagenomische *sequencing*, en metagenomische sequentie analyse met een *probe* set waarbij alle bekende virussen waarmee gewervelden organismen geïnfecteerd kunnen worden, aanwezig zijn. De sensitiviteit met het gebruik van *probes* was 100 procent t.o.v. de initiële PCR-test, en de hoeveelheid *sequence reads* nam 100 tot 10.000x toe, waardoor ook de genoomsequenties van de virussen beter konden worden bekeken. De verbeterde dekking van het genoom zorgde ook voor nauwkeuriger diagnostiek met betrekking tot de specificiteit van de virale metagenomische test. Na de technische vergelijkingen is er in **hoofdstuk 3** een aantal patiëntenmonsters nader bekeken van hematologische patiënten met een verdenking op encefalitis waar geen causaal virus of bacterie uit de initiële diagnostische testen naar voren kwam. Daar kwamen in 12,2 procent van de gevallen niet eerder bij de patiënt aangetoonde virussen uit naar voren, zoals: BK polyomavirus, hepatitis E-virus, humaan herpes virus-6 en Epstein-Barr-virus. De populatie patiënten met een onverklaarde encefalitis zou een goede patiëntenpopulatie zijn om de metagenomische sequentie analyse test in te zetten.

Reizigers uit het buitenland die terugkomen met onverklaarbare koorts behoren eveneens tot een doelgroep die bij het stellen van een diagnose mogelijk baat hebben bij de inzet van virale metagenomische sequentie analyse. In **hoofdstuk 4** is het onderzoek bij deze groep patiënten beschreven. Bij verschillende patiënten werden virussen gevonden waar eerder niet op was getest of waar de initiële antigeentest negatief was. Pathogene virussen die in 5 procent additioneel werden gevonden waren: denguevirus en hepatitis C virus. Tevens werden met deze brede benadering alle initieel gevonden virussen teruggevonden en enkele additionele niet-pathogene virussen. Verder maakt de test het mogelijk om de verschillende virussen verder te typeren, bijvoorbeeld als *subspecies*.

In deze eerdere hoofdstukken is gekeken naar de diagnostische waarde, de verbetering van het laboratoriumprotocol en de toepasbaarheid bij enkele verschillende ziektebeelden van de metagenomische test. In **hoofdstuk 5** is er gekeken naar wat de meest betrouwbare bioinformatische software is voor het accuraat detecteren van virussen bij metagenomische sequentie analyse. Hiervoor is er luchtwegmateriaal van 88 patiënten met een chronische longaandoening getest op verschillende luchtweginfectievirussen met in totaal 1120 PCR-testen. Hetzelfde luchtwegmateriaal van deze patiënten is ook bestudeerd met metagenomische sequentie analyse, om vervolgens met vijf verschillende softwareprogramma's de sensitiviteit en specificiteit van de metagenomische software te bepalen. De bioinformatische programma's hadden een sensitiviteit van 83 tot 100 procent.

Als extra onderzoek is er in **hoofdstuk 5** gekeken of het aantal *sequence reads* correleert met de ct-waarde van de PCR-test, een mate van kwantiteit. Hierbij kwam naar voren dat de meest accurate *tool* de *reads* op basis van aminozuren classificeert, al waren er grote verschillen tussen bepaalde soorten virussen. Vooral bij divergente virussen zoals rhinovirussen, bleek de kwantiteit in een patiëntenmonster het minst goed lineair te modelleren.

Hoofdstuk 6 rapporteert over een verdieping waarbij de metagenomische analyse wordt toegepast voor het kwantificeren van virussen op basis van een test die gebruikt maakt van kalibratiemonsters. Dit is gedaan in een cohort van transplantatiepatiënten, de *virusloads* in het plasma van invloed waren op de behandeling, waarbij alle patiëntenmonsters succesvol werden gekwantificeerd met metagenomische analyse.

Om na te gaan hoe metagenomische sequentie analyse functioneert bij de detectie van nieuwe of geëvolueerde virussen in patiëntenmonsters, is er in **hoofdstuk 7** onderzocht of SAR -CoV, MERS-CoV en SARS-CoV-2 rechtstreeks in patiënten materiaal gedetecteerd kon worden. Al deze corona virussen konden worden gedetecteerd door het gebruik van *capture probes*. SARS-CoV-2 monsters gaven een zeer hoge genoom dekking met gebruik van *probes* die waren ontworpen jaren vóór de ontdekking van het virus. Monsters met de desbetreffende virussen werden geanalyseerd met gebruikmaking van software waarvan de database uitsluitend virussen bevatte van alvorens de initiële ontdekking. Met de reguliere software bleken er steeds enkele *sequence reads* in te delen in de juiste virus familie. Middels *de novo assembly* werd een nieuw (deel) virusgenoom gebouwd op basis van de sequenties in het patiëntenmonster. Dat nieuwe (deel) virusgenoom werd vergeleken met de op dat moment al bekende virusgenomen, waardoor het homologie percentage berekend kon worden. Zolang een nieuw virus een zekere mate van homologie vertoont met een al bekend virus, zal het met metagenomische *sequencing* gedetecteerd kunnen worden.

Niet lang na het uitbreken van de SARS-CoV-2 pandemie zijn de specificaties van verschillende sequence technieken voor het SARS-CoV-2 genoom onderzocht. De resultaten van dat onderzoek staan beschreven in **hoofdstuk 8**. Een panel van 26 respiratoire patiëntenmonsters werden met vijf verschillende technieken en sequencing platformen behandeld. Zo is het metagenomische sequentie analyse protocol vergeleken met vier amplicon sequence technieken van drie verschillende fabrikanten. Metagenomische sequentie analyse was bij een hoge ct-waarde minder gevoelig in vergelijking met de amplicon methodes, maar de amplicon methodes leidden soms tot aspecifieke sequence reads en leverden om deze reden minder informatie over het SARS-CoV-2 genoom op. Wel is metagenomische sequentie analyse in vergelijking met amplicon methodes geschikter om onontdekte virussen te detecteren, aangezien er geen nieuw protocol voor hoeft te worden opgezet. Dit is daardoor van nut in de vroege fase van een eventuele pandemie.

Metagenomische sequentie analyse kent voordelen: met één test kan tegelijkertijd gekeken worden naar alle bekende virussen, maar eveneens naar onverwachte, en nog niet ontdekte virussen. De test kan in de toekomst eventueel nog verder verbeterd worden, zo zal een verbeterde sensitiviteit van shotgun metagenomische *sequencing* een nog betere dekking geven, en indien een real-time sequentie analyse test wordt gebruikt zal dit de doorlooptijd aanzienlijk bekorten. Metagenomische sequentie analyse zou in ieder geval uitstekend als tweede test ingezet kunnen worden in het geval bij een patiënt wel een sterke verdenking bestaat op een virus-infectie, maar de initiële testen negatief blijken.

In de toekomst kan de test mogelijkerwijs worden gecombineerd worden met virus transcriptie factor analyse om zo te bepalen of een virus actief is of slechts latent aanwezig. Samenwerking met verschillende aanverwante medische disciplines biedt eveneens waardevolle perspectieven. Zo zou het wellicht mogelijk zijn om aan de hand van metagenomische sequentie analyse genetische aandoeningen in het immuunsysteem te detecteren of om simultaan te screenen op pathologische markers of farmacogenetische informatie.

Metagenomische sequentie analyse heeft duidelijk potentie binnen en buiten de microbiologie, waarbij de ontwikkeling en potentieel multidisciplinaire toepassing nog in de kinderschoenen staat.

List of Publications / Publicaties

1. Glen R. Monroe, Gerardus W. Frederix, Sanne M.C. Savelberg, Tamar I. de Vries, Karen J. Duran, Jasper J. van der Smagt, P.A. Terhal, Peter M. van Hasselt, Hester Y. Kroes, Nanda M. Verhoeven-Duif, Ies J. Nijman, **Ellen C. Carbo**, Koen L. van Gassen, Nine V.A.M. Knoers, Anke M. Hövels, Mieke M. van Haelst, Gepke Visser, Gijs van Haaften. Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. Genetics in medicine: official journal of the American College of Medical Genetics. 2016. Sep;18(9):949-56.
   DOI: 10.1038/gim.2015.200.

2. Iris M. de Lange, Marco J. Koudijs, Ruben van 't Slot, Boudewijn Gunning, Anja C.M. Sonsma, Lisette J.J.M van Gemert, Flip Mulder, **Ellen C. Carbo**, Marjan J.A. van Kempen, Nienke E. Verbeek, Ies J. Nijman, Robert F. Ernst, Sanne M.C.Savelberg, Nine V.A.M., Knoers, Eva H. Brilstra, Bobby P.C. Koeleman. Mosaicism of de novo pathogenic SCN1A variants in epilepsy is afrequent phenomenon that correlates with variable phenotypes. Epilepsia. 2018. Vol. 59, issue 3.
   DOI: 10.1111/epi.14021.

3. Marielle E. Van Gijn, Isabella Ceccherini, Yael Shinar, **Ellen C. Carbo**, Mariska Slofstra, Juan I. Arostegui, Guillaume Sarrabay, Dorota Rowczenio, Ebun Omoyımn, Banu Balci-Peynircioglu, Hal M. Hoffman, Florian Milhavet, Morris A. Swertz, Isabelle Touitou. New workflow for classification of genetic variants' pathogenicity applied to hereditary recurrent fevers by the International Study Group for Systemic Autoinflammatory Diseases (INSAID). Journal of Medical Genetics. 2018. 2018;55:530–537.
   DOI: 10.1136/jmedgenet-2017-105216

4. Iris M. de Lange, Marco J. Koudijs, Ruben van 't Slot, Anja C. M. Sonsma, Flip Mulder, **Ellen C. Carbo**, Marjan J.A. van Kempen, Isaac J. Nijman, Robert F. Ernst, Sanne M.C. Savelberg, Nine V.A.M. Knoers, Eva H. Brilstra, Bobby P.C. Koeleman. Assessment of parental mosaicism in SCN1A-related epilepsy by single-molecule molecular inversion probes and next-generation sequencing. Journal of Medical Genetics. 2019. 2019;56:75-80.
   DOI: 10.1136/jmedgenet-2018-105672

5. Benjamin Kant, **Ellen C. Carbo**, Iris Kokmeijer, Jelske J.M. Oosterman, Joost Frenkel, Morris A. Swertz, Johannes K. Ploos van Amstel, Juan I. Aróstegui, Marco J. Koudijs, Mariëlle E. van Gijn. Gene Mosaicism Screening Using Single-Molecule Molecular Inversion Probes in Routine Diagnostics for Systemic Autoinflammatory Diseases. The Journal of Molecular Diagnostics. 2019. Vol. 21, No. 6.
DOI: 10.1016/j.jmoldx.2019.06.009

6. Anneloes L. van Rijn, Sander van Boheemen, Igor Sidorov, **Ellen C. Carbo**, Nikos Pappas, Hailiang Mei, Mariet Feltkamp, Marianne Aanerud, Per Bakke, Eric C. J. Claas, Tomas M. Eagan, Pieter S. Hiemstra, Aloys C. M. Kroes, Jutte J. C. de Vries. The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease. PLoS ONE. 2019. 14(10): e0223952.
DOI: 10.1371/journal.pone.0223952

7. Sander van Boheemen, Anneloes L. van Rijn, Nikos Pappas, **Ellen C. Carbo**, Ruben H.P. Vorderman, Igor Sidorov, Peter J. van \`t Hof, Hailiang Mei, Eric C.J. Claas, Aloys C.M. Kroes, Jutte J.C. de Vries. Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients. The Journal of Molecular Diagnostics. 2020. Vol. 22, No. 2.
DOI: 10.1016/j.jmoldx.2019.10.007

8. **Ellen C. Carbo**, Emilie P. Buddingh, Evita Karelioti, Igor A. Sidorov, Mariet C.W. Feltkamp, Peter A. von dem Borne, Jan J.G.M. Verschuuren, Aloys C.M. Kroes, Eric C.J. Claas, Jutte J.C. de Vries. Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics. Journal of Clinical Virology. 2020. Vol.130, 104566.
DOI: 10.1016/j.jcv.2020.104566.

9. Tehmina Bharucha, Clarissa Oeser, Francois Balloux, Julianne R. Brown, **Ellen C. Carbo**, Andre Charlett, Prof. Charles Y. Chiu, Eric C. J. Claas, Marcus C. de Goffau, Jutte J. C. de Vries, Prof. Marc Eloit, Susan Hopkins, Jim F. Huggett, Duncan MacCannell, Sofia Morfopoulou, Prof. Avindra Nath, Denise M. O'Sullivan, Lauren B. Reoma, Liam P. Shaw, Igor Sidorov, Patricia J. Simner, Le Van Tan, Prof Emma C. Thomson, Lucy van Dorp, Michael R. Wilson, Prof. Judith Breuer, Nigel Field. STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies. The Lancet Infectious Diseases. 2020. Vol. 20, Issue 10, E251-E260.
DOI: 10.1016/S1473-3099(20)30199-7

10. **Ellen C. Carbo**, Igor A. Sidorov, Jessica C. Zevenhoven-Dobbe, Eric J. Snijder, Eric C. Claas, Jeroen F.J. Laros, Louis C.M. Kroes, Jutte J.C. de Vries. Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness. Journal of Clinical Virolology. 2020. Vol. 131, 104594.
DOI: 10.1016/j.jcv.2020.104594.

11. Jutte J.C. de Vries, Julianne R. Brown, Natacha Couto, Martin Beer, Philippe Le Mercier, Igor Sidorov, Anna Papa, Nicole Fischer, Bas B. Oude Munnink, Christophe Rodriquez, Maryam Zaheri, Arzu Sayiner, Mario Hönemann, Alba Pérez-Cataluña, **Ellen C. Carbo**, Claudia Bachofen, Jakub Kubacki, Dennis Schmitz, Katerina Tsioka, Sébastien Matamoros, Dirk Höper, Marta Hernandez, Elisabeth Puchhammer-Stöckl, Aitana Lebrand, Michael Huber, Peter Simmonds, Eric C.J. Claas, F. Xavier López-Labrador. Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting, Journal of Clinical Virology. 2021. Vol. 138, 104812, ISSN 1386-6532.
DOI: 10.1016/j.jcv.2021.104812.

12. Alhena Reyes, **Ellen C. Carbo**, Johan S. van Harinxma thoe Slooten, Margriet E.M. Kraakman, Igor A. Sidorov, Eric C.J. Claas, Aloys C.M. Kroes, Leo G. Visser, Jutte J.C. de Vries. Viral metagenomic sequencing in a cohort of international travellers returning with febrile illness. Journal of Clinical Virology. 2021. Volume 143, 104940.
DOI: 10.1016/j.jcv.2021.104940.

13. **Ellen C. Carbo**, Ivar Blankenspoor, Jelle J. Goeman, Aloys C.M. Kroes, Eric C.J. Claas, Jutte J.C. de Vries. Viral metagenomic sequencing in the diagnosis of meningoencephalitis: a review of technical advances and diagnostic yield. Expert Review of Molecular Diagnostics. 2021. 21:11, 1139-1146.
DOI: 10.1080/14737159.2021.1985467.

14. **Ellen C. Carbo**, Anne Russcher, Margriet E.M. Kraakman, Caroline S. de Brouwer, Igor A. Sidorov, Mariet C.W. Feltkamp, Aloys C.M. Kroes, Eric C.J. Claas, Jutte J.C. de Vries. Longitudinal Monitoring of DNA Viral Loads in Transplant Patients Using Quantitative Metagenomic Next-Generation Sequencing. Pathogens. 2022. 11, 236.
DOI: 10.3390/pathogens11020236

15. **Ellen C. Carbo**, Igor A. Sidorov, Anneloes L. van Rijn-Klink, Nikos Pappas, Sander van Boheemen, Hailang Mei, Pieter S. Hiemstra, Tomas M. Eagan, Eric C.J. Claas, Aloys C.M. Kroes, Jutte J.C. de Vries. Performance of Five Metagenomic Classifiers for Virus Pathogen Detection Using Respiratory Samples from a Clinical Cohort. Pathogens. 2022. 11, 340.
DOI: 10.3390/pathogens11030340

16. **Ellen C. Carbo**, Kees Mourik, Stefan A. Boers, Bas Oude Munnink, David Nieuwenhuijse, Marcel Jonges, Matthijs R.A. Welkers, Sebastien Matamoros, Joost van Harinxma thoe Slooten, Margriet Kraakman, Evita Karelioti, David van der Meer, Karin Ellen Veldkamp, Aloys C.M. Kroes, Igor A. Sidorov, Jutte J.C. de Vries. A comparison of five Illumina, Ion torrent, and nanopore sequencing technology-based approaches for whole genome sequencing of SARS-CoV-2. European Journal of Clinical Microbiol Infectious Diseases. 2023.
DOI: 10.1007/s10096-023-04590-0

Curriculum Vitae

Ellen Carbo was born in Utrecht, the Netherlands, on February 26th, 1982. She finished her secondary education at the ‘Niels Stensen College’ in 2001, after which she went to the Technical University of Eindhoven to study Architecture and Civil Engineering. After first pursuing a career in the field of building engineering, she decided to change direction and started an undergraduate study of ‘Life Sciences’ at the Hogeschool Utrecht. During her Bachelor studies she completed two internships: one at the Department of Clinical Genetics at the University Medical Center Utrecht (UMCU), sequencing a novel breast cancer mutation gene, and one at the Department of Human Genetics at UMCU, working on next-generation sequencing (NGS) of ion channel genes in epilepsy patients. In 2013 she obtained her Bachelor degree in ‘Biomolecular Research’, after which she was hired at the place of her first internship at the UMCU to work on next-generation sequencing at the Department of Clinical Genetics.

While working at the Department of Clinical Genetics at the UMCU, she was eager to continue studying, so she started a part-time postgraduate study in biostatistics and epidemiology at the University of Amsterdam. Meanwhile, she focused her efforts on NGS data analysis (‘dry lab’), instead of laboratory work (‘wet lab’), and she gained experience in bioinformatics. For her postgraduate internship she combined her interest in genetics with her biostatistics and epidemiology study and worked on complex trait analysis. This research was performed on a large cohort of ALS patients at the UMCU Department of Neurogenetics, all while still working for the Department of Clinical Genetics at both the ‘wet’ and ‘dry’ laboratory. She obtained her Master of Science degree in 2016, after which she started working on a PhD project on auto inflammatory diseases for the Departments of Clinical Genetics at both the UMCU and the UMCG in Groningen.

To focus again on her original preferred subjects of NGS and bioinformatics, in 2018 she switched to a PhD project on viral metagenomic sequencing at the Department of Medical Microbiology at the Leiden University Medical Center (LUMC). Previously ignorant about the world of microbiology or viruses, besides occasionally experiencing annoying infections herself, she decided to dive into this new field and new adventure. She found out that she loved to unravel the nucleotides of all species, from viruses as well as bacteria, and from cucumber to her own human genome (that she analyzed herself in 2020). During her PhD she worked on improving the laboratory and bioinformatics protocols of viral metagenomics and on the clinical application of this novel approach.

Ellen is also the coordinator of the Dutch special interest group of bioinformatics in medical microbiology, as a part of the *Dutch Society of Medical Microbiology*. This group was created to form an expertise network to improve the bioinformatics within the field, and for this group she organizes symposia, online meetings, and develops courses for experts at a national level. Recently, she started at the Department of Medical Microbiology & Infection Prevention at the Amsterdam University Medical Centers and began her training as a medical molecular microbiologist. In this role, she will continue to work in both the 'wet' and 'dry' laboratory to support patient care.

We do not follow maps to buried treasure,

and “X” never, ever marks the spot.

Indiana Jones, Indiana Jones and the Last Crusade, 1989