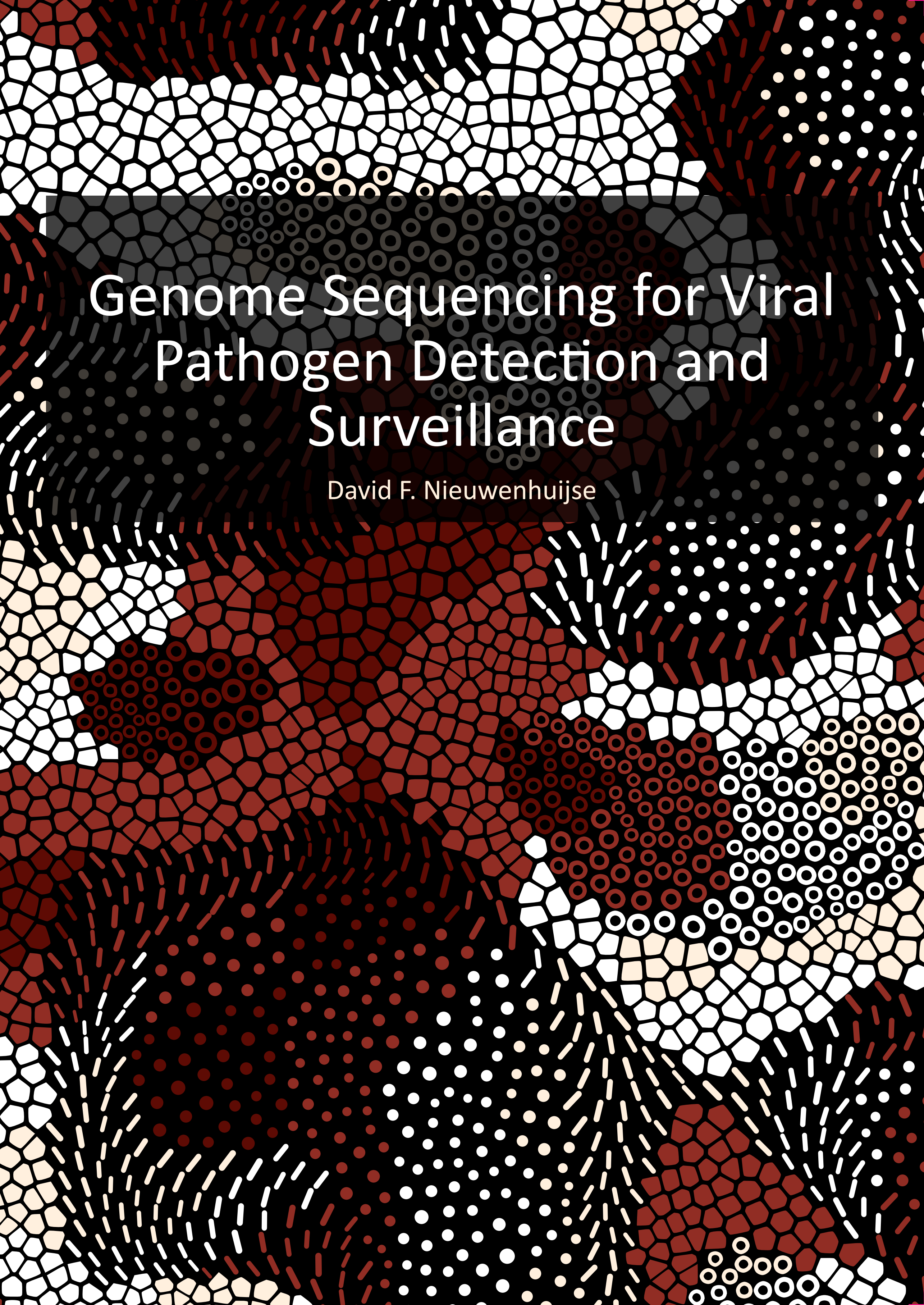


Genome Sequencing for Viral Pathogen Detection and Surveillance

David F. Nieuwenhuijse

Genome Sequencing for Viral Pathogen Detection and Surveillance

Genoom sequencing voor virale pathogeen detectie en surveillance

David Frederik Nieuwenhuijse

The research described in this thesis was performed at the Department of Viroscience, Erasmus Medical Center, Rotterdam, the Netherlands, and within the Postgraduate School of Molecular Medicine.

ISBN: 978-94-93330-01-6

Cover design by: David Nieuwenhuijse, to reproduce the front cover visit:
https://codepen.io/dnieuw/pen/rNKLKbp
Layout & printed by: Proefschriftspecialist.nl – Zaandam

The printing of this thesis was financially supported by Viroclinics Biosciences B.V.

Genome Sequencing for Viral Pathogen Detection and Surveillance

Genoom sequencing voor virale pathogeen detectie en surveillance

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus
Prof.Dr. A.L. Bredenoord
and in accordance with the decision of the Doctorate Board

The public defence shall be held on
Tuesday, 16 May 2023 at 13:00 hours

by

David Frederik Nieuwenhuijse

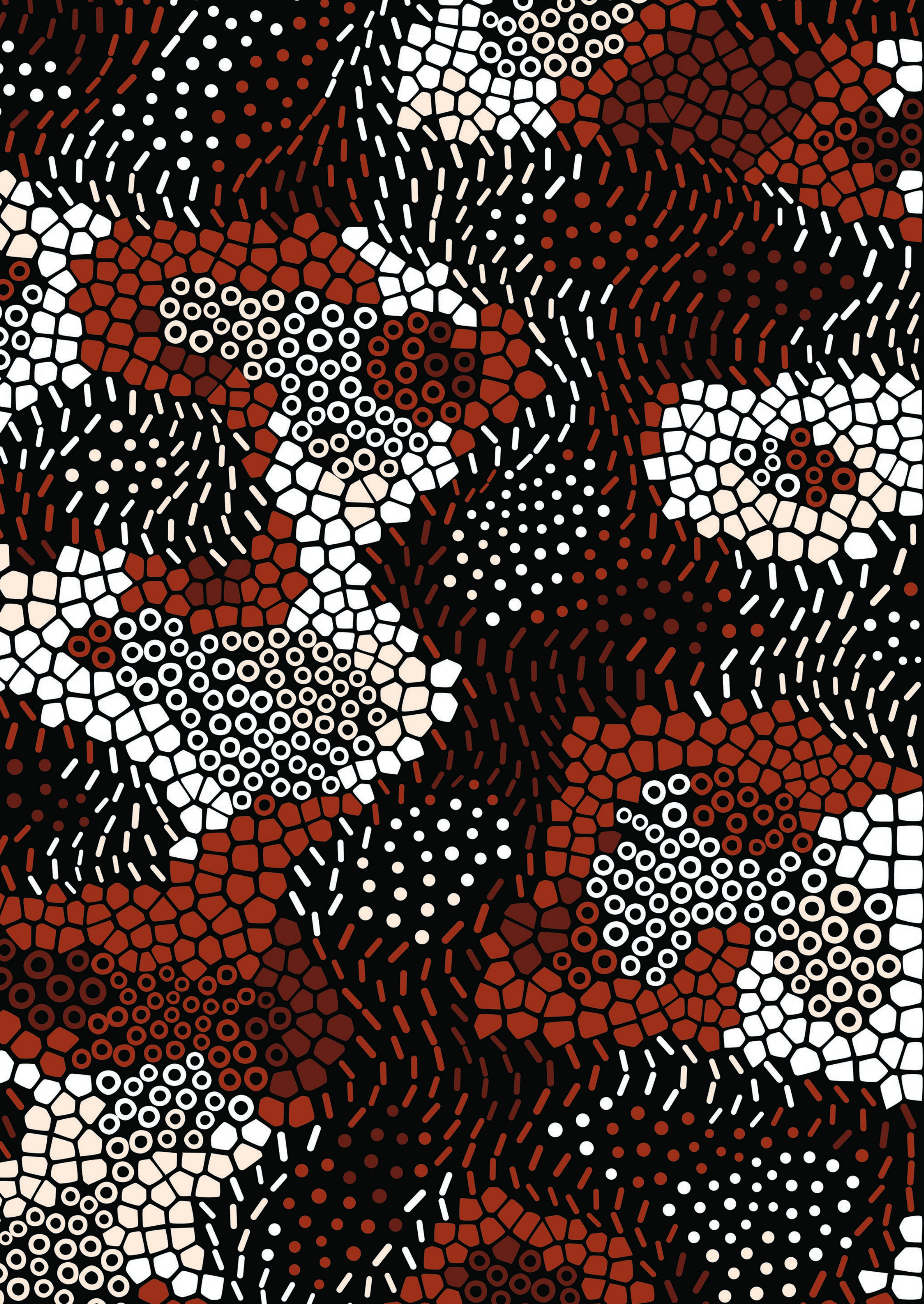
born in Woerden, The Netherlands

ERASMUS UNIVERSITEIT ROTTERDAM

Doctoral Committee

Promotor:	Prof. Dr. M.P.G. Koopmans
Other members:	Prof. Dr. R.A.M. Fouchier
	Dr. C.M. van der Hoek
	Prof. Dr. M. Kayser
Copromotor:	Dr. B.B. Oude Munnink

Table of contents



General introduction

1

1.1 Changing human population dynamics and infectious disease risk

Over the last decades, rural to urban migration has caused a population growth in cities leading to a concentration of people with more and close-contact interactions between them, which can strongly influence the spread of diseases[1]. The expansion of large cities is made possible by rapid development of transportation infrastructure which does not always go hand in hand with the expansion of public health infrastructure[2]. Although urbanization in conjunction with increasing wealth has been associated with a decreased burden of disease[3], impoverished suburbs and rural areas with a less developed health and sanitation infrastructure are now well connected to an expanding and concentrated urban population. In addition, the necessary changes in land use for urbanization, agricultural development and deforestation, with a loss of 314 million hectares between 2001 and 2015[4], has been associated with increased risk of infectious disease emergence[5]. The increased connectedness of concentrated human populations also impedes infectious disease containment, as the affected areas quickly become national and international. Altogether, the increased risk of emerging disease spreading and infecting large populations calls for better infectious disease detection and surveillance methods to aid public health governance.

1.2 Recent viral epidemics

In the last two decades several emerging viruses have caused large outbreaks in the human population. In 2002 an outbreak of SARS-CoV-1 occurred starting in mainland China and thought to have originated from bats[6]. It infected 8,098 people killing 774 and spread across 29 countries before it was contained[7]. In 2009 an H1N1 variant of the infuenza virus emerged in Mexico and caused a pandemic killing an estimated 209,300 individuals[8]. A second emerging bat related zoonotic virus[9], Ebola virus, caused a large outbreak in West Africa in 2013. Although its spread across the globe could be contained, Ebolavirus infected 28,646 and killed 11,323 individuals over a period of three years across several countries in West Africa[10]. It was during this outbreak that for the first time in real time large scale genomic and epidemiological data was collected to reconstruct the spread of the virus[11]. Later, in 2015, Brazil experienced a large outbreak of the mosquito borne Zika virus, infecting a modelled 132,3 million individuals[12] and caused a confirmed 3,720 cases of congenital syndromes in newborns[13]. Using genomic epidemiology the time of introduction of the virus could be estimated to already have occurred as early as two years before the first detected case[14]. These examples show what genomic surveillance can do to detect and monitor viral epidemics and to guide public health decision making.

1.3 Current virus surveillance and detection

In developed countries, infectious disease surveillance is mainly performed in hospitals by clinical diagnostics and shared via national communicable disease surveillance systems with the national public health centers[15]. In the early 2000s innovations in new DNA polymerases allowed for the development of easy to use and reliable real-time polymerase chain reaction (RT-PCR) based assays[16]. Since then this technology has become main method for viral detection and characterization and the golden standard in clinical microbiology laboratories for detecting single and also multiple pathogens in parallel[17,18].

Depending on the range of clinical symptoms caused by specific pathogens, the proportion of cases that are captured by hospital- or case-based surveillance can vary greatly. Infections that lead to severe disease in a large proportion of infected individuals will be easier to monitor, whereas infections for which a large proportion of cases remain asymptomatic or only develop mild symptoms are much more difficult to track resulting in an underestimation of their prevalence and burden of disease. Underestimation and under-ascertainment of infectious disease burden hampers effective decision making and resource allocation to preventative measures[19].

The surveillance bias caused by under-ascertainment is compounded by the current focus on targeted diagnostic tests, which risks underdetection of uncommon etiologies. Broadening the surveillance requires implementing multiple targeted tests, thus raising the costs. Moreover, besides detection, further characterization of the virus can be important for its surveillance[20] requiring additional typing assays and genome sequencing. These clinical surveillance systems lead to biased and fragmentated infectious disease surveillance.

1.4 The genomics-based approach

1.4.1 The development of modern sequencing platforms

Concurrent with the further development of RT-PCR assays, genome sequencing made its entry in the field of diagnostics. The colloquially called “first generation” Sanger sequencing has been the golden standard for sequence generation for a long time but does not have the throughput to generate complete genomes without unreasonable effort. Second or “next” generation sequencing platforms have alleviated that issue by generating large quantities of sequence data. However, apart from the high investment costs one of the issues of second-generation sequencing is the short size of the produced sequences hampering genome reconstruction with large insertions, deletions, and repetitive elements. The third generation of sequencing platforms is hallmarked by the increased length of the produced sequences, lower investment costs, real-time sequencing and simplified library preparation in the lab. These longer sequences however do not have the same quality as the shorter sequences, requiring deep coverage and extensive error correction post-processing. Despite the challenges, these sequencing platforms have been increasingly used in practice for virus detection and characterization[21–23].

1.4.2 Sequencing methods that benefit clinical diagnostics

Although genome sequencing has not reached the same status as RT-PCR in clinical diagnostic settings, it has several advantages over RT-PCR, one of which is the possibility to detect a pathogen without any prior knowledge of the virus. This so-called agnostic or metagenomic sequencing is a sequencing approach of which much is expected for clinical use and has been applied with success for detection of viruses in clinical samples[24]. Beside detection of known and unknown viruses, metagenomic sequencing can also detect multiple viruses at the same time, making it possible to detect co-infections[25]. One step beyond, metagenomic sequencing can be used to detect a community of viruses in complex sample types such as feces, sewage, or other environmental samples comprising viral pathogens and commensal viruses, the so called “virome”, and unknown “dark matter”[26].

The catch-all benefit of metagenomic sequencing is conversely also its main weakness as samples generally contain a lot of “background” genomic material, which, by using random amplification, can overshadow the targets of interest. This background genomic material can consist of host DNA/RNA but also for instance plant viruses and bacteriophages which share similar biological properties to human pathogenic viruses making it difficult to discriminate between them based on physicochemical properties. To that end, besides random amplification, target enrichment approaches have been recently developed to increase the sensitivity and specificity of metagenomic sequencing without losing its catch-all benefit[27].

Alternatively, when the genome of the virus of interest is known, an amplicon-based approach can be used which targets, amplifies and sequences multiple overlapping parts of a viral genome of interest to obtain the complete sequence. The advantages are that amplicon sequencing is very specific and more sensitive compared to metagenomic sequencing. However, this approach has the drawback that only the targeted virus will be sequenced and unanticipated pathogens will be missed.

The main advantage of genome sequencing is that beyond detection of the virus, much more information becomes available with which a range of follow up analyses can be performed. For example, detailed typing of the virus can be performed[28], the time of entry into the population can be estimated[11], and the source of the virus can be tracked and linked to an outbreak[29]. In view of its potential benefits, viral genome sequencing would be a valuable addition to routine viral diagnostics and surveillance.

1.4.3 Challenges of data generation

Despite the recent advances, there are still several challenges that prevent genome sequencing from being widely adopted in practice. The main challenge is the relatively high time and financial cost of genome sequencing compared to RT-PCR. Consequently, the question remains whether the added benefits of sequence information outweigh the added time and monetary investment[30]. To succeed in the field of virus diagnostics and surveillance it is therefore crucial to reduce these barriers.

The concept is therefore to replace a multitude of targeted assays with a metagenomic or agnostic sequencing approach which will simplify and unify several diagnostic assays. However, the sensitivity of metagenomic sequencing in relation to RT-PCR needs to be further investigated, as well as the standardization and optimization of sample preparation. Sequencing of samples in multiplex is also possible, which can reduce the costs and increase the number of samples that can be processed in parallel. In that case, the balance between sequencing depth and the number of parallel samples must be optimized to obtain good quality sequencing data for each sample. As mentioned before, virus enrichment protocols can potentially increase the sensitivity and specificity of metagenomic sequencing, but these gains should be compared to the additional cost and effort. Novel cheaper platforms such as the Nanopore platform could provide a cost-efficient alternative, but sequence data quality must be assessed in comparison with more established sequencing methods. Cost and time efficient sequencing comes with several challenges that, to make optimal use the added information that it can provide, must be addressed before it can be implemented in practice.

1.5 Virus sequence data analysis

With the introduction and development of novel sequencing approaches, also the approaches to analyze the their data have evolved. Because of the initial interest in the human genome many methods that are currently used were developed for large eukaryote genomics. Therefore, the underlying algorithms of these methods have often originally been created for the analysis of eukaryote or prokaryote sequence data and the assumptions made in those algorithms may not always hold true for virus sequence data. However, a multitude of methods have been developed recently specifically for the data analysis of viral genomes[31].

Virus analysis workflows are generally designed to give an overview of all viruses in a sample or zoom into a certain set of viruses based on a predefined research question. The bioinformatic workflow is then often published separately as a means to credit the data analysis method and describe it in more detail. This often results in workflows are built to be used once and for a specific task instead of focusing on broad application. Moreover, especially when a project has finished, there is little incentive to keep maintaining and updating the workflow when new versions of the software it uses or new databases appear that potentially cause errors. The fact that workflows are generally designed for a specific research question also makes it difficult to select a workflow that fits the analysis requirements for a different dataset and research question. Therefore, there is a need for generalized and customizable data analysis workflows that are aimed at reusability[32]. Besides data processing, the interpretation of genome sequencing data can be complex. Especially metagenomic and agnostic sequencing has the challenge that often only fragments of viruses of interest are found, which hinders sequence annotation, which is further complicated by incomplete or misannotated viral reference databases[33]. Correct interpretation of viral genome sequencing data therefore requires strong tools for data visualization and interrogation that give a broad perspective but can focus on a specific research question or virus of interest.

1.6 Outline of this thesis

Chapter 2 is an introductory review about the challenges of the implementation of metagenomic sequencing for viral surveillance in food- and water borne viruses. In **chapter 3** a comparison of viruses present in global sewage is described which can be used as a baseline for sewage based viral surveillance. **Chapter 4** addresses the issue of viral metagenomic data analysis and introduces a user-friendly tool to facilitate the annotation and the comparison of different samples based on certain metadata. **Chapter 5** focuses on the comparison of several methods and platforms for whole genome sequencing for public health purposes and **chapter 6** further validates the usage of the Nanopore platform for viral whole genome sequencing. In **chapter 7** the use of near to real-time whole genome sequencing on the Nanopore platform is shown to be beneficial for public health decision making at the start of the SARS-CoV-2 outbreak in 2020 in the Netherlands. Further application of Nanopore sequencing in tracking and tracing SARS-CoV-2 in hospital settings is demonstrated in **chapter 8**. **Chapter 9** describes a tool that can be used to check the primers used in viral diagnostic RT-PCR or amplicon based viral genome sequencing.

Metagenomic sequencing for surveillance of food- and waterborne viral diseases

David F. Nieuwenhuijse[1] and Marion P. G. Koopmans[1*]

Microbiology 8, 230, 2017

2

2.1 Abstract

A plethora of viruses can be transmitted by the food- and waterborne route. However, their recognition is challenging because of the variety of viruses, heterogeneity of symptoms, the lack of awareness of clinicians, and limited surveillance efforts. Classical food- and waterborne viral disease outbreaks are mainly caused by caliciviruses, but the source of the virus is often not known and the foodborne mode of transmission is difficult to discriminate from human-to-human transmission. Atypical food- and waterborne viral disease can be caused by viruses such as hepatitis A and hepatitis E. In addition, a source of novel emerging viruses with a potential to spread via the food- and waterborne route is the repeated interaction of humans with wildlife. Wildlife-to-human adaptation may give rise to self- limiting outbreaks in some cases, but when fully adjusted to the human host can be devastating. Metagenomic sequencing has been investigated as a promising solution for surveillance purposes as it detects all viruses in a single protocol, delivers additional genomic information for outbreak tracing and detects novel unknown viruses. Nevertheless, several issues have to be addressed in order to apply metagenomic sequencing in surveillance. First, sample preparation is difficult since genomic material of viruses is generally overshadowed by host- and bacterial genomes. Second, several data analysis issues hamper the efficient, robust and automated processing of metagenomic data. Third, interpretation of metagenomic data is hard, because of the lack of general knowledge of the virome in the food chain and the environment. Further developments in virus specific nucleic acid extraction methods, bioinformatic data processing applications and unifying data visualization tools are needed to gain insightful surveillance knowledge from suspect food samples.

2.2 Introduction

Transmission via the food- and waterborne route is a common mode of spread of a wide range of viruses. Many commonly recognized food- and waterborne infections are caused by viruses that are transmitted by the fecal-oral route and can cause diarrhea and vomiting, particularly caliciviruses (norovirus, sapovirus), and less commonly astroviruses, rotaviruses, and adenoviruses. Other viruses cause symptoms resulting from extra-intestinal spread, like hepatitis A (HAV), and hepatitis E (HEV). High levels of viral shedding through stool and vomit lead to dispersal in the environment, and the stability of many food- and waterborne viruses allows for prolonged persistence in the environment. Food- and water associated transmission is also suspected to enhance the spread and emergence of zoonotic viruses (e.g. Middle East Respiratory Syndrome-coronavirus and Nipah virus) and facilitate the occurrence of zoonotic events though the handling of bushmeat (Ebola virus).

Challenges of detecting viruses transmitted by the food- and waterborne route are their diversity, and the frequent secondary person-to-person transmissions, which may mask an initial food- or waterborne introduction. In addition, there is a lack of awareness among clinicians, as the symptoms caused by food-borne viruses are not specific to the viruses causing the illness, and there is limited coverage in surveillance of food- and waterborne viral disease, hampering detecting and tracing.

In the past years, high throughput sequencing technologies have increased the ability to measure genomic material from diverse samples tremendously, and these methods will most likely continue to improve in the future. Specifically, metagenomic analysis using untargeted sequencing, has received a lot of attention, because the high throughput of current sequencing technologies has made it possible to obtain multiple high coverage genomes from highly complex samples[34,35]. Even though it is still a developing field, metagenomics is starting to become mature enough for applications outside of the research environment.

With the development of multiplex real-time polymerase chain reaction (RT-PCR) protocols came the realization that unraveling etiologies of main disease syndromes is more complex than previously recognized, leading to questions about the detection of viruses for which the role as causes of illness remains to be evaluated, the importance of co-infections and recognition of less common disease etiologies. Similarly, high throughput metagenomic sequencing broadens the scope of detectable viruses, which, apart from making it more complex, make us further understand the role of viruses in health and disease. The biggest promise, however, is that of routine application of metagenomic sequencing in diagnostic context, facilitating viral detection and offering huge potential for tracing of viruses in (food-borne) outbreaks.

2.3 Recognizing food- and waterborne viral disease

Given the number of different viral pathogens potentially associated with food- and waterborne transmission, and the lack of cell culture systems that are sensitive and robust enough for application in routine settings for many of these pathogens, their detection has not been straightforward. The entry point for disease-based surveillance of viruses spreading by food and water is the reporting of patients presenting to a clinician. However, patients only present themselves in case of a severe symptomatic infection, or in case self-help is not sufficient. Mild symptoms are therefore generally not registered creating a bias in surveillance. This phenomenon is captured in the surveillance pyramid (**Figure 1**), and the full extent of disease can only be captured through epidemiological studies addressing incidence and etiology at community level coupled with severity of a range of enteric pathogens[36–38]. Additionally, it is challenging to distinguish between food-borne outbreaks and outbreaks caused by direct contact between humans. Classic clinical symptoms of food-borne disease vary, ranging from diarrhea and vomiting to abdominal cramps and general malaise which makes it hard for clinicians to pinpoint the exact causative agent and leads to misdiagnosis if the diagnostic work-up is selective, and if there are no obvious signs of food related exposure. Moreover, heterogeneity in clinical interpretation can be caused by host factors, such as differences in the expression of histo-blood-group antigens that are receptors for rota- and noroviruses[39,40]. Susceptibility to fecal-orally transmitted viruses may also be influenced by the established microbiome and virome in the host population, of which the prior is shown to differ between different locations and age groups[41]. It is reasonable to think that the differences in the gut environment are more pronounced between countries with larger social and economic differences such as first and third world counties, which often differ in their resident pathogens. The role of the gut virome, in addition to the gut microbiome, is a relatively new concept and has been described as potentially having influence on gut health and therefore expression of disease[42]. Because of under and miss-diagnoses, clinical surveillance likely only captures the tip of the iceberg of food- and waterborne viral disease cases.

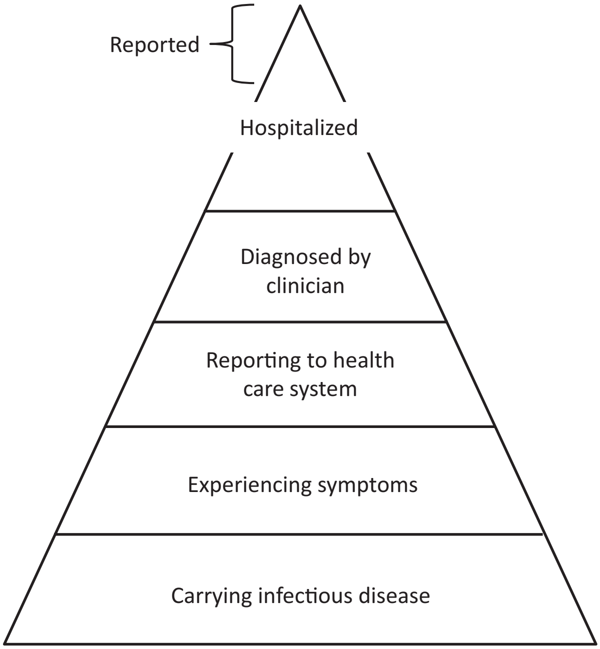


Figure 1 Schematic representation of the phenomenon known as the "surveillance pyramid'. Layers represent different categories of infected individuals. Width of the layers represents the estimated number of individuals in that category. As indicated, individuals reported by surveillance programs generally originate from the hospitalized category.

2.4 Detection of food- and waterborne viral disease outbreaks

In cases where a cluster of patients with similar symptoms presents, there can be an investigation to look for epidemiological clues of the link between the cases. Additional information is garnered from the use of viral genome sequencing, making it possible to track origins of outbreaks, and to estimate how much of the observed human disease is attributable to foodborne infection by computerized linking of epidemiologic data to aligned viral genomic sequences[43]. However, often the original source or evidence of it being food- or waterborne cannot be found, which means that outbreaks often are merely registered. Of the 941 viral disease outbreaks reported as foodborne in the joint ECDC-EFSA surveillance report of 2015, only 9.1% had robust evidence of food- or waterborne transmission[44]. Routine application of genotyping of HAV in newly diagnosed cases quadrupled the number of cases in which food was the most likely source of infection a 3 year enhanced surveillance study in The Netherlands, but this is not commonly done[45]. In an investigation of 1794 food- and waterborne outbreaks in Korea, roughly 75% of the outbreaks reported in schools and public restaurants were attributed to an unknown origin[46]. Availability and costs of molecular testing combined with sequencing, additional to the limited success of virus detection in food products, are likely further limiting their use in food and water surveillance. This is demonstrated by the fact that formal confirmation of a viral outbreak associated to food- and waterborne transmission still requires extensive epidemiological analysis or confirmation of a virus in the infected individual, or both[47]. However, due to the increase of genomic information of viruses, sequence data is increasingly used to support and strengthen outbreak investigations. Nevertheless, the surveillance programs for these viruses in the human food chain is limited, in contrast with the American CDC[1] and the European ECDC[2] surveillance programs for bacteria and parasitic pathogens causing food- and waterborne diseases[48] and does not have wide spread coverage. As an example, to comply to European food safety regulations, shellfish, a well-known source of food-borne pathogens, need to be tested for enteric bacteria. However, it has been well documented that shellfish that pass quality control based on bacterial counts may still contain human pathogenic viruses[49]. To be able to recognize food- waterborne viral disease outbreaks and stop underestimation of its disease burden there should be innovations in the current food-born surveillance system.

2.5 Classical viruses associated with food- and waterborne diseases

Although the list of viruses causing acute gastroenteritis is long, norovirus ranks among the top causes of diarrheal disease[50]. Reporting of outbreaks suggests that the food- and waterborne disease transmission route is relatively rare, but provides an underestimate, bearing in mind that it may be hard to recognize a food- and waterborne transmission route in community acquired diarrheal disease. To quantify the burden of all diarrheal disease attributable to foodborne transmission, the World Health Organization commis-sioned a study that combined data from surveillance and exhaustive literature reviews with a systematic approach to calculation of the fraction of disease attributable to food contamination[51]. This ranked the burden of norovirus illness among the top causes of foodborne disease, along

[1] http://ecdc.europa.eu/en/healthtopics/food_and_waterborne_disease/surveillance/Pages/index.aspx

[2] http://www.cdc.gov/ncezid/dfwed/keyprograms/surveillance.html

with *Campylobacter*, and listed HAV associated disease among other significant causes of foodborne disease, along with *Salmonella* and *Taenia solium*.

For bacterial foodborne pathogens, the analysis of systematically collected surveillance data has been used as the basis of attribution analysis[52]. A popular approach has been to quantify the proportion of foodborne disease of humans to their likely origin, by comparing diversity of strains found in human disease outbreaks with that found in animal and environmental reservoirs[53]. While this model does not allow estimating the foodborne disease where food is a vehicle for person-to-person transmission, which is common for noroviruses, it has been used with some success to quantify the contribution of foodborne viral disease stemming from environmentally contaminated food (e.g. associated with shellfish;[54]). This builds from the observation that there is a large discrepancy between the norovirus variants in clinical settings and environmental samples[55,56]. Norovirus GII.4, found in clinical setting, is generally related to person-to-person transmission, however several other norovirus genotypes and genogroups were found in environmental samples in the same area. However, food associated AGE is not limited to norovirus infections. In a large retrospective study of oyster related AGE outbreaks in Osaka City in Japan 30.7% of the cases were attributed to other pathogens such aichivirus, astrovirus, sapovirus rotavirus A and enteroviruses[57]. Furthermore, outbreaks can be caused by a mixture of these viruses and viral variants[58].

2.6 Other viruses transmitted via the food- and waterborne route

Apart from viruses causing gastro-enteritis, there are viruses causing food- and waterborne diseases that are associated with a variety of other syndromes. The second most common disease syndrome is hepatitis, caused by HAV,a fecal-orally transmitted virus[51]. By decreasing natural exposure in regions with low endemicity, the susceptibility of the population for outbreaks of HAV disease in these regions is increasing[59]. Because of increased globalization, contamination of food products by viruses prevalent in food producing regions can increase the risk of outbreaks in these regions. Several outbreaks of HAV infection have been reported in recent years both in the USA and Europe[60]. Most of these outbreaks could be identified as food-borne infections after intense investigations[61]. Especially fresh (imported) food products (e.g. fresh frozen berries, pomegranate seeds and sun-dried tomatoes) have been identified as sources of the virus[60,62]. Tracking the foodborne source of infection is challenging for HAV due to the long incubation period in infected individuals, underestimating the contribution of food as a source of infection[45].

Another foodborne virus gaining increased attention is zoonotic HEV, associated with genotype 3 and 4 HEV. HEV is widespread in commercially held pigs, as well as in wild pigs and deer[63]. Human disease with genotype 3 HEV is increasingly recognized, but in the large majority of the cases the source of the virus is unknown[64]. There is clear evidence that food can be a source of zoonotic HEV infections. Outbreaks that have been confirmed to be caused by food-borne transmission of the virus by consumption of wild meat from boar, deer and rabbit[63,65,66]. Several studies have shown the zoonotic potential of HEV from pigs[67], HEV can also be readily detected in pork products such as dried meats and liver sausages[68]. A large proportion of food related HEV infections, however, does not lead to hospitalization of the patient, leading to under reporting and unrecognized risk and burden of the disease[63].

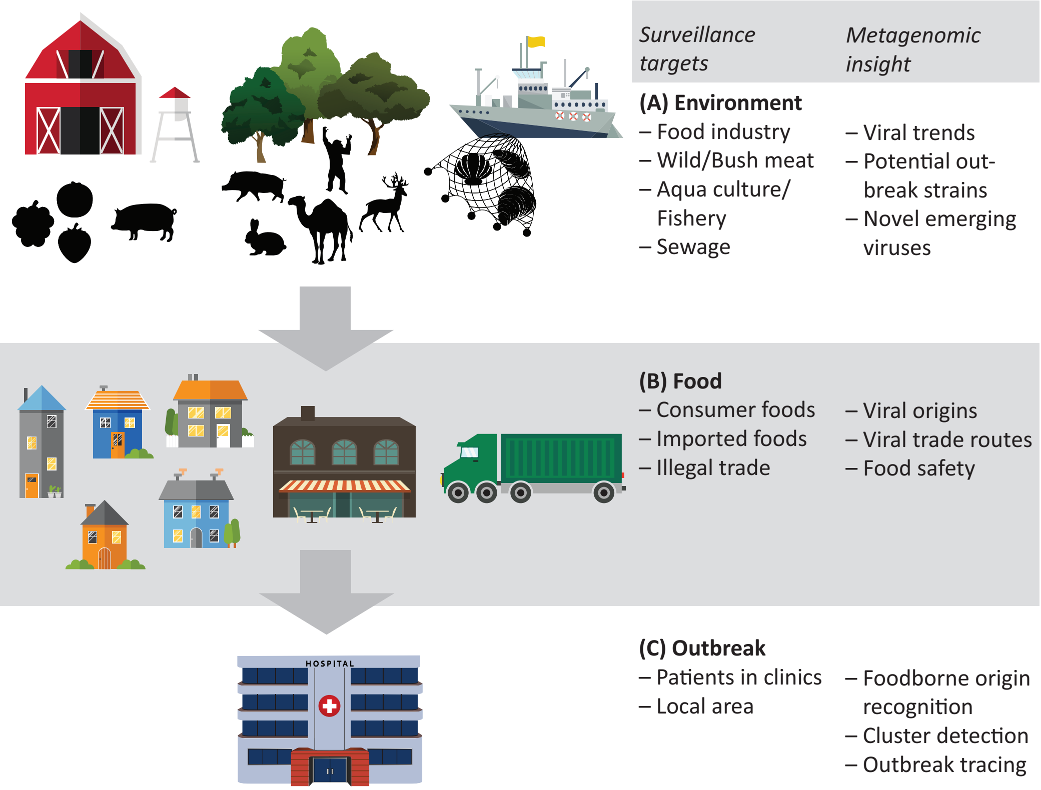
Beside viruses circulating in livestock, wildlife has the potential to be a large reservoir of unknown zoonotic viruses. Hunting, trading, preparing and consuming so called "bush-meat" is one of the routes by which novel viruses can be introduced into the human population[69]. It may be difficult to disentangle foodborne infection from direct zoonotic exposure, but it is important to consider local practices before ruling out food as a source of human infection. A special example are the occasional introductions of Nipah viruses from bats into humans through contamination of date palm sap which is collected in open containers to which bats that harbor these viruses have access[70]. Not proven but certainly interesting is the practice of drinking unprocessed camel urine which may contain MERS coronavirus, a practice that came to light during the investigations into sources of MERS coronavirus infection in humans[71]. Even if limited in scale, small food-borne infections, originating from human-wildlife interaction, constitute as many incidents potentially pushing wildlife viruses to become human-to-human transmissible[72,73]. In the cases of Monkeypox and Nipah this only led to small epidemics, but when the virus is well adapted to spread from human to human this can lead to larger outbreaks, as seen during the Ebola crisis in 2015[73,74]. Continuing deforestation, increasing population and continued trade of bushmeat brings more humans in contact to wildlife and increases the risk of zoonosis[69]. Urbanization and globalization of travel and trade provides ample and increasing opportunity for further spread. Therefore, even anecdotal zoonotic introductions may constitute a public health risk, and ideally should be investigated in conjunction with the animals these humans were exposed to. As the ability to spread between humans is a key property for successful further spread, enhancing the capacity to investigate clusters of disease (in humans and animals) is important[75].

2.7 Unknown fecal-oral passengers

Bacteriophages, although not directly pathogenic to humans, could play a role in human health and disease by influencing the gut microbiome. Sequencing data from human gut samples presents a large diversity of bacteriophages in the human gut[76]. In addition to bacteriophages, untargeted sequencing of sewage samples has shown the presence of large quantities of different plant viruses[77]. Because of the presence of numerous infectious plant virus particles in human fecal waste there is ongoing research of the effect of these viruses in human health and disease[78]. Similarly, there is ongoing research into the impact of bacteriophages on human health through their modulating effect on the gut microbiome[76], and thereby, gut immunity[79]. In what way bacteriophages protect or expose the human gut to bacterial or viral pathogens has yet to be further investigated. However, using metagenomic sequencing it will at least be possible to recognize the presence of unknown fecal-oral passengers.

2.8 Metagenomics for food- and waterborne viral disease surveillance

Metagenomics is a term used for experiments in which all nucleic acids in a certain sample are sequenced. For bacteria, historically, the diversity of a sample used to be expressed by performing phylogenetic analyses based on 16S ribosomal RNA[80]. However, since viruses lack such a universally conserved motif, viral metagenomics refers to the attempt to recover full and partial genomes of all viruses present in the sample. Viral metagenomic analysis protocols generally start with procedures to remove host and bacterial cells followed by nuclease treatment to remove free nucleic acids. Often, the remaining nucleic acids are amplified using randomly primed (RT-) PCR and finally sequenced using high-throughput sequencing technology. Viral metagenomics has great potential in surveillance of viruses in the global food-chain because of its sensitivity, broad detection range and detailed information of the detected virus.



Surveillance targets	*Metagenomic insight*
(A) Environment	
– Food industry – Wild/Bush meat – Aqua culture/ Fishery – Sewage	– Viral trends – Potential outbreak strains – Novel emerging viruses
(B) Food	
– Consumer foods – Imported foods – Illegal trade	– Viral origins – Viral trade routes – Food safety
(C) Outbreak	
– Patients in clinics – Local area	– Foodborne origin recognition – Cluster detection – Outbreak tracing

Figure 2 Food- and waterborne viral surveillance targets for metagenomic sequencing approaches. (A) Environmental surveillance of food industry, wild meat and bushmeat habitat and aquaculture and fishery environment. (B) Food surveillance of consumer and imported foods, including illegally imported foods. (C) Surveillance of food- and waterborne outbreaks, in clinic and locally. Potential of metagenomic sequencing based surveillance is listed next to each category.

2.8.1 Environmental Surveillance

Metagenomic sequencing has already been used in the sampling of the world’s oceans to estimate the global viral diversity[81]. Similarly, metagenomics can be used in environments associated with viruses spread via the food- and waterborne route (**Figure 2A**), which gives an overview of all these viruses and circumvents the mentioned sampling biases. The potential of such an approach for food related purposes was exemplified by Hellmér and colleagues who conducted a multi-species viral surveillance study and, albeit not metagenomic sequencing based, were able to detect several food- and waterborne viruses in sewage. Interestingly, norovirus and HAV, detected in sewage, could be related to hospitalized patients diagnosed with the viral infection in the catchment area of the sewage system. Moreover, they detected a peak in the level of norovirus several weeks before the outbreak was reported in the hospital in that area[82]. This demonstrates the po-tential power of shifting the scope of surveillance of food- and waterborne viruses from the hospital to the environment. Untargeted metagenomic sequencing has been shown to be able to capture a multitude of viruses in sewage samples in several studies. Moreover, comparison between sewage viromes from Nigeria, Nepal, Bangkok and California, four geographically distant locations, showed distinct differences in the subsets of detected human viruses[83]. Interestingly, the average sequence similarity between the reference sequences stored on the NCBI GenBank and the human viruses detected in the samples from California was higher than those from the other locations. This may indicate a bias towards American viruses in view of human virus diversity in this database[83]. A study in which a human epithelial cell culture was inoculated with sewage, to increase the relative number of human viruses capable of infecting these cells, was able to detected, apart from a large number of bacteriophages, several different species of the *Polyomaviridae, Picornaviridae* and *Papillomaviridae* viral families[84]. Another more recent evaluation of untargeted metagenomic sequencing for surveillance purposes retrieved full genomes of Adeno-associated virus-2 as the most prominent mammalian virus in the sample. This virus is generally not associated with any pathology and cannot be grown in cell cultures, possibly underestimating its role in diarrheal disease[85]. A striking fact of these studies is the number of sequencing reads that are found that share no sequence similarity with current reference databases. Percentages of unmapped sequences range from 37% to 66%[83,86]. Whether these sequences represent novel viruses that can be transmitted via the food- and waterborne route remains to be determined. Nevertheless, these preliminary studies show the potential of untargeted metagenomic sequencing to detect novel and known human pathogens. Sampling a larger variety of locations, performing longitudinal studies of the same environment and deeper sequencing will provide more information on what environmental metagenomic sequencing can contribute to the monitoring of viral trends and viral diversity.

2.8.2 Food surveillance

Analogous to the environment in which it has been produced, food itself can benefit from metagenomic surveillance. Food contamination in combination with international trade, changing eating habits and food processing practices all contribute to the spread of food-

and waterborne viruses and making food itself a valuable target of metagenomic surveillance (**Figure 2B**). Sentinel screening of imported foods, especially risk foods such as fresh fruits and vegetables, dried meats and seafood, could prevent foodborne viral outbreaks such as the international HAV outbreak in Europe from 2012 to 2013[87]. Successful application of metagenomic sequencing of viruses has been shown in a study isolating viruses in the family of *Reoviridae* and *Picobirnaviridae* from field-grown lettuce[84].

Apart from legal trade, illegal import of food products, such as bushmeat, could also be screened. Untargeted metagenomic sequencing is especially suited for these types of screenings, as the origin and potential viral content of these samples is often completely unknown. In one example, metagenomic sequencing was performed on bushmeat seized by the customs officers of a French airport. Although no viruses with potential threat for human health could be detected[88], these initial attempts should be looked at as potentially interesting surveillance approaches, given that relative large quantities of raw bushmeat are estimated to enter Europe and the Americas annually[74].

Another source of known and potentially unknown food-borne disease causing viruses are shellfish. Mainly the consumption of oysters is associated with food-borne outbreaks[89]. However, oysters, cockles and clams have been shown to accumulate norovirus, sapovirus and HAV[90]. To our knowledge, there are no published studies performing untargeted virome sequencing of these shellfish. Surveillance by metagenomic sequencing can be beneficial for aquaculture, also for monitoring seafood health, as in aquaculture, large numbers of animals are kept in a confined environment for an extended period, increasing opportunities for spread of infections. Cultivated fish and other sources of seafood can be infected with a wide variety of viruses[91].

2.8.3 Outbreak surveillance

One of the main promises of surveillance using metagenomic sequencing is that of concomitant clinical application (diagnosis of patients) and public health application (typing and cluster analysis to trace of food- and waterborne outbreaks) (**Figure2C**). Using metagenomic sequencing, the effort of detecting and genotyping of a virus can be combined to trace an outbreak, regardless of prior knowledge of the virus, provided the data is analyzed in combination with relevant metadata. The use of this integrative approach has been demonstrated in an investigation of a hospital outbreak of human parainfluenza virus, which was investigated using high-throughput metagenomic sequencing[92]. Both the detection of the virus, the diagnosis of the disease and the establishment of viral clusters and transmission routes could be derived from the metagenomic sequencing data. A similar approach should enable investigation of viruses related with food- or waterborne diseases and distinguishing between a food- and waterborne and a person-to-person transmission route. In such investigations, speed is of the utmost importance, therefore on-site sequencing strategies, enabled by novel portable sequencing platforms such as the Oxford Nanopore MinION[93], have potential in fast local outbreak detection and disease monitoring[94]. Recent reports have shown potential in metagenomic detection of hepatitis C, chikungunya, Ebola and Zika virus in hospital settings[95,96]. The development of on-site sequencing technology is still in its infancy, however, and it remains to be investigated if food related viral outbreaks will be traceable, and can deliver whole-genome based viral dynamics analysis analogous to the investigation of the Ebola outbreak of 2014[97,98]. However, the same on-site technology has been shown to be beneficial in tracing foodborne salmonella[99]. Aspects of current on-site sequencing technologies that need to be improved for viral metagenomic sequencing are the limited throughput and sequence quality, which limit the detection of low level viral genomes and minor variants. Nevertheless, the use of near-real-time sequencing of Ebola and Zika during the recent outbreaks has received a lot of attention and has shown that the technology works.

2.9 Challenges in complex viral sample preparation

The initial barrier to reconstructing whole-genomes from all viruses in complex samples is obtaining enough high quality viral nucleic acid for reliable viral genome coverage. Approaches which try to enrich all viral sequences are being investigated, ranging from different extraction protocols[34,100] to using a virome specific capturing chip[27] or blood-derived antibodies to capture viral particles[101]. Paradoxically, however, the sequencing capacity of high throughput metagenomic sequencing is sensitive enough to pick up contaminants from the lab reagents, of from previous experiments[102]. These pose a challenge to the interpretation of metagenomic data. To limit contamination, laboratories in which samples are processed are often separated from those in which nucleic acids are amplified and equipment is UV treated and cleaned with bleach. Additionally, alternating the sample-specific DNA barcodes in multiplex sequencing experiments reduces contamination from previous runs. Nevertheless, it is recommended to include both negative control samples, which have been processed similarly, but are believed to contain no viruses, and positive control samples, which contain known quantities of a variety of viruses[103]. Alternatively, bioinformatics tools such as DeconSeq[104], have been developed to check for signals of regularly found lab contamination in the sequencing data. In conclusion, as contamination of samples and equipment may not be avoidable, its likelihood should be taken in consideration when using metagenomic sequencing technology for food-related surveillance applications.

2.10 Difficulties of metagenomic data analysis

Aside from lab-based technical difficulties there are several challenges concerning data analysis of metagenomic sequencing experiments. First, due to the high and increasing read output of sequencing machines, data analysis of high throughput sequencing projects generally requires strong computational infrastructure, which, can require large investments and technological expertise[105]. However, metagenomic data analysis tools have been improving, optimizing the ratio between computing resources needed and their speed and accuracy. Sequence annotation tools based on k-mer lookup tables such as UBLAST[106], Kraken[107], Kaiju[108] and Diamond[109] have increased the speed of sequence assignment to reference database with several orders of magnitude, while requiring relatively modest processing power.

Second, the assembly of millions of genomic fragments into thousands of different individual genomes is a daunting task. Historically, short-read assemblers were developed and optimized to assemble a single genome out of a set of sequencing reads. These assemblers are therefore not suited for the reconstruction of metagenomes, and are prone to creating synthetic chimeric genomes[110]. Various assemblers have since been developed specifically aimed at metagenome assembly, like MetaSPAdes[111], Ray-Meta[112], MetAMOS[113], MetaVelvet[114], and IDBA-UD[115]. Nevertheless, metagenome assembly is still a challenging task, often requiring manual editing to resolve miss-assemblies.

Third, assigning all assembled genomes to a reference genome is hampered by miss-annotations and incomplete reference databases. One example is “non-A, non-B hepatitis virus”, a sequence present in the NCBI GenBank, which was miss-annotated and the sequence was shown to belong to a bacteriophage[86]. The volume of sequencing databases is increasing rapidly, however sequence annotations and metadata are of varying levels of quality and the speed of analysis decreases with increasing reference datasets. Therefore, there is a tradeoff between the rate of success of annotation of a sequence against a smaller curated reference dataset, and reliability of annotation using a large reference database with less-well curated annotation data.

Sequence homology of multiple reference genomes can lead to spurious assignment of sequencing reads to one of these genomes. An example of the impact of spurious read annotation was the alleged detection of genomic material of *Yersinia pestis* in the New York subway system. Further inspection showed that the reads mapping to *Yersinia pestis* could have mapped with similar likelihood to other bacterial species[116]. Such miss-annotations of metagenomic sequences need to be anticipated and carefully addressed before using metagenomics in surveillance and diagnostic applications.

2.11 Metagenomic data interpretation

The final challenge of metagenomic sequencing based surveillance is the interpretation of the annotated sequences. There is still little knowledge of the presence and dynamics of viruses in the environment and the food-chain, which is of influence on the interpretation of food- and waterborne viral surveillance samples. Various factors are expected to influence the virome, and without knowledge of the typical viral content of a sample, the relevance of the detection of a virus is hard to determine. An example of this is a study showing a large discrepancy between the levels of HAV genotypes detected in sewage samples compared to the genotype infecting patients in the clinic in the same time[117]. A potential sampling bias and asymptomatic shedding of one of the variants was proposed as an explanation of the discrepancy. However, this shows that a lack of knowledge of viral diversity in a population under surveillance could potentially lead to wrong conclusions in environmental surveillance studies. It is becoming increasingly clear that integration of different data sources and experimental results is crucial for the interpretation of metagenomic sequencing experiments. Therefore, browsing of these data and visualization of relationships between genome datasets and metadata should be facilitated. In the recent years, interactive web-based data browsing and visualization tools have increased in popularity to facilitate the interaction with and the browsing through highly complicated data in a user-friendly manner. Further development of tools that facilitate interaction with and visualization of metagenomic sequencing results, such as Kronatools[118] and Taxonomer[119], and frameworks for so-called data analysis “dashboards”[3,4,5], should make the interpretation of metagenomics experiments easier in the future.

[3] http://shiny.rstudio.com/

[4] https://plot.ly/

[5] http://jupyter.org/

2.12 Conclusion

In our current society, there is much attention for the diseases that are causing occasional outbreaks. However, there are multiple strong signs that there are viruses hiding below the radar, due to a focus on viruses with direct clinical impact. As such, the disease burden of food- and waterborne viral infections is mainly recorded in outbreaks, signified by severe symptoms and hospitalization. However, it is estimated that the large abundance of viral infections causing mild symptoms, and thus not being recorded, carry a large portion of the global food- and waterborne disease burden. Moreover, this disease burden is expanded by the consequential infections and outbreaks of these viruses in susceptible populations. Global food trading, diversification of food sources and interactions with animals and other reservoirs of food- and waterborne disease related viruses complicate the capability of investigators to detect the original source and to determine the transmission pattern of viruses causing foodborne outbreaks. Therefore, surveillance efforts should look to metagenomic sequencing technologies, bioinformatics analysis tools and data sharing initiatives to get a more realistic insight in the global burden of food- and waterborne viral disease, and to make informed decisions on how to reduce this burden.

Author Affiliations

[1] Department of Viroscience, Erasmus Medical Center, P.O. Box 2040, 3000 CA, Rotterdam, The Netherlands.

**** Correspondence***

M.P.G. Koopmans, Department of Viroscience, Erasmus Medical Center, P.O. Box 2040,
3000 CA Rotterdam, The Netherlands, tel. +31 (0) 10 7044066, fax. +31 (0) 10 7044760,
m.koopmans@erasmusmc.nl

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

DN and MK designed the focus of the review, DN did the literature search and drafted the manuscript, MK reviewed and revised.

Funding

European Union's Horizon 2020 research and innovation program under grant agreement No 643476 (COMPARE). ZonMw TOP project 91213058

Acknowledgments

We thank Matt Cotten, Miranda de Graaf, Bas Oude Munnink and My Phan for their critical review of the manuscript and valuable discussions.

We would like to give credit to www.vecteezy.com, who provided the vector illustrations used to create the figures in this manuscript.

Setting a baseline for global urban virome surveillance in sewage

David F. Nieuwenhuijse[1]*, Bas B. Oude Munnink[1]*, My V. T. Phan[1]*, the Global Sewage Surveillance project consortium, Patrick Munk[2], Shweta Venkatakrishnan[1], Frank M. Aarestrup[2], Matthew Cotten[1], Marion P. G. Koopmans[1]

Scientific Reports 10 (1), 1-13, 28, 2020

3

3.1 Abstract

The rapid development of megacities, and their growing connectedness across the world is becoming a distinct driver for emerging disease outbreaks. Early detection of unusual disease emergence and spread should therefore include such cities as part of risk-based surveillance. A catch-all metagenomic sequencing approach of urban sewage could potentially provide an unbiased insight into the dynamics of viral pathogens circulating in a community irrespective of access to care, a potential which already has been proven for the surveillance of poliovirus. Here, we present a detailed characterization of sewage viromes from a snapshot of 81 high density urban areas across the globe, including in-depth assessment of potential biases, as a proof of concept for catch-all viral pathogen surveillance. We show the ability to detect a wide range of viruses and geographical and seasonal differences for specific viral groups. Our findings offer a cross-sectional baseline for further research in viral surveillance from urban sewage samples and place previous studies in a global perspective.

3.2 Introduction

The increasing connectivity of the modern world, changing demographics, and climate change increase the potential for novel and known viral pathogens to emerge and rapidly spread in new and unexpected areas, as could be seen during the emergence and global threat of Ebola virus in recent outbreaks[120]. Early detection or ruling out of high impact (emerging) infections as causes of disease is a hallmark of preparedness, but research in response to recent outbreaks of Ebola, Zika and yellow fever has shown that these pathogens circulated for extended periods of time before being recognized, leading to costly delays in public health response[121–124]. One of the key challenges is how to prioritize local investments in detection capacity, given the diversity of emerging diseases, the unpredictable nature of outbreaks, and the limited resources available for outbreak preparedness. Understandably, surveillance of infectious diseases mainly targets common conditions and is scaled up in response to the emergence of pathogens and in particular disease outbreaks, rather than the costlier approach of broad range testing for any relevant infectious disease. The changing dynamics of infectious diseases related to global change, however, require rethinking of this model for public health preparedness, as incidence-based surveillance provides a fragmented and limited scope of which pathogens are circulating in the general population, particularly in low resource settings where access to healthcare and laboratory diagnostics is restricted[125,126]. Therefore, in its reorganization in response to the West African Ebola outbreak, the World Health Organization has launched the term “Disease X” to call for novel ideas for preparedness to unpredictable disease outbreaks[127]. Thus, there is a need for novel approaches to viral surveillance providing a broader and less biased insight into the circulation of viral pathogens to supplement the more targeted surveillance. Genomic epidemiology using real-time pathogen sequencing has become part of the routine toolbox for outbreak tracking once the cause of the outbreak is known[128,129]. In addition, metagenomic sequencing has been put forward as a potential catch-all surveillance tool, but the step from research to routine implementation is extremely challenging[130,131], and thus, careful validation is needed to avoid overpromise and wasting of resources.

Here, we set out to explore the potential use of metagenomic sequencing of urban sewage as an add-on strategy for global disease preparedness. One key driver of emergence is the amplification of rare zoonotic and vector-borne diseases in densely populated regions where infrastructure needs are outpaced by rapid urban developments. This leads to the formation of slums, favorable conditions for viral disease vectors, disparity in access to clean water, sanitation and healthcare, and an increase in close human-animal interaction due to deforestation[132,133]. The advantage of using sewage-based surveillance is that it represents the entire population of the catchment area, sample collection is straightforward, and the anonymization by default makes it less challenging to use than patient-based surveillance regarding privacy laws. Using sewage to detect viruses with low case fatality rate but overall high population level impact has been tested successfully to monitor the progress of the global polio-elimination program, particularly in regions where non-replicating polio-virus vaccines are used[134,135]. The huge potential of environmental surveillance was illustrated when a silent epidemic of wild-poliovirus type 1 in Israel was detected, which led to a mop-up vaccination campaign and resolution of the epidemic, without a single case of paralytic poliomyelitis[136]. In addition, small-scale studies have already shown the potential for using metagenomic sequencing of sewage extracts for the detection of a range of virus families[83,84,86] (**Extended Data Table 1**). While these studies have largely focused on viruses with a replication phase in the gastro-intestinal tract, the fecal and/or urinary shedding of, for instance, measles virus, yellow fever virus, Zika virus, West Nile virus, Ebola virus, SARS coronavirus, and MERS coronavirus suggests the potential utility of sewage testing to capture circulation of these pathogens as well[137–141]. Moreover, metagenomic sequencing has the potential to detect any viral genomic material in the sample, without targeting a specific viral pathogen or limiting for only known viral pathogens. In this study, we pilot the use of metagenomics to describe a comparative snapshot of the virome from sewage samples of high-density urban areas across all continents. We provide a critical appraisal of technical and analytical biases and discuss the potential utility for human and animal disease monitoring and surveillance, as well as the additional steps needed to go towards routine implementation.

3.3 Results

3.3.1 Data quality evaluation

Urban sewage samples and associated metadata (**Supp. File 1**, available in online version) were obtained from 62 countries across all continents between January and April 2016 from the influent of wastewater treatment plants prior to treatment or from open sewage systems in low- and middle-income countries. All samples were previously processed for the detection of bacterial antimicrobial resistance genes using DNA metagenomics[142]. Here we focus solely on viral DNA and RNA metagenomics (methods) and the analysis of the viral data. Sewage samples are highly variable in terms of composition and DNA abundance and therefore potential biases that might impact the final read abundance and diversity of the sewage virome were evaluated.

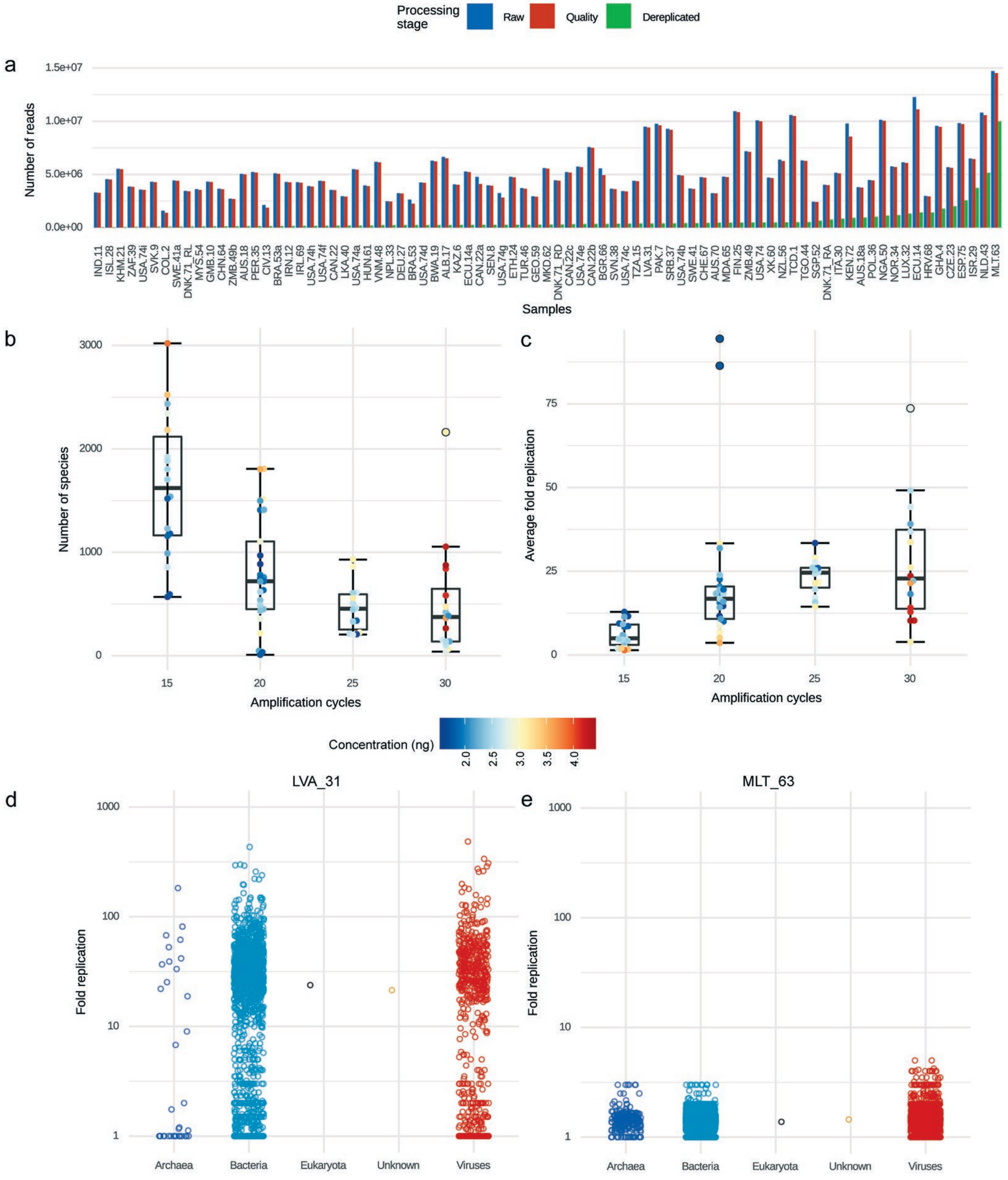


Figure 1 Effect of read preprocessing on data interpretation. (a) Number of reads before preprocessing (blue bars) after quality control (red bars) and read dereplication (green bars). The x axis shows sample identifiers ordered by number of dereplicated reads. (b and c) Effect of number of PCR replication cycles on library concentration (color), species diversity (b) and read replication rate (c).(d and e) Fold replication of raw reads by species level annotation (points). X axis separates superkingdom or "Unknown" annotations. d shows sample LVA_31 with a high replication rate and panel e shows sample MLT_63 with a low replication rate.

Initially, an extensive evaluation of the technical factors that may impact the resulting data to gain a deeper understanding of potential pitfalls was performed. First, read abundance was evaluated as a proxy for viral abundance. Sequencing protocols for virome analysis in sewage typically require an amplification step to provide enough DNA input for sequencing, which can result in artificial duplication of sequence reads and thereby impact the quantitative interpretation of the data substantially (**Figure 1a**). Indeed, the observed viral species richness was negatively correlated with the number of amplification cycles needed to obtain enough DNA as input for sequencing (**Figure 1b**), while the average fold replication of a read was

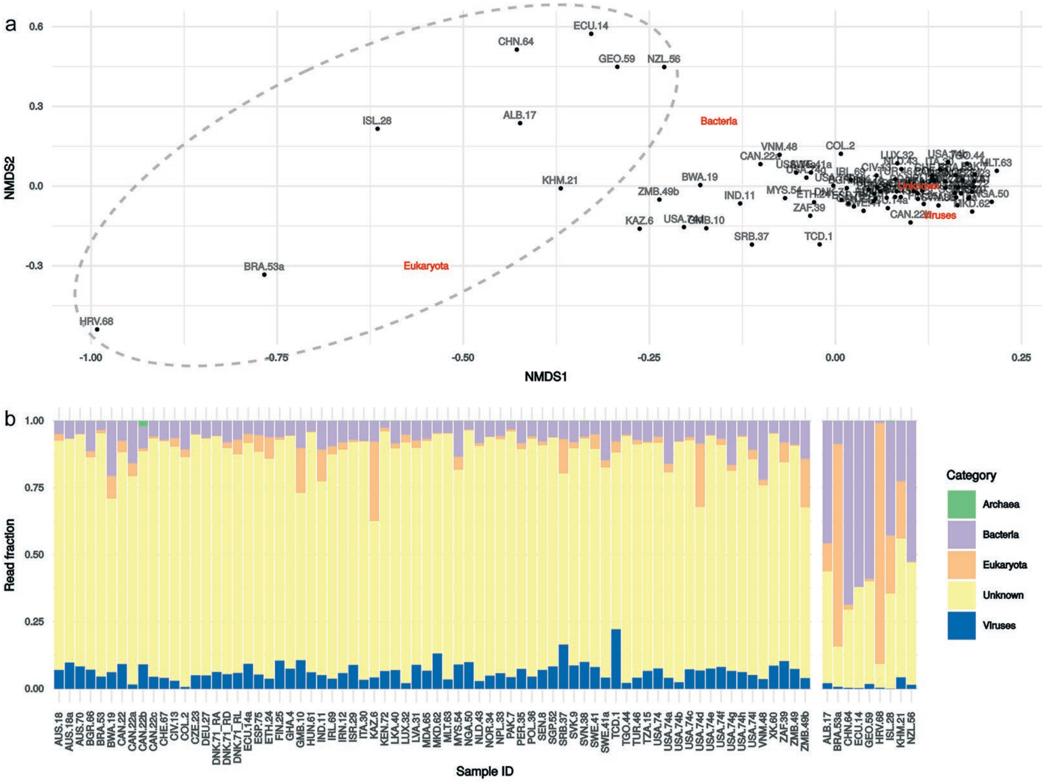


Figure 2 Effect of non-viral background read abundances on viral read abundance and the chosen outlier samples in the sewage metagenome data. (a) a multidimensional scaling of Bray-Curtis dissimilarity between samples based on the normalized read counts of bacterial, archaeal, eukaryote (human), and viral content. "Unknown" indicates reads that could not be assigned any annotation. The red labels indicate the effect of the different annotations on the position of a sample in the plot. Gray circle indicates the samples that were manually assigned to be outliers. (b) A scaled bar chart of relative read abundance showing the outliers in a separate facet to the right.

positively correlated (**Figure 1c**). The impact of dereplication on the individual species level read counts varied greatly within a sample. Especially in samples with a low number of reads after dereplication (**Figure 1d**) the decrease in read counts for a species ranges from 600 to 5-fold. These differences have a profound effect on the species distribution within the sample, and thus the interpretation thereof. The effect of dereplication is much less variable between species in samples with a high number of reads after dereplication (**Figure 1e**). Therefore, the optimal use of virome sequencing depends on the initial abundance of viral sequences in the sample and extra amplification may only increase the coverage of the same viruses, but does not increase the richness of the virome, which needs to be carefully considered when designing and interpreting sewage metagenomics studies.

Besides the influence of read replication on read abundance, the richness of the virome can be impacted by the presence of non-viral sequences. Typically, the metagenomic data contain a large fraction of unknown reads, and, despite the virus specific sample preparation, non-viral reads, including archaeal, bacterial, and eukaryote DNA.

While the overall proportion of reads for the different domains was comparable in most samples, multidimensional scaling of the non-viral read counts showed that some samples were very divergent from the central cluster and were manually marked as outliers (**Figure 2a**, dashed line). Viral read abundance was low in these outlier samples (**Figure 2b**, right panel). There was no significant correlation between the concentration of human or bacterial read fractions with any of the measured sample characteristics, such as pH, conductivity, and type of sewer system.

3.3.2 Exploration of the sewage virome

Based on the data quality assessment, we analyzed viral diversity in the samples after dereplication and following annotation by both Kajiu and Centrifuge as described. Between 0.09% and 22% of the reads could be annotated as viral (median of 6%), with high abundances of bacteriophages, plant- and insect viruses (**Figure 3**). Most abundant were bacteriophages, representing on average 77% (ranging from 9 to 94%) of the annotated viral reads in the sewage. In particular *Microviridae* (median of 18%, range 0.5% to 51% of reads), *Siphoviridae* (median of 17%, range 0.22% to 67% of reads), *Myoviridae* (median of 9%, range 0.08% to 41% of reads), and *Podoviridae* (median of 4%, range 0.02% to 25% of reads), were highly abundant. These bacteriophage families could be found around the globe without obvious regional differences when using read annotations at this taxonomic level. Although specific

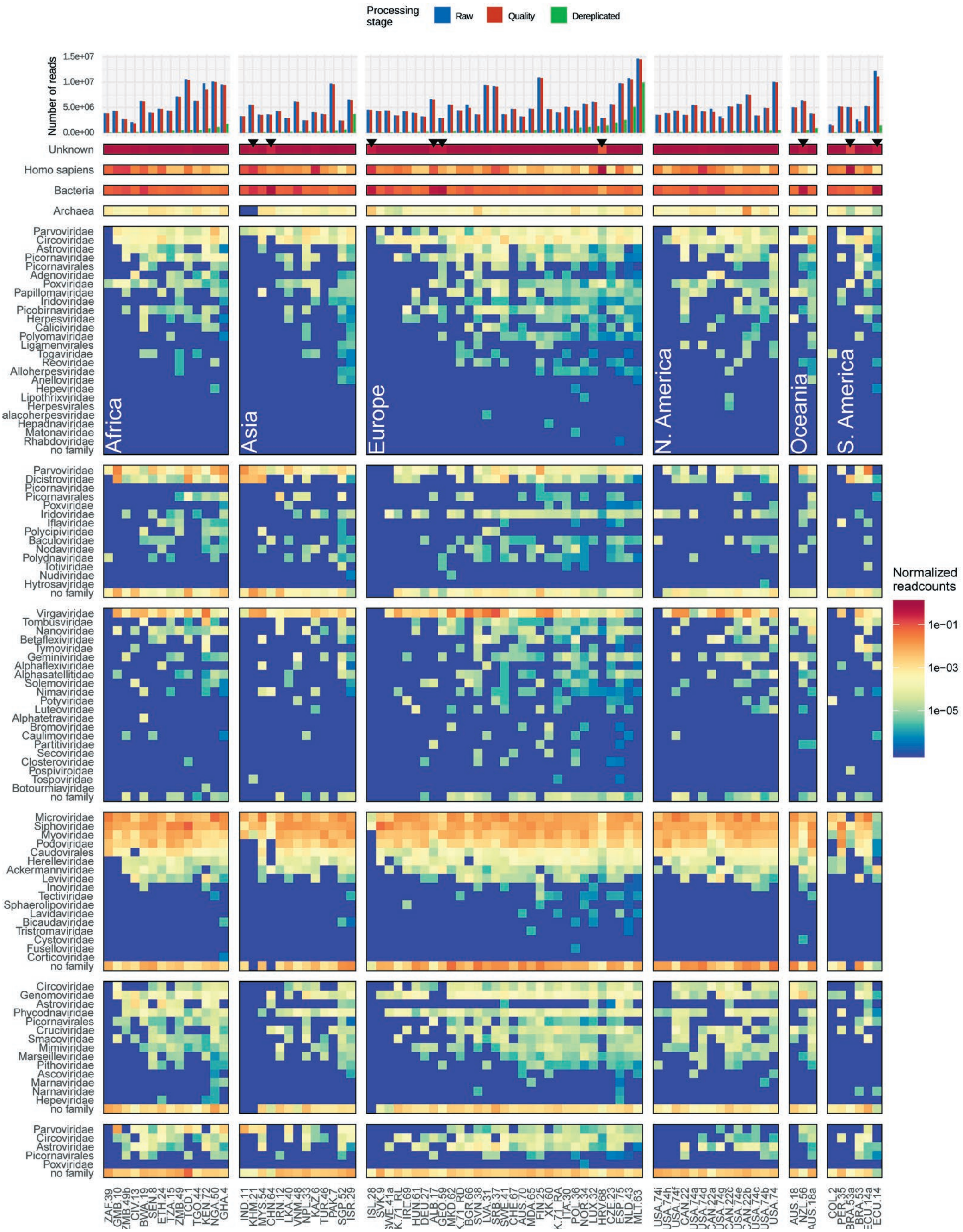

Processing stage
Raw
Quality
Dereplicated
Number of reads
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Unknown
Homo sapiens
Bacteria
Archaea
Parvoviridae
Circoviridae
Astroviridae
Picornaviridae
Picornavirales
Adenoviridae
Poxviridae
Papillomaviridae
Iridoviridae
Picobirnaviridae
Herpesviridae
Caliciviridae
Polyomaviridae
Ligamenvirales
Togaviridae
Reoviridae
Alloherpesviridae
Anelloviridae
Hepeviridae
Lipothrixviridae
Herpesvirales
alacoherpesviridae
Hepadnaviridae
Matonaviridae
Rhabdoviridae
no family
Africa
Asia
Europe
N. America
Oceania
S. America
Parvoviridae
Dicistroviridae
Picornaviridae
Picornavirales
Poxviridae
Iridoviridae
Iflaviridae
Polycipiviridae
Baculoviridae
Nodaviridae
Polydnaviridae
Totiviridae
Nudiviridae
Hytrosaviridae
no family
Virgaviridae
Tombusviridae
Nanoviridae
Betaflexiviridae
Tymoviridae
Geminiviridae
Alphaflexiviridae
Alphasatellitidae
Solemoviridae
Nimaviridae
Potyviridae
Luteoviridae
Alphatetraviridae
Bromoviridae
Caulimoviridae
Partitiviridae
Secoviridae
Closteroviridae
Pospiviroidae
Tospoviridae
Botourmiaviridae
no family
Microviridae
Siphoviridae
Myoviridae
Podoviridae
Caudovirales
Herelleviridae
Ackermannviridae
Leviviridae
Inoviridae
Tectiviridae
Sphaerolipoviridae
Lavidaviridae
Bicaudaviridae
Tristromaviridae
Cystoviridae
Fuselloviridae
Corticoviridae
no family
Circoviridae
Genomoviridae
Astroviridae
Phycodnaviridae
Picornavirales
Cruciviridae
Smacoviridae
Mimiviridae
Marseilleviridae
Pithoviridae
Ascoviridae
Marnaviridae
Narnaviridae
Hepeviridae
no family
Parvoviridae
Circoviridae
Astroviridae
Picornavirales
Poxviridae
no family
Normalized readcounts
1e−01
1e−03
1e−05
ZAF.39
GMB.10
ZMB.49b
CIV.13
BWA.19
SEN.8
ETH.24
TZA.15
ZMB.49
TCD.1
TGO.44
KEN.72
NGA.50
GHA.4
IND.11
KHM.21
MYS.54
CHN.64
IRN.12
LKA.40
VNM.48
NPL.33
KAZ.6
TUR.46
PAK.7
SGP.52
ISR.29
ISL.28
SVK.9
SWE.41a
K.71_RL
IRL.69
HUN.61
DEU.27
ALB.17
GEO.59
MKD.62
K.71_RD
BGR.66
SVN.38
LVA.31
SRB.37
SWE.41
CHE.67
AUS.70
MDA.65
FIN.25
XK.60
K.71_RA
ITA.30
POL.36
NOR.34
LUX.32
HRV.68
CZE.23
ESP.75
NLD.43
MLT.63
USA.74i
USA.74h
USA.74f
CAN.22
USA.74a
USA.74d
CAN.22a
USA.74g
CAN.22c
USA.74e
CAN.22b
USA.74c
USA.74b
USA.74
AUS.18
NZL.56
AUS.18a
COL.2
PER.35
BRA.53a
BRA.53
ECU.14a
ECU.14

Figure 3 Heatmap of the viral diversity at viral family level (when available) and non-viral fraction. The read abundance after quality control and dereplication is shown ordered by total read abundance after preprocessing and facetted by continent. The heatmap follows the same ordering. Color gradient represents log-transformed relative abundance of reads belonging to the taxonomic groups indicated. The top four rows of the heatmap show read abundances of non-viral annotations, the other rows show read abundance by viral family, or “no family” if only genus or species level annotation was available. Vertical facets represent subdivision of the viral families based on their inferred host. Black arrows indicate outlier samples based on an overabundance of background sequences.

bacteriophages have been studied extensively as potential indicators of human fecal pollution, bacteriophage taxonomy is relatively poorly defined, making accurate classification challenging at genus and species level[143,144]. Hence, geographical patterns at a more fine-grained level of annotation may be lost in our analysis. Moreover, interpretation of patterns of bacteriophage abundance could be obscured by the fact that bacteriophages can encounter bacterial hosts in the sewage in which they can multiply. As described elsewhere, the analysis of the bacterial resistomes of the same samples showed clear segregation of sequences from Africa and Asia versus those from Europe and the US[142]. A more detailed analysis is needed to assess if there is a relation between specific bacteriophages and the resistomes, as environmental viromes have been shown to be a potential reservoir for antimicrobial resistance genes[145].

3.3.3 Global patterns of viruses related to vegetable consumption and to insects detected in urban sewage

The second largest fraction of the virome (0.02% to 69%, median 3.4%) consisted of plant-related viruses. On average, more than 84% of these reads belonged to the *Virgaviridae* family. Especially viral species related to infections of cucumber, tomato, tobacco and pepper plants could be detected in sewage, as indicated by species level taxonomy (**Figure 4b**). Apart from a sample from Kenya, the abundance of vegetable-consumption-related viruses was higher in samples from Europe and North America compared to samples from the rest of the world (**Figure 4a**) (Welch’s t-test, p-value=0.06). The global presence and high abundance of plant viruses has led to the proposal that they may be good indicators for human fecal contamination alike specific bacteriophage populations[146]. However, this remains to be validated given the geographic variation observed in our dataset, which could reflect differences in diet and/or agricultural practices in these countries.

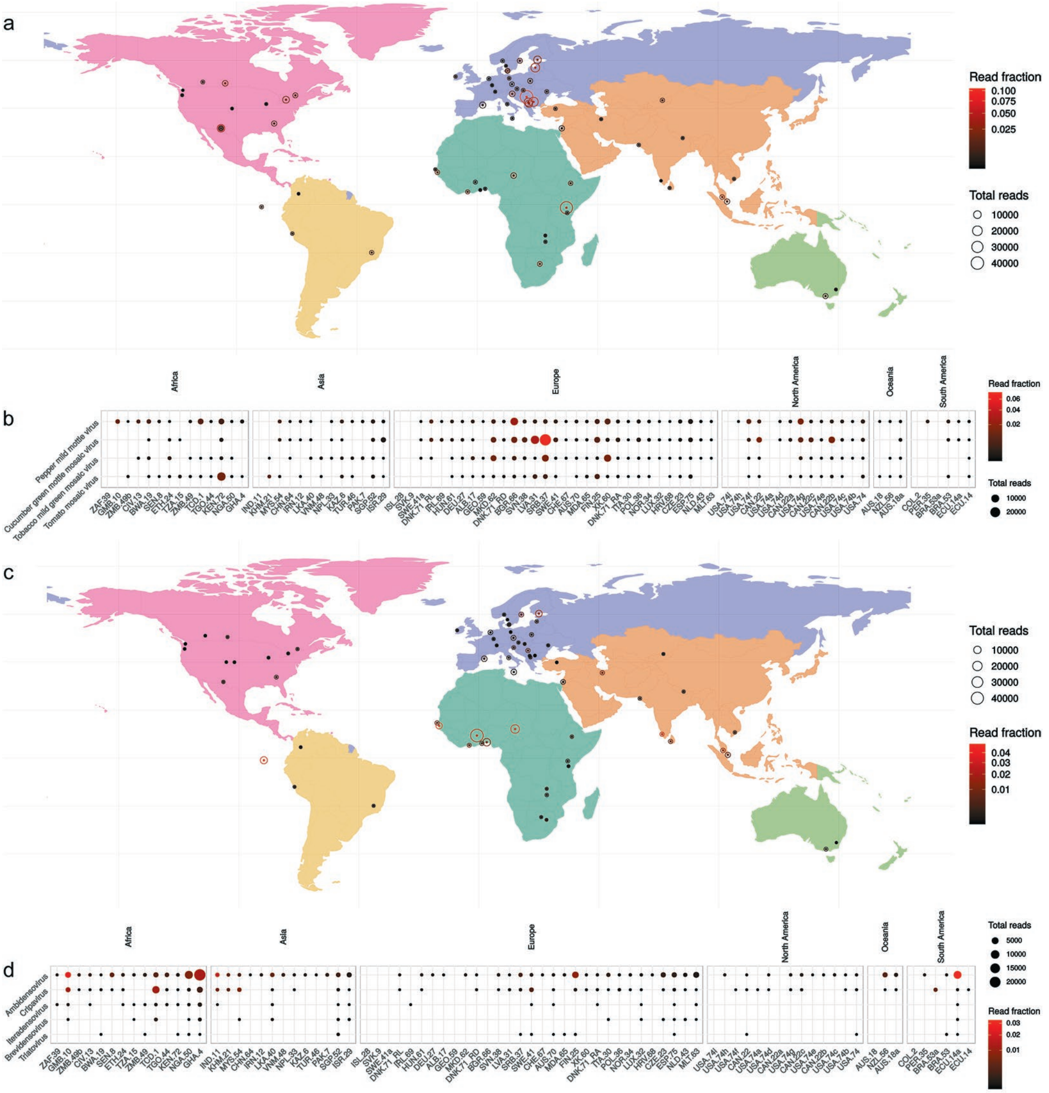


Figure 4 Overview of the global distribution and abundance of plant viruses and insect viruses in urban sewage. (a) Global distribution of all plant viruses (b) The four most abundant plant virus species and their global spread. (c) Global distribution of all insect related viruses. (d) Top 5 most abundant insect virus genera. Datapoints represent absolute read numbers and read fraction by varying size and color respectively. Viral species are ordered by summed read abundance across samples and samples are ordered by total read abundance from left to right. Facets represent continent of sample origin.

A median of 1.4% (ranging 0.1% to 74%) of the sewage virome consisted of viruses associated with insects, comprising mainly species from the genera *Ambidensovirus, Cripavirus,* and *Brevidensovirus* (**Figure 4d**), known to infect a range of crickets, cockroaches, fruit flies, and mosquitos[147]. In the global distribution there was an increased abundance of insect viruses in samples from around the equator, mainly in samples from Africa (**Figure 4c**) (Welch's t-test, p-value=0.0004). One exception was the sample from Finland, which had a high abundance of insect virus reads (13.7%) in comparison with samples from other European countries (1.5%). Several reads were found to be annotated as "Aedes albopictus densovirus 2", "Aedes aegypti Thai densovirus", and "Anopheles gambiae densonucleosis virus". There is some evidence that these densoviruses may be associated with *Aedes aegypti*, *Aedes albopictus* and *Culex* mosquitos[148,149]. Current data are not sufficient to meet the requirements for sewage surveillance, but these findings show the potential to track mosquitos by looking for mosquito specific viruses.

3.3.4 Detection of vertebrate viruses and investigation of known human pathogens

About 1.7% (ranging 0.01% to 11%) of the virome consisted of vertebrate viruses. Most abundant were small ssDNA viruses from the families *Circoviridae* and *Parvoviridae,* and members of the *Picornaviridae, Astroviridae* and *Adenoviridae* families (**Figure 5a**). Vertebrate viruses were detected widely across the samples, but did not show distinct geographical patterns of abundance. Circoviruses were especially highly abundant across most sewage samples and, as novel variants of circoviruses have been associated with several diseases in pigs[150]. Further longitudinal sewage surveillance could potentially be used to detect epidemiological patterns of emerging circovirus variants.

A selection of viral taxa was analyzed containing human pathogenic viruses from the *Astro-*, *Entero-*, *Noro-*, *Sapo-*, *Adeno-* and *Rotaviridae* families that are known to be abundant across the world as causes of diarrheal disease (**Figure 5b**). Most abundant and widespread were the astroviruses. Enteroviruses were present to a lesser extent but could be detected in sewage samples from across the globe as well. Members of the noro-, sapo-, adeno-, and rotaviruses were only sporadically detected. Further investigation of samples with high human astrovirus content showed mostly evidence of the classic Human Astrovirus 1, 2 and 4 that are common causes of diarrheal disease, and sporadic detection of other clades such as Human Astrovirus MLB and Human Astrovirus VA for which less is known regarding clinical impact[151]. Mapping of human enterovirus reads resulted in 102 small contiguous sequences which were typed using

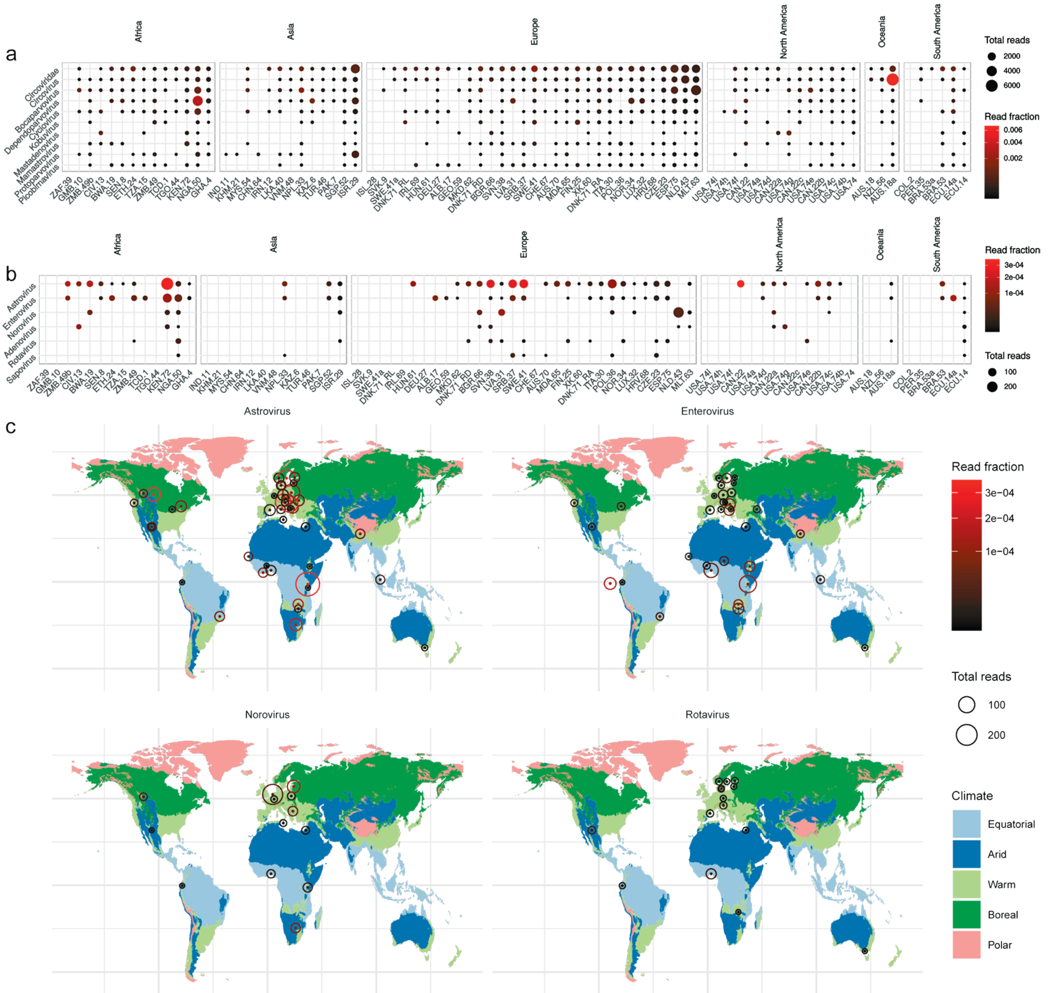


Figure 5 Overview of the most abundant vertebrate viruses and specific human viruses and their distribution worldwide in urban sewage. (a) Distribution of the top ten most abundant vertebrate viral families. (b) Relative abundance of viruses encountered in clinical surveillance. (c) World maps showing distribution of viruses encountered in clinical surveillance. Coloring of the maps delineates differences in climate by geographical location. Datapoints represent absolute read numbers and read fraction by varying size and color respectively. Viral families are ordered by summed read abundance across samples and samples are ordered by total read abundance from left to right. Facets represent continent of sample origin.

the enterovirus typing tool[28]. Mainly Enterovirus C (46%) and B (9%) were detected. Further subtyping of for instance poliovirus was not possible because of a lack of coverage of the standardized genotyping region VP1. The same mapping was done for norovirus, resulting in 13 contigs of 84 to 962 nucleotides in length. Most norovirus sequences were typed as either GII, with capsid type 6, 10 and 17, and GIV, all viruses that are commonly found in outbreak based surveillance[152]. Sapovirus sequences, all belonging to type GI, were found in seven of the samples. Adenovirus and rotavirus hits were sporadically detected across all sampling sites and upon further investigation showed mainly adenovirus C and rotavirus A hits.

It is known that noroviruses, astroviruses and rotaviruses follow a winter seasonality and enteroviruses follows a summer seasonality pattern[153–155]. The time of sampling of the sewage was in a 3-month timeframe between January and March, which corresponds to the winter period in the northern hemisphere, therefore a higher prevalence of winter seasonal viruses was expected in those. When looking at the global distribution of viruses, the average abundance of astro- and noroviruses was higher in the northern hemisphere, and the reverse pattern was observed for enteroviruses, with higher average abundance in the southern hemisphere during the sampling period (**Figure 5c**). Given the cross-sectional nature of our study we acknowledge that these seasonal patterns will have to be confirmed using longitudinal sampling which would allow for meaningful statistical analysis, but our first observations align with what is generally expected at that time of the year.

3.4 Discussion

This global sewage study gives, for the first time, a catch-all metagenomic comparison of the urban sewage virome of major cities across the world. We show that it is possible to detect a wide diversity of viruses in sewage samples and we identify geographical and seasonal differences in abundance for specific viral groups, including those that are currently targeted by surveillance for diarrheal and neurological disease, as well as viruses that could be used as indicators for presence of specific mosquito species. In addition, we provide the global scientific community with a geographically very broad resource for searching for novel virus sequences as novel pathogens continue to emerge. The pilot study also highlights some important challenges that need to be addressed to take the technology forward, such as how to deal with low input samples and the overabundance of phages, plant, and insect viruses in the sample. Metagenomic sequencing of viruses is a complex and evolving technology which is currently far from being standardized. Differences in sample preprocessing, sequencing technology, and data analysis can have a major impact on the viral read abundance, diversity, and the proportion of sequences that are annotated[35,156]. In our study, we eliminated lab-to-lab variability by performing all sample preparation, sequencing and analysis at the same location, which, apart from the analysis, is obviously not feasible for global surveillance. Further work is ongoing, including the development of fieldable sample treatment and sequencing protocols, comparison of effects of sample preparation on viral richness and further exploration of applicability, by longitudinal sampling and sampling in the presence of known ongoing outbreaks.

A critical challenge of using metagenomic sequencing for surveillance purposes remains the interpretation of sequence annotations. With the development of high-speed k-mer based annotation tools such as the ones used in this study, annotation can be performed rapidly and with few false negatives. However, erroneous and mis-annotated entries in public databases, together with inconsistency in the sequence-based taxonomic classification of viruses, make annotation to the species level challenging. Major steps have been taken to create a more consistent sequence based viral taxonomy[143,157], but these approaches have not yet been integrated in fast viral annotation tools. Also, deposits of large volumes of virus sequences without a clear host association or pathogenicity data in public databases[158] make it difficult to interpret the relevance of such findings. In our data, many of these “environmental viruses” could be identified. Given the increase in virus diversity in reference databases, it is striking how many sequence reads can remain unclassified with the currently used methods. This is in line with previous observations, where 40-90% of the sequence reads could not be classified[159]. It can very well be that the currently unclassified sequence reads represent potential new viruses, including novel pathogens.

In conclusion, we show the potential of global viral surveillance using metagenomic sequencing of sewage without ignoring the complexity of the approach. However, with improvements in sample preprocessing, sequencing methods and interpretability of viral sequence annotation this potential will increase.

3.5 Methods

3.5.1 Urban sewage sample and metadata collection

Samples were obtained from 62 countries from all continents as previously described[142]. All samples were taken before wastewater treatment. A questionnaire was filled in with information on sampling site, sample consistency and sample temperature, including transport time, storage time, and temperature before shipping. All samples were taken in a timeframe of 3 months from January until March 2016. In addition to sample specific data, additional metadata (Supp. File 1) was collected such as demographics, type of industry in the surrounding area, weather conditions and catchment area of the sewer. Upon arrival, samples were thawed at room temperature and 250 ml of the raw sewage was taken and centrifuged at 10,000 g for 10 minutes. The pellet was removed for bacterial content determination and DNA metagenomic sequencing[142] and the supernatant was used to perform the virus specific sample pretreatment and sequencing.

3.5.2 Sample processing for sequencing

Viral extraction was performed on 40 ml of sewage supernatant as previously described[160]. In short, the conductivity was measured to exceed 2000 µs and the pH of the samples was adjusted to pH 4. Afterwards 10 ml PEG 6000 was added and the samples were incubated overnight at 4˚C under agitation.

After incubation the samples were centrifuged a 13,500 g for 1.5 hours at 4˚C. The supernatant was removed, the pellet was dissolved in warm glycine buffer and 1 mL of chloroform-butanol (50/50) was added. After mixing, the sample was centrifuged for 5 minutes at 13,000 g at 4˚C. The filtrate was collected through a series of filters with 5 µm, 1.2 µm, 0.45 µm and 0.22 µm pore size.

Unprotected free DNA was removed by incubation with Ambion Turbo DNase for 30 minutes at 37˚C. Total nucleic acid content was extracted using Roche NA isolation kit and cDNA was made using superscript III (Invitrogen) using random hexamers that avoids amplification of human rRNA[161]. dsDNA was made using Klenow (NEB) and samples were sheared using Ion Shear Plus Enzyme Mix II. Libraries were amplified for 15 cycles using High Fidelity Platinum PCR reaction. The library concentration was determined using Ion Torrent quantification kit (Thermo Fisher). If the concentration was below 20 nM, extra amplification cycles were performed. Sequencing was performed on the Ion Torrent S5XL platform to generate around 10 million sequence reads per sample.

3.5.3 Data preprocessing

Raw fastq files were quality trimmed using FastP[162]. Read ends were trimmed to mean quality 25 with a sliding window of 5. Reads were trimmed to 400 nucleotides by default because the chemistry of Ion Torrent sequencing technology allows for reads of maximally 400 nucleotides long and longer reads were observed to contain high Phred score non-sense repetitive patterns in the tail region. Reads shorter than 50 nucleotides were discarded as well as reads with an average Phred score below 25. Duplicate reads were removed using CD-HIT[163] by clustering reads that start at the exact same position in the genome and have over 90% sequence identity in the first 50 nucleotides of the read, because of variable read length and observed insertion and deletion errors in the beginning of the reads.

3.5.4 Read based analysis

Due to the expected high diversity of viruses present in the sewage samples, a read based annotation of the data was chosen, contrary to an assembly-based approach. Annotation was performed using two taxonomic annotation tools: Kaiju[108] and Centrifuge[164]. Kaiju performs taxonomic annotation based on an amino acid (AA) level which provides a higher sensitivity. This is especially important for the annotation of viral sequences given the high mutation rate of viruses[165] compared to other organisms. In parallel with Kaiju, Centrifuge was run, which uses nucleotide (nt) identity for taxonomic annotation. Combining a nucleotide and an amino acid based matching approach ensures that both coding and non-coding read sequences can be annotated. In addition, the combination of two read annotation tools with different annotation strategies was chosen to give more robust mapping results.

The databases used for taxonomic annotation consisted of archaeal, bacterial and human RefSeq sequences and were extended with all viral and phage entries in GenBank version 230[166] because of the limited viral and phage sequence diversity in the RefSeq database.

Recommended quality thresholds and parameters for metagenomic data were used for both Kaiju and Centrifuge. Kaiju was run in greedy mode with a score cutoff of 70 and an error of 5. Centrifuge was run with a score threshold of 300 and a hit length cutoff of 50. If neither method produced a hit the read was annotated as “Unknown”. BASTA[167] was used to determine the last common ancestor (LCA) of each hit given by both methods without restrictions on hit quality.

The final read counts passing QC were determined by the sum of read annotations at a certain taxonomic level and were normalized by total dereplicated read count to adjust for differences in sequencing depth and data quality[104,168,169]. The LCA taxon was used if the annotation at a certain taxonomic level was absent. Manual regrouping of taxonomic levels was performed to calculate read counts of human pathogenic viruses and read counts by host group. For sample comparison, read counts were normalized by Hellinger transformation[170]. Sample-wise comparison was done by calculating the Bray-Curtis dissimilarity between the normalized read counts using the R package Vegan[171]. Further investigation of the annotation of specific viral species was performed by mapping the reads against a redundant set of reference genomes using KMA with default parameters[172]. The maps of global read distribution were created using the continent subdivision from the "rnaturalearthdata" R package and the Köppen-Geiger climate classification[173].

Data Availability

Raw sequence data that support the findings of this study have been deposited in the European Nucleotide Archive with the study accession code PRJEB23496.

Author affiliations

[1] Viroscience Department, Erasmus Medical Center, The Netherlands

[2] National Food Institute, Technical University of Denmark, Denmark

*: These authors contributed equally to this work

The Global Sewage Surveillance project consortium author list

Rene S. Hendriksen[2], Artan Bego[3], Catherine Rees[4], Elizabeth Heather Neilson[5], Kris Coventry[6], Peter Collignon[7], Franz Allerberger[8], Teddie O. Rahube[9], Guilherme Oliveira[10], Ivan Ivanov[11], Thet Sopheak[12], Yith Vuthy[12], Christopher K. Yost[13], Djim-adjim Tabo[14], Sara Cuadros-Orellana[15], Changwen Ke[16], Huanying Zheng[16], Li Baisheng[16], Xiaoyang Jiao[17], Pilar Donado-Godoy[18], Kalpy Julien Coulibaly[19], Jasna Hrenovic[20], Matijana Jergović[21], Renáta Karpíšková[22], Bodil Elsborg[23], Mengistu Legesse[24], Tadesse Eguale[24], Annamari Heikinheimo[25], Jose Eduardo Villacis[26], Bakary Sanneh[27], Lile Malania[28], Andreas Nitsche[29], Annika Brinkmann[29], Courage Kosi Setsoafia Saba[30], Bela Kocsis[31], Norbert Solymosi[32], Thorunn R. Thorsteinsdottir[33], Abdulla Mohamed Hatha[34], Masoud Alebouyeh[35], Dearbhaile Morris[36], Louise O'Connor[36], Martin Cormican[36], Jacob Moran-Gilad[37], Antonio Battisti[38], Patricia Alba[38], Zeinegul Shakenova[39], Ciira Kiiyukia[40], Eric Ng'eno[41], Lul Raka[42], Aivars Bērziņš[43], Jeļena Avsejenko[44], Vadims Bartkevics[44], Christian

Penny[45], Heraa Rajandas[46], Sivachandran Parimannan[46], Malcolm Vella Haber[47], Pushkar Pal[48], Heike Schmitt[49], Mark van Passel[49], Milou G.M. van de Schans[50], Tina Zuidema[50], Gert-Jan Jeunen[51], Neil Gemmell [51], Kayode Fashae[52], Astrid Louise Wester[53], Rune Holmstad[54], Rumina Hasan[55], Sadia Shakoor[55], Maria Luz Zamudio Rojas[56], Dariusz Wasyl[57], Golubinka Bosevska[58], Mihail Kochubovski[58], Cojocaru Radu[59], Amy Gassama†[60], Vladimir Radosavljevic[61], Moon Y.F. Tay[62], Rogelio Zuniga-Montanez[63], Stefan Wuertz[63], Dagmar Gavačová[64], Marija Trkov[65], Karen Keddy[66], Kerneels Esterhuyse[67], Marta Cerdà-Cuéllar[68], Sujatha Pathirage[69], D. G. Joakim Larsson[70], Leif Norrgren[71], Stefan Örn[71], Tanja Van der Heijden[72], Happiness Houka Kumburu[73], Ana Maria de Roda Husman[74], Berthe-Marie Njanpop-Lafourcade[75], Pawou Bidjada[76], Somtinda Christelle Nikiema-Pessinaba[77], Belkis Levent[78], John Scott Meschke[79], Nicola Koren Beck[79], Chinh Dang Van[80], Doan Minh Nguyen Tran[80], Nguyen Do Phuc[80], Geoffrey Kwenda[81]

[3] Institute of Public Health, Tirana, Albania

[4] Melbourne Water Corporation, Docklands, Australia

[5] University of Copenhagen, Frederiksberg C, Australia

[6] Applied Research, Docklands, Australia

[7] Canberra Hospital, Canberra, Australia

[8] Austrian Agency for Health and Food Safety (AGES), Vienna, Austria

[9] Botswana International University of Science and Technology , Palapye, Botswana

[10] Vale Institute of Technology, Sustainable Development, Belém, Brazil

[11] National Center of Infectious and Parasitic Diseases, Sofia, Bulgaria

[12] Institut Pasteur du Cambodge, Phnom Penh, Cambodia

[13] University of Regina, Regina, Canada

[14] University of N'Djamena, N'Djamena, Chad

[15] Centro de Biotecnología de los Recursos Naturales, Universidad Católica del Maule, Talca, Chile

[16] Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China

[17] Shantou University Medical College, Shantou, China

[18] Corporacion Colombiana de Investigacion Agropecuaria (AGROSAVIA), Mosquera, Colombia

[19] Institut Pasteur de Côte d'Ivoire, Abidjan, Côte d'Ivoire

[20] University of Zagreb, Faculty of Science, Zagreb, Croatia

[21] Andrija Stampar Teaching Institute of Public Health, Zagreb, Croatia

[22] Veterinary Research Institute, Brno, Czech Republic

[23] Renseanlæg Lynetten, København K, Denmark

[24] Addis Ababa University , Addis Ababa, Ethiopia

[25] University of Helsinki, Helsinki, Finland

[26] Instituto Nacional de Investigación en Salud Pública-INSPI (CRNRAM), Quito, Galápagos, Ecuador

[27] National Public Health Laboratories, Ministry of Health and Social Welfare, Kotu Layout, Kotu, Gambia

[28] National Center for Disease Control and Public Health, Tbilisi, Georgia

[29] Robert Koch Institute, Berlin, Germany

[30] University for Development Studies, Tamale, Ghana

[31] Semmelweis University, Institute of Medical Microbiology, Budapest, Hungary

[32] University of Veterinary Medicine, Budapest, Budapest, Hungary

[33] Institute for Experimental Pathology, University of Iceland, Keldur, Reykjavík, Iceland

[34] Cochin University of Science and Technology, Cochin, India

[35] Pediatric Infections Research Center, Research Institute for Children's Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[36] National University of Ireland Galway, Galway, Ireland

[37] School of Public Health, Ben Gurion University of the Negev and Ministry of Health, Beer-Sheva, Israel

[38] Istituto Zooprofilattico Sperimentale del Lazio e della Toscana, Rome, Italy

[39] National Center of Expertise, Taldykorgan, Kazakhstan

[40] Mount Kenya University, Thika, Kenya

[41] Kenya Medical Research Institute, Nairobi, Kenya

[42] University of Prishtina "Hasan Prishtina" & National Institute of Public Health of Kosovo, Pristina, Kosovo

[43] Institute of Food Safety, Animal Health and Environment "BIOR", Riga, Latvia

[44] Institute of Food Safety, Riga, Latvia

[45] Luxembourg Institute of Science and Technology, Belvaux, Luxembourg

[46] Centre of Excellence for Omics-Driven Computational Biodiscovery, Faculty of Applied Sciences, AIMST University, Kedah, Malaysia

[47] Environmental Health Directorate, St.Venera, Malta

[48] Agriculture and Forestry University, Kathmandu, Nepal

[49] National Institute for Public, Health and the Environment (RIVM), Bilthoven, Netherlands

[50] Wageningen Food Safety Research, Wageningen, Netherlands

[51] University of Otago, Dunedin, New Zealand

[52] University of Ibadan, Ibadan, Nigeria

[53] Norwegian Institute of Public Health, Oslo, Norway

[54] VEAS, Slemmestad, Norway

[55] Aga Khan University, Karachi, Pakistan

[56] National Institute of Health, Lima, Peru

[57] National Veterinary Research Institute, Puławy, Poland

[58] Institute of Public Health of the Republic of Macedonia, Skopje, Republic of Macedonia

[59] State Medical and Pharmaceutical University, Chişinău, Republic of Moldova

[60] Institut Pasteur de Dakar, Dakar, Sénégal

[61] Institute of Veterinary Medicine of Serbia, Belgrade, Serbia

[62] Nanyang Technological University Food Technology Centre (NAFTEC), Nanyang Technological University (NTU), Singapore, Singapore

[63] Nanyang Technological University, Singapore Centre for Environmental Life Sciences Engineering (SCELSE), Singapore, Singapore

[64] Public Health Authority of the Slovak Republic, Bratislava, Slovakia

[65] National Laboratory of Health, Environment and Food, Ljubljana, Slovenia

[66] University of the Witwatersrand, Johannesburg, South Africa

[67] Daspoort Waste Water Treatment Works, Pretoria, South Africa

[68] IRTA, Centre de Recerca en Sanitat Animal (CReSA, IRTA-UAB), Bellaterra, Spain

[69] Medical Research Institute, Colombo, Sri Lanka

[70] The Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden

[71] Swedish University of Agricultural Sciences, Uppsala, Sweden

[72] Ara region bern ag, Herrenschwanden, Switzerland

[73] Kilimanjaro Clinical Research Institute, Moshi, Tanzania

[74] Centre for Infectious Disease Control, Bilthoven, the Netherlands

[75] Agence de Médecine Préventive, Dapaong, Togo

[76] National Institute of Hygiene, Lome, Togo

[77] Division of Integrated Surveillance of Health Emergencies and Response, Lomé, Togo

[78] Public Health Institution of Turkey, Ankara, Turkey

[79] University of Washington, Seattle, United States

[80] Institute of Public Health in Ho Chi Minh City, Ho Chi Minh, Viet Nam

[81] University of Zambia, Lusaka, Zambia

Acknowledgements

This study has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 643476 (COMPARE), the World Health Organization, and The Novo Nordisk Foundation (NNF16OC0021856: Global Surveillance of Antimicrobial Resistance). We would like to thank Miranda de Graaf for the technical assistance at Erasmus MC.

Author Information

Contributions

D.F.N., B.O.M., M.V.T.P. and M.C. designed the study. B.O.M., M.V.T.P., S.V. and M.C. performed the experiments. D.F.N. performed the data analysis, data interpretation and wrote the manuscript. The Global Sewage Surveillance project consortium provided the samples. The Global Sewage Surveillance project consortium, P.M. and F.M.A. coordinated sampling and sample transportation. M.P.G.K., B.O.M., M.V.T.P., The Global Sewage Surveillance project consortium, P.M. and F.M.A. revised the manuscript.

The authors declare no competing interests.

Correspondence and requests for materials should be addressed to

m.koopmans@erasmusmc.nl

Extended data

Extended Data Table 1 List of viral families and viral species detected in other metagenomic sewage surveillance studies.

Virus family	Virus species	Reference
Adenoviridae	Human adenovirus B	84
	Human adenovirus C	84
	Human adenovirus F7 201–332	84
	Human adenovirus 41	86
Astroviridae	Human astrovirus 1	83,86
	Human astrovirus 3	83
	Human astrovirus 4	83
	Human astrovirus 8	83
	Astrovirus MLB1	86
Caliciviridae	Norwalk virus	83,86
	Sapporo virus	83,86
Hepeviridae	Hepatitis E virus	83
Papillomaviridae	Human papillomavirus 112	86
	Papillomaviridae	84
Parvoviridae	Adeno-associated virus	83,86
	Human bocavirus 2	83,86
	Human bocavirus 3	83,86
Picobirnaviridae	Human picobirnavirus	83,86
Picornaviridae	Human Enterovirus B	84
	Aichi virus	83,86
	Human cosavirus D	83
	Human coxsackievirus B2	83
	Human coxsackievirus B6	83
	Human echovirus 11	83
	Human enterovirus 76	83

Virus family	Virus species	Reference
	Human enterovirus 97	83
	Human parechovirus 1	83
	Human poliovirus 2	83
	Saffold virus	83
	Salivirus NG-J1	83
	Human klassevirus 1/Salivirus NG-J1	86
	Human parechovirus 1	86
	Human parechovirus 3	86
	Human parechovirus 4	86
	Human parechovirus 7	86
Polyomaviridae	JC polyomavirus	84
	BK polyomavirus	84
	Polyomavirus HPyV6	86
Reoviridae	Banna virus	83

viromeBrowser: A Shiny app for browsing virome sequencing analysis results

David F. Nieuwenhuijse[1], Bas B. Oude Munnink[1], Marion P. G. Koopmans[1]*

Viruses 13 (3), 437, 2021

4

4.1 Abstract

Experiments in which complex virome sequencing data is generated remain difficult to explore and unpack for scientists without a background in data science. The processing of raw sequencing data by high throughput sequencing workflows usually results in contigs in FASTA format coupled to an annotation file linking the contigs to a reference sequence or taxonomic identifier. The next step is to compare the virome of different samples based on the metadata of the experimental setup and extract sequences of interest that can be used in subsequent analyses.

The viromeBrowser is an application written in the opensource R shiny framework that was developed in collaboration with end-users and is focused on three common data analysis steps. First, the application allows interactive filtering of annotations by default or custom quality thresholds. Next, multiple samples can be visualized to facilitate comparison of contig annotations based on sample specific metadata values. Last, the application makes it easy for users to extract sequences of interest in FASTA format.

With the interactive features in the viromeBrowser we aim to enable scientists without a data science background to compare and extract annotation data and sequences from virome sequencing analysis results.

4.2 Introduction

The use of metagenomic sequencing to explore the virome of complex samples such as stool, sewage and environmental samples has increased over the years[174]. Alongside the increase in virome studies a plethora of viral metagenomics processing workflows have been created[31]. The results of experiments in which complex virome sequencing data is generated are difficult to visualize and unpack for a person without programming experience. After processing of the raw sequencing data by a next generation sequencing (NGS) processing workflows the usual output consists of contiguous sequences (contigs) in FASTA format and an annotation file linking the contigs to a reference sequence or taxonomic identifier in tabular format[175–178]. For example, Virusfinder outputs several text files containing annotations and contig sequences as results[175] and VirFind outputs tabular annotation and FASTA files for viral and non-viral annotations[177]. Further unpacking of results and getting an overview of what viruses are found in which samples is especially difficult when this data is spread over multiple tables containing many annotations. Many of the workflows generate a summary file and figures containing a selection of the annotated sequences from the sample based on predefined selection criteria. As example SURPI generates a summary annotation file specifying read counts and contig coverage, but relies on Excel for annotation summarizing and comparison[176]. Interestingly Virus Identification Pipeline (VIP) produces an html report which can be browsed interactively for individual samples and viruses, but does not allow for comparison of samples[178]. A first and routine step in virome analysis is an annotation quality check. If quality thresholds are set upfront in the sequence annotation workflow, depending on the virus and the intentions of the user, these settings can be too stringent or too relaxed resulting in either false negatives or false positives[179–181]. Therefore, manual inspection and manual filtering using custom quality criteria is often needed which means that the workflow has to be rerun with different settings or, if possible, the unfiltered annotation results can be filtered using custom scripts and further analyzed using visualization tools.

Several visualization tools have been made to view annotation results while staying relatively agnostic of the metagenomic sequencing workflows such as MEGAN, Pavian, Krona, PanViz, MetaViz and Anvi'o[118,182–186]. MEGAN and Pavian perform very extensive analyses, but only accept specific input formats making it less flexible to use with different kinds of analysis workflows. PanViz, MetaViz and Anvi'o are tailored to the analysis of bacterial metagenomic data and are less well suited for virus data. Krona is very flexible and easy to use but cannot be used to easily compare multiple samples side by side. Paid software such as Geneious (https://www.geneious.com), CLC bio (https://digitalinsights.qiagen.com) are also available but require expensive licensing and cannot be customized or cannot visualize BLAST outputs in a concise manner.

In collaboration with end-users we developed viromeBrowser, a virome browsing app, that works through the steps they commonly use when interpreting sequencing outputs. The browser imports multiple annotation files and a corresponding FASTA files containing annotated contigs, addressing three common processes: (1) annotation quality assessment, (2) dataset visualization and interpretation, and (3) extraction of sequences with a specific annotation. An iterative design process was employed with end-users to allow intuitive browsing, selecting and exporting of specific sequences and – selections, as well as visualization of these sequences combined with metadata.

4.3 Materials and Methods

The virome browser was written in the R programming language and makes use of the Shiny[187] web application R package. The interactive visualizations are created using the rbokeh package[188]. Rsamtools[189] is used to handle the FASTA files and open reading frame prediction is performed by splitting the contig sequence based on the presence of a stop condon using the Biostrings package[190]. The application was packaged for easy installation following the guidelines of the O'Reilly R Packages book[191] and is available on the R CRAN platform. Examples of virome datasets with metadata were used to present the tool to end-users and invite feedback for further optimization. In total, we went through several iterations before finalization of the application.

4.4 Results

The viromeBrowser is implemented in Shiny[187] a web application framework and depends on several other R packages as shown in the package description at https://CRAN.R-project.org/package=viromeBrowser. Even though Shiny can be used to create web applications, the server and client part of the application can also be set up locally for analysis of data in situations where sharing may not be not allowed, such as in clinical settings. This setup also allows an institute to run the computational heavy part on a centralized powerful computer while running the lightweight client on the user's computer. The user interface (UI) elements of the viromeBrowser have been made modular to allow easier expansion of the app, using the Shiny functionalities for module development. A video demonstrating the complete application is provided in Supplementary File S1.

4.4.1 Data input

The app is separated into tree main parts which are listed as separate menu items. The first part allows data to be loaded into the app by selecting a FASTA file containing contig sequences, a contig annotation file, a BAM file containing mapped reads to contigs and a metadata file (**Figure 1**). These files need to be in the format specified in the packages vignette. Briefly, the contigs have to be in regular FASTA format with the FASTA headers exactly matching the contig identifiers in the annotation and read mapping files, the contig annotation files need to be in BLAST tab-separated format the read mapping files need to contain the mapped reads to the aforementioned contigs and the metadata file needs to be in tab-separated format in which the file names are in the first column and the sample characteristics are listed in each additional column. Currently only default tabular BLAST-like output is supported, but other tabular formats can be implemented by creating a novel import function. To determine the taxonomic lineage of each annotation based on the associated taxonomic identifier the application uses the "rankedlineage.dmp" file which is downloaded automatically from the NCBI taxdump database at ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/. For contigs with multiple associated annotations, which is not unusual for BLAST output, a lowest common ancestor (LCA) taxonomic lineage is determined by finding the lowest order taxon which is present in all annotations. If no LCA can be found the contig is annotated as "root". The LCA calculation is performed after annotation quality filtering to avoid spurious hits from interfering in the process.

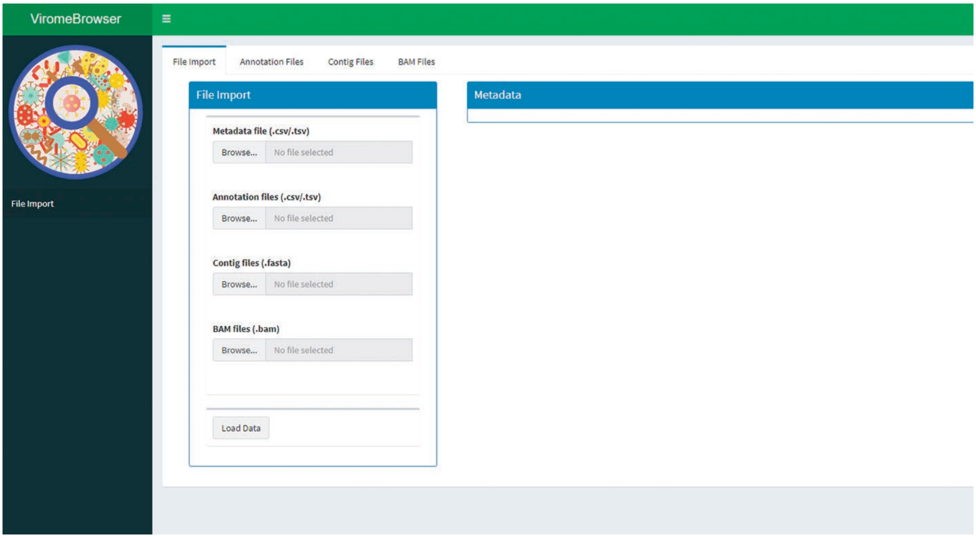


Figure 1 Example screenshot of the file import screen. Metadata, annotation files and contig files are uploaded and processed in the file import screen. An excerpt of the metadata, the annotation files and the contig files can be visualized under the corresponding tabs.

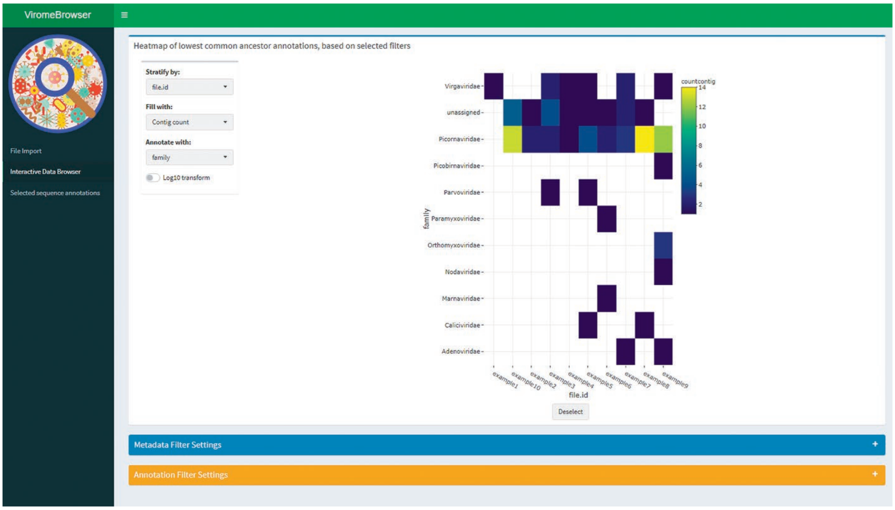
4.4.2 Interactive quality assesment

The second part of the app displays an interactive heatmap based on the annotations provided in the contig annotation and metadata files. The initial data of the heatmap is based on default quality settings aimed at highlighting contigs larger than 500 nucleotides with more than 90 percent identity over a length of 500 nucleotides or more to the chosen references. Workflows may differ with regards to annotation quality parameters, and therefore, the default quality settings are part of the data import function and can be customized accordingly. Users can override the default quality thresholds by filling in other values in the quality threshold tab. This gives full control over the filtering of the data allowing users to browse contigs for virus discovery or set the filters such that only results with high confidence are visible. Annotation settings could for instance be set to specifically target low sequence identity annotations to filter out novel viruses, or high sequence identity to continue with high certainty annotations only. On the other hand, annotation settings can be made stricter for diagnostic purposes to prevent false positives.

Further selection of specific annotations of interest can be done by using the rectangular selection tool in the interactive heatmap. The selection made in the heatmap will propagate to the rest of the app enabling the user to zoom in to a specific annotation, sample or sample characteristic of interest. Deselecting all tiles in the heatmap will reset the selection resulting in the selection of all annotations in the current view.

4.4.3 Metadata guided sample comparison

Once the quality settings have been defined, the next step is the interpretation of the obtained results in combination with the metadata of the prepared experiment or sample cohort. This part was implemented in the viromeBrowser by an interactive heatmap which can be used to stratify, filter and group samples based on the provided metadata file. The interactive heatmap can also be used to compare multiple sample annotation files in a single overview and from different points of view. The interactivity is enabled by three dropdown menus (**Figure 2**). In the first menu the factor can be chosen by which the samples in the heatmap will be stratified. In the second two menus the value by which the heatmap tiles are colored and the taxonomic level by which the heatmap has to be drawn can be selected. The fill options are either by number of contigs, absolute number of reads, or relative number of reads, scaled by the total number of reads in the BAM read mapping file.



Heatmap of lowest common ancestor annotations, based on selected filters

Stratify by: file.id

Fill with: Contig count

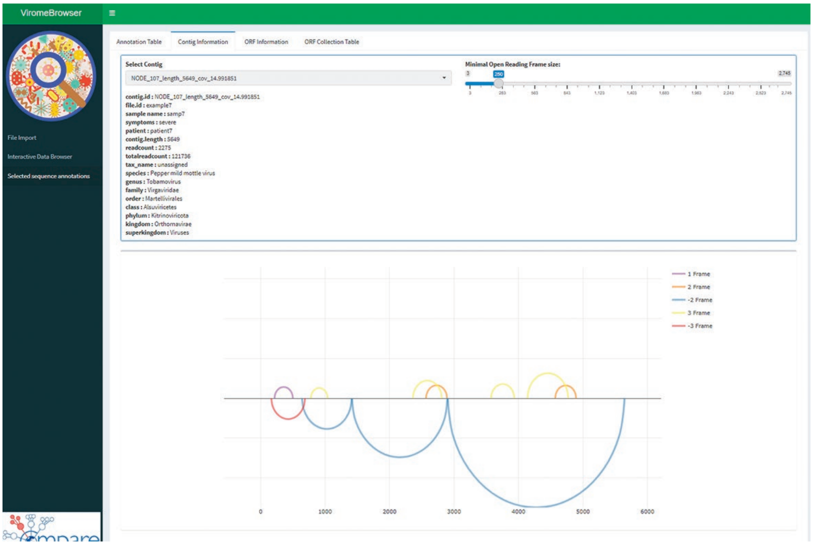
Annotate with: family

Figure 2 Screenshot of the interactive data browsing heatmap. Stratification variable and filter criteria can be selected in the browser settings window. Annotation quality filter settings are available as a drop-down menu in the bottom of the page. Hovering over the heatmap shows the contig annotations and the number of contigs or read counts. Selecting tiles of the heatmap results in selection of only those contigs for further analysis and downloading.

4.4.4 *Sequence extraction and downloading*

Another functionality of the app is further inspection and downloading of specific contigs sequences. A table shows the annotations based on the selections made in the interactive heatmap and quality thresholds in the previous tab. Sequences can be selected from the table for further inspection or saving, but it is also possible to save all sequences from the table by selecting the download all filtered button.

Users can continue to zoom in on a single contig in the app by selecting one or more contigs from the table and continuing to the contig information tab (**Figure 3**). In this tab a single contig is displayed for which open reading frames (ORF) are predicted by performing a canonical stop codon lookup and using these to split each frame into ORF fragments. The ORFs are represented by arcs on the contig and are predicted for all 6 frames and small spurious ORFs can be filtered by setting a minimal ORF size. The visualization allows users to perform a quick check of the expected ORF structure of an annotated virus. Individual ORFs can be viewed under the ORF information tab and nucleotide or amino acid sequences can be directly selected and copied. Alternatively, all displayed ORFs can be collected in the ORF collection table and saved in FASTA format. For further analysis it is sometimes useful to only extract certain ORFs from a genome, which is possible by separately saving ORFs in FASTA format.



Select Contig: NODE_107_length_5649_cov_14.991851

Minimal Open Reading Frame size: 250

contig.id : NODE_107_length_5649_cov_14.991851
file.id : example7
sample name : samp7
symptoms : severe
patient : patient7
contig.length : 5649
readcount : 2275
totalreadcount : 121736
tax_name : unassigned
species : Pepper mild mottle virus
genus : Tobamovirus
family : Virgaviridae
order : Martellivirales
class : Alsuviricetes
phylum : Kitrinoviricota
kingdom : Orthornavirae
superkingdom : Viruses

Figure 3 Example screenshot of the sequence information tab. Contigs can be selected from a table of previously selected annotations. Further inspection can be by visualization of the open reading frame (ORF) structure and downloading of individual ORFs or the complete contig.

4.5 Discussion

The viromeBrowser can be used to interrogate the analysis results of complex metagenomic sequencing experiments such as viromes of stool, wastewater or environmental samples. A unique feature is that the viromeBrowser starts with filtering of analysis results by quality allowing less or more stringent selection of annotations which is important given the large diversity differences of viruses and the different interests of users, be it virus discovery or viral diagnostics, bridging the gap between research setting and diagnostic usage. The former requires a broader view in which quality parameters can be more lenient which is useful for virus discovery while the latter requires stringent quality threshold to prevent false positive results. This limits the use of viromeBrowser to workflows that output unfiltered annotation results of contigs as part of their analysis, which makes it incompatible with read based annotation workflows such as SURPI[176].

A feature that was added based on end-user input was the possibility to compare viromes based on added metadata parameters, for instance in a setting where the goal is to find differences in the virome of patient and control groups, linking annotations with the associated contig sequences and custom extraction of selected sequences for further analysis. This allows interactive visualizations to aid the user in making a manual selection of the data which can be used in further analyses such as primer design, phylogenetic analyses or variant analysis. An alternative tool with strong visualization and result selection features is Genome Detective[192], which provides user-friendly NGS data processing and visualization via a web-interface. However, for settings where data cannot be shared externally, local installation of Genome Detective, can only be done by paying a licence fee and paying a fee per sample analyzed.

Future improvements can be made on the implementation of other annotation formats creating a more flexible interface for annotation data input. Additional improvements can be added for more visualization options for detailed contig analysis such as contig coverage plots and variant visualization by importing read alignment files.

To keep the application interactive and lightweight the pre-required raw data processing and annotation steps are not performed by the viromeBrowser, which is a disadvantage for users without a running NGS processing workflow. To address the need for data processing by users without their own raw data processing capacity the European Bioinformatics Institute has developed datahubs in which raw data can be uploaded and analyzed with several standardized workflows[193]. After the preprocessing and annotation steps have been performed, the viromeBrowser can be used to inspect the results.

4.6 Conclusions

In conclusion, here we present viromeBrowser, an interactive application to browse through the annotation results of viral metagenomic sequencing experiments. Interactively selecting viral annotations of choice and manually tuning quality thresholds should make it easier for scientists with little programming experience to analyze complex metagenomics data. Facilitating separating and downloading of contigs of interest will make it easier to perform follow up analyses with these sequences. The viromeBrowser is implemented as an R package which is distributed by the R CRAN platform. Future updates will be available from the same platform.

Author affiliations

[1] Viroscience Department, Erasmus Medical Center, Doctor Molewaterplein 40, 3015 GD, Rotterdam, The Netherlands

* Correspondence: m.koopmans@erasmusmc.nl

Author Contributions

Software, D.F.N.; writing – review and editing, D.F.N., M.K., B.O.M.; All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by COMPARE, grant number 643476

Conflicts of Interest

The authors declare no conflict of interest.

Towards reliable whole genome sequencing for outbreak preparedness and response

David F. Nieuwenhuijse[1], Anne van der Linden[1], Robert H. G. Kohl[2], Reina S. Sikkema[1], Marion P. G. Koopmans[1], Bas B. Oude Munnink[1*]

BMC genomics 23 (1), 1-9, 2022

5

5.1 Abstract

To understand the dynamics of infectious diseases, genomic epidemiology is increasingly advocated, with a need for rapid generation of genetic sequences during outbreaks for public health decision making. Here, we explore the use of metagenomic sequencing compared to specific amplicon- and capture-based sequencing, both on the Nanopore and the Illumina platform for generation of whole genomes of Usutu virus, Zika virus, West Nile virus, and Yellow Fever virus.

We show that amplicon-based Nanopore sequencing can be used to rapidly obtain whole genome sequences in samples with a viral load up to Ct 33 and capture-based Illumina is the most sensitive method for initial virus determination.

The choice of sequencing approach and platform is important for laboratories wishing to start whole genome sequencing. Depending on the purpose of genome sequencing the best choice can differ. The insights presented in this work and the shown differences in data characteristics can guide labs to make a well informed choice.

5.2 Background

Due to the increased connectivity of the modern world, deforestation and climate change, viral pathogens which used to be restricted to certain geographic areas or hosts have increased potential to spread to previously naïve populations. This is especially true for arthropod-borne (arbo) viruses like members of the genus *Flavivirus* as could be seen during the large Zika virus (ZIKV) outbreak in Brazil[194] and the recent expansion of Usutu virus (USUV) and West Nile virus (WNV) to Western Europe[195–198]. In Europe, also the risk of the introduction of other *Flaviviruses* like Yellow Fever virus (YFV) increases due to the expanding establishment of competent vectors along with other factors[199].

Whole genome sequencing (WGS) is increasingly advocated as important public health tool and has proven to be valuable during viral outbreaks to identify transmission chains, determine epidemic links and detect specific mutations[94,194,200]. Especially in the beginning of outbreaks this information may help to inform public health officials, provided that the data is generated and analysed in a timely fashion as was done for the recent SARS-CoV-2 outbreak[201]. This, however, can be challenging sometimes as it was for the Zika virus outbreak[202]. The successful and timely generation of WGS depends on the types of infections, sample types, instruments used for sequencing, costs, and quality of data and analysis. For instance for Zika virus, the viral load decreases rapidly after onset of symptoms[203], a phenomenon commonly observed during infections by members of the *Flaviviridae* family[204,205] requiring protocols that work with low viral loads. The same applies when trying to sequence viruses from small volume samples, for instance when specimens from birds or mosquito pools are used[197].

During the more recent virus outbreaks, amplicon-based approaches were used to generate full length sequences of emerging viruses[94,194,201,206]. This approach is specific and highly sensitive up to a Ct value of around 35[207]. Nonetheless, the main limitation of this approach is that specific primer sets have to be developed for different (sets of) pathogens which are based on our current knowledge about virus diversity. This is not the case when using metagenomic sequencing, where all RNA and/or DNA present in the sample will be sequenced. However, this approach is sensitive to the presence of host background and/or bacterial DNA, decreasing the detection limit[208]. Capture probes can be used to increase the sensitivity of the metagenomic approach while still benefiting of the broad coverage of virus diversity. These probe sets can be designed to target a large spectrum of genomes of viral taxa that are known to infect vertebrates, thus providing potential to detect a wide range of pathogens. Briese et al. showed that a capture probe set (VirCapSeq-VERT) resulted in a 100- to 10,000 fold increase in viral reads compared to direct metagenomic sequencing[27].

There are several high-throughput second and third generation sequence machines available at the moment. The widely accepted golden standard is the sequencing by synthesis platform developed by Illumina, but novel platforms have been developed such as nanopore based sequencing by Oxford Nanopore. There are several differences between Illumina and Nanopore based sequencing. Compared to the Illumina sequencing machines, the cost of the Nanopore sequencing hardware is relatively low, the machine is portable, and data is generated in real-time. This gives Nanopore sequencing a benefit over Illumina sequencing in a setting where costs need to be kept to a minimum and speed is key[30]. However, the sequence method of choice is also dependent on the specific research question. For example, for early detection in the beginning of an outbreak, time to result is an important parameter, while later in the outbreak more detailed analysis using high quality sequencing reads to identify minority variants within patients may become important. For this application deep coverage with high quality reads can be preferred over speed.

The reported high error rate[209] compared to Illumina might limit the application of Nanopore sequencing, depending on how the data is analysed. For WGS the error rate can be compensated by creating a consensus sequence based on a larger number of overlapping reads compared to what is standard for Illumina sequencing. Previously it was shown that a read coverage of 100x is sufficient to compensate for the errors generated by Nanopore sequencing when using an R9.4 flowcell[23], which can go down to 20x using the recently released R10 flowcell[210].

Here, the performance of whole genome sequencing is compared for four members of the *Flaviviridae* family in three different concentrations using five different sequencing approaches. Cell culture supernatants of USUV, WNV, YFV and ZIKV were diluted to a Ct value of 25, 29 and 33 and sequenced on the Illumina and Nanopore sequencing platform. The samples were sequenced using an amplicon-based approach and a metagenomic approach on both platforms and, due to technical constraints, a capture-based approach on Illumina only (**Figure 1**).

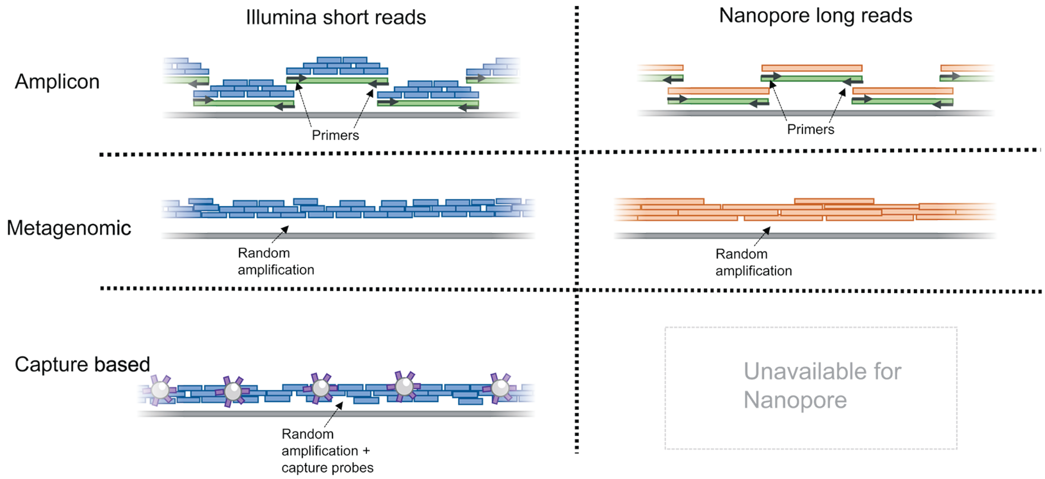


Figure 1 Overview sequencing approaches. Grey bars represent the to be sequenced genome, blue and orange are short and long reads respectively, generated either directly or from amplicons (green) or captured nucleic acid using capture probes (purple).

5.3 Results

5.3.1 Amplicon based sequencing on the Nanopore and Illumina platform

When generating complete genomes using amplicon sequencing, there was little difference between Nanopore and Illumina. As expected, most reads belonged to the targeted virus (**Figure 2**). On average, the total number of reads did not vary much for the different Ct values. Read counts for Illumina sequencing were around 3M per sample with a single 5,4M exception, while Nanopore produced around 400k reads with a 682k and 917k as exceptions. Using the amplicon approach, both sequencing platforms were able to generate complete or near complete genomes from most samples. Illumina sequencing resulted in more or equal percentage coverage in most cases (10 out of 12), but performed worse in some cases (2 out of 12): for WNV Ct 29 and Ct 33 Nanopore covered 5 and 9 percent more of the genome respectively (**Figure 3**). In all the amplicon based results the coverage depth varied greatly along the genome resulting in a spiky pattern which was mostly caused by differences in the performance of individual amplicons. In all cases the coverage difference between performant amplicons and failing amplicons was more than 51,000x (**Figure 4**). Extreme differences in coverage were especially visible in high Ct samples such as Illumina WNV Ct 33 where one part of the genome had 800,000x coverage and other parts were not covered at all.

5.3.2 Metagenomic sequencing on the Nanopore and Illumina platform

With the metagenomic approaches, using Nanopore sequencing resulted in, on average, 0.8% more virus specific reads than using Illumina, on an average of 1.8% viral reads (**Figure 2**). The total number of reads per sample was heavily influenced by Ct value for Nanopore but not for Illumina, which produced around than 3 million reads for all samples. With Nanopore sequencing, the total number of reads dropped with 25% on average between the highest and the lowest Ct value, despite the equimolar pooling of the samples. Illumina sequencing resulted in reliable (more than 5x coverage) complete genome sequences for all samples with Ct 25 and 29 except for WNV. The samples with Ct 33 had between 3% and 46% of the genome covered at at least 5x. With Nanopore sequencing only the highest viral load USUV and ZIKV samples resulted in reliable (100x coverage) complete genome sequences. For the other viruses there was only partial coverage of the genome and the Ct 29 and Ct 33 samples had little to no coverage, but did allow idenficitation of all the viruses. The coverage profile of the Ct 25 Nanopore genomes was relatively smooth with high coverage across the entire genome. The coverage of the genomes generated by Illumina was less smooth resulting in an occasional drop of coverage below the 5x coverage threshold (**Figure 4**).

5.3.3 Capture-based sequencing on the Illumina platform

When using a capture-based approach for Illumina sequencing (here VirCapSeq-VERT) the percentage of viral reads was much higher compared to the metagenomic approach and comparable to the amplicon based approach. The total number of sequence reads was heavily influenced by the viral load in the sample as a result of the pooling all samples before capture (**Figure2**). The number of generated complete and near complete genome sequences was comparable to the amplicon-based Illumina approach although it performed worse for samples with Ct 33 (**Figure 3**). The coverage profile of the genomes was relatively even, although some regions seemed to be preferably sequenced, resulting in coverage spikes. These spikes are especially noticeable in the Ct 33 samples and resemble the metagenomic Illumina coverage profile (**Figure 4**).

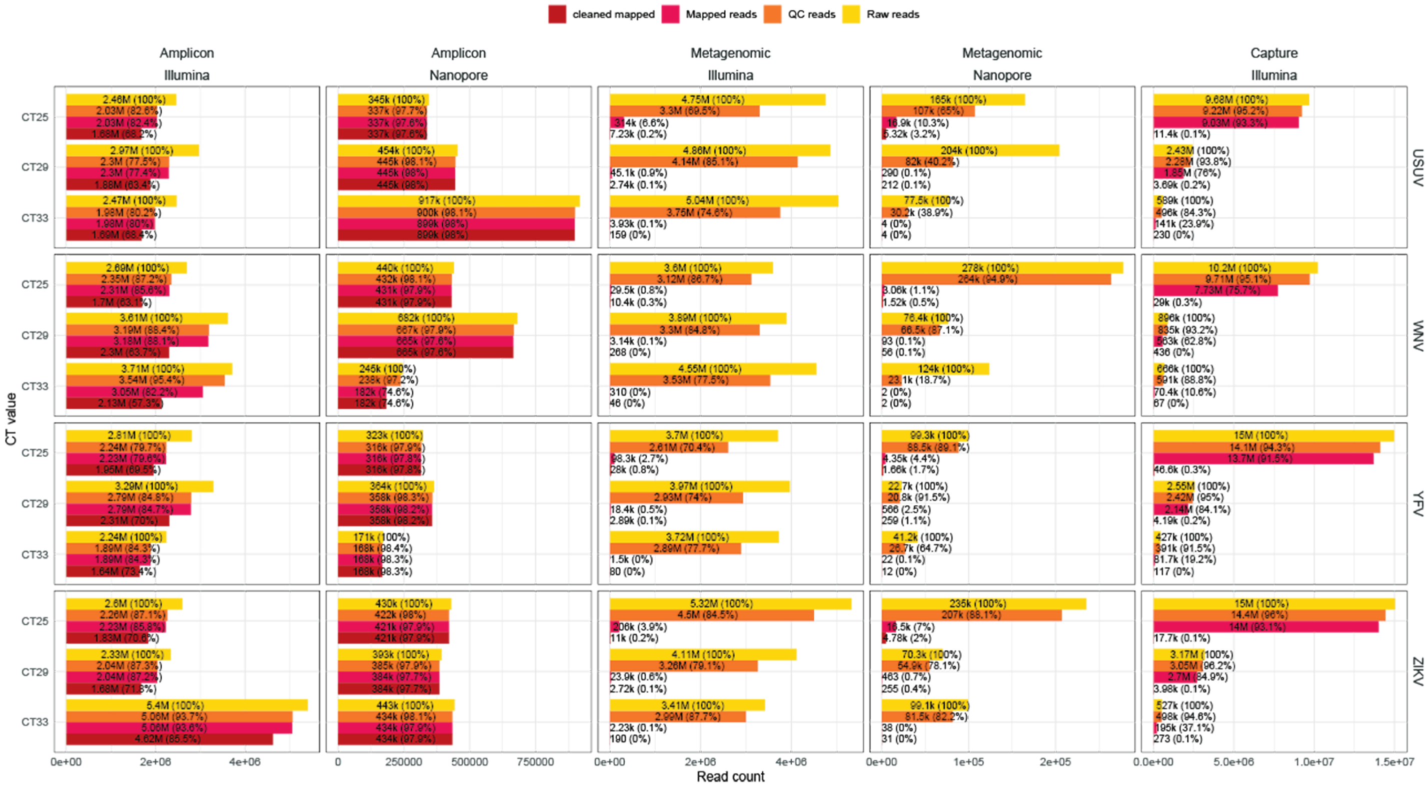


Figure 2 Overview of read numbers by virus and sequencing approach. The bars indicate the total number of sequence reads at different stages of the analysis process. Values and percentages of the number of reads are indicated in text, abbreviated using SI units.

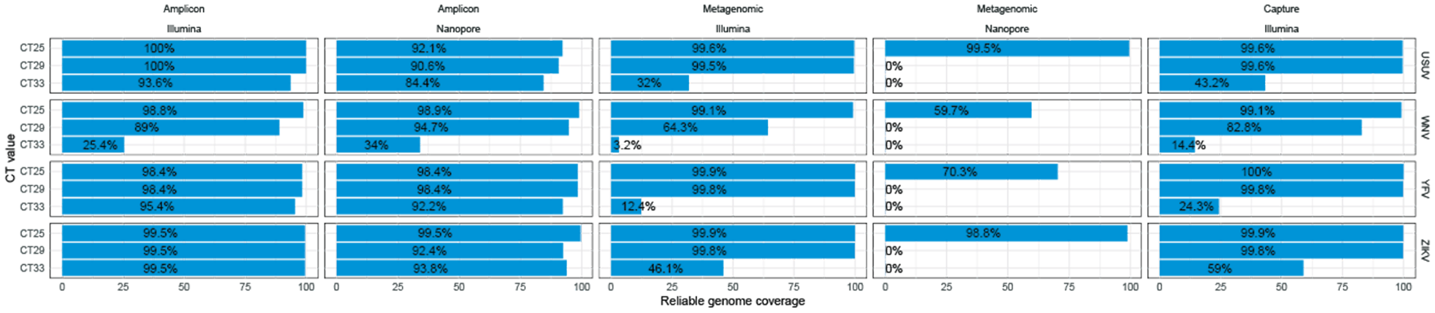


Figure 3 Overview of the percentage of reliable genome coverage. The bars indicate the percentage of coverage of the respective viral genomes with the respective approach and platform. The coverage percentage of amplicon data is calculated based on the amplified region of the genomes. The other percentages are calculated based on the size of the complete reference genome.

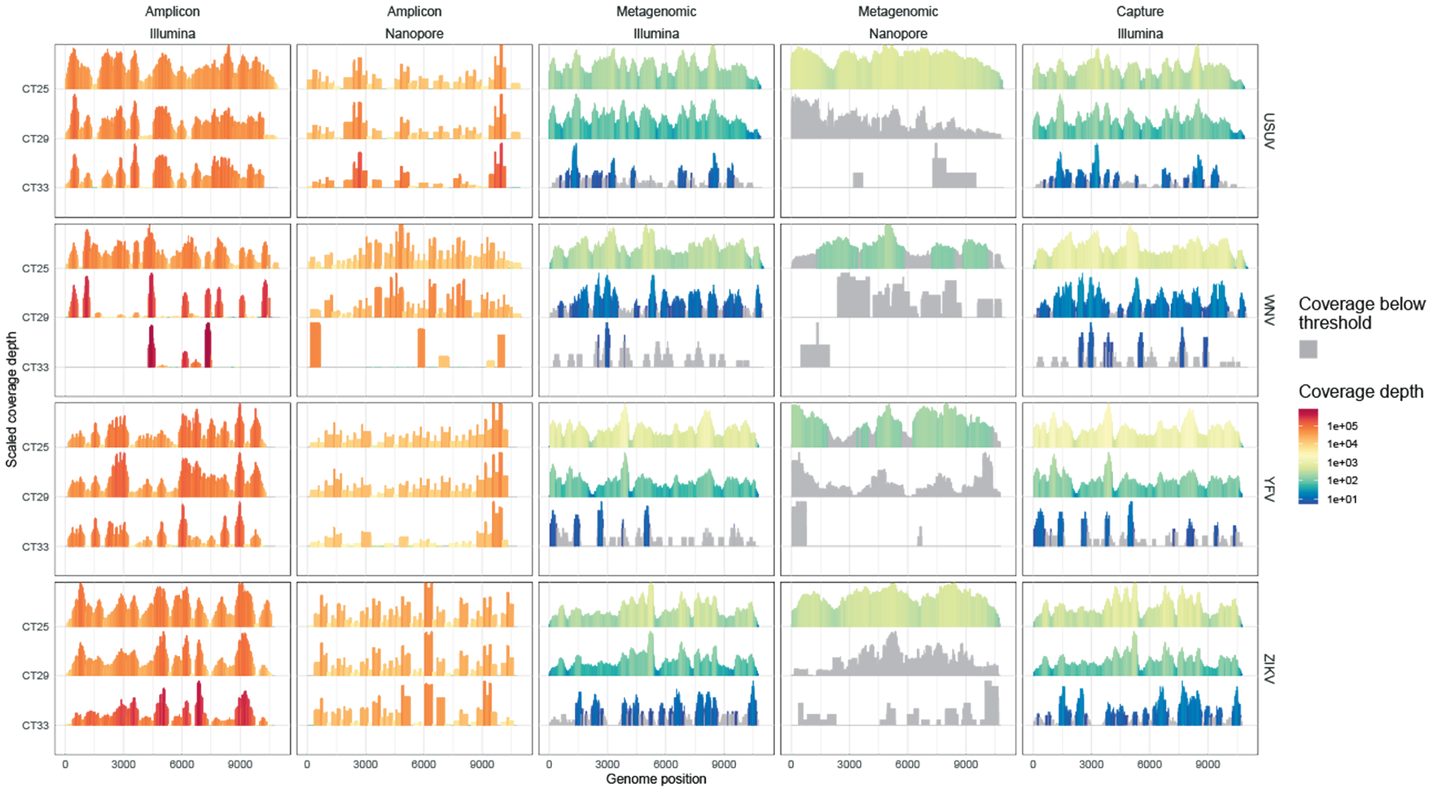


Figure 4 Overview of read coverage across the genomes. The height of the peaks represents the relative coverage profile at that position scaled by the maximum coverage of across the sample. The color scale represents the log transformed coverage depth. Profiles and coverage depth represent the cleaned mapping results. A grey coverage color indicates coverage below the coverage threshold (5x Illumina, 100x Nanopore).

5.3.4 Influence of virus concentration and sequencing approach on consensus sequence quality

The quality of the complete and partial genomes retrieved from the different sequencing approaches was evaluated by comparing the individual sequence variations found with each method. Using PySam[211], the major variants were extracted at each position in the alignment. Those variants that were present across all approaches (except for those that did not have sufficient coverage at the variant position) were accepted as true variants **(Figure S2)**. The other variants were considered errors and where investigated **(Figure S1)**. Most errors could be traced back to poorly trimmed primer sequences. Therefore we developed a custom script to better trim the primer sequences from the BAM file using the primer's coordinates. In addition, several errors were found in the consensus sequences generated with the metagenomic or capture approach presumably resulting from PCR amplification errors. Dereplicating the reads in these alignments using Sambamba's "markdup" method[212] resolved these errors, but reduced the number of mapped reads with 88% on average, showing that with these approaches many technical replicates are generated. Strongly softclipped reads (>10%) were also removed as these were not dereplicated by Sambamba and often contained errors. After resolving these issues the remaining errors seemed to be related to the sequencing technology and the viral load **(Figure 5)**. Nanopore sequencing has difficulty with calling insertions and deletions, as multiple single nucleotide deletions were found, albeit at low variant fraction, resulting from the variable number of deletions present in the reads at these positions. For the same reason it was difficult to automatically call the large deletion at position 10,390 in the WNV genome. Also, several erroneous substitutions were found in the Nanopore data at a low frequency, which could be attributed to the error rate which caused systematic errors in some positions that were difficult to distinguish from real variants, especially since some of the true variants were also present at a relatively low frequency in the Nanopore data **(Figure S2)**. The only errors in the Illumina results were five false positive substitutions at positions WNV 4509, 5034, 7088 and YFV 822, 3711 (**Figure 5**). These errors could not be attributed to primer errors, PCR amplification, or softclipping, but were located in low coverage areas of high Ct value samples indicating that the coverage may be too low for reliable variant calling with high Ct value samples.

The effect of these few errors on the interpretation of the genome sequences may seem unimportant, but can be crucial when using WGS for analyses that rely on the detection of only several variants, such as source tracing, lineage assignment and vaccine escape. In those cases, great care should be taken to rule out any false positive or false negative variants in the genomeby manual curation, especially at high Ct values. The errors in the Nanopore data are even more complicated to resolve automatically because real variants and erroneous variants can have similar variant frequencies. Especially insertions and deletions are hard to correctly interpret automatically and manual curation is necessary in most cases.

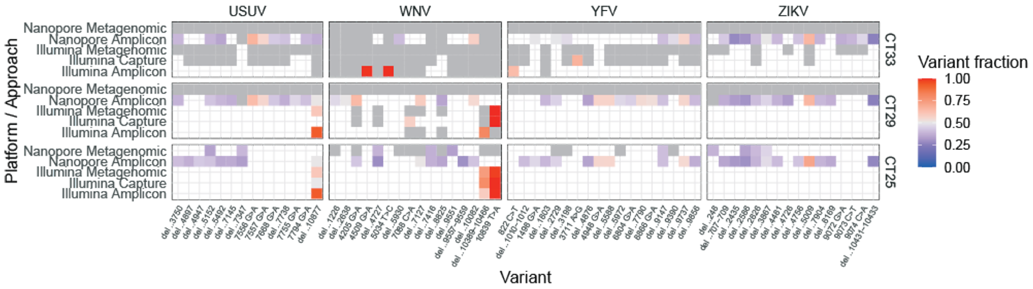


Figure 5 Overview of variants across the genomes. The x axis shows the variants found across the 4 genomes. The color scale represents the fraction of mapped reads containing the indicated variant in the read alignment. A darkgrey tile color indicates coverage below the coverage threshold (5x Illumina, 100x Nanopore).

5.3.5 Usability of different sequencing approaches for public health decision making

The use of the different sequencing approaches was also evaluated based on the time to result, the costs of the sequencing instrument, the costs per sample sequenced, the specificity, the suitability for sensitive WGS, and the application for initial detection of a pathogen (**Table 1**). Both the metagenomic and the capture-based approach can be used for initial detection of an unknown pathogen in the early days of an outbreak. Metagenomic Illumina sequencing is more suited for samples with lower viral loads, but the sequencing takes much longer and is more expensive. The cost and duration of metagenomic data analysis are also higher and longer, because assembling and annotating the obtained genomes is much more complex without a known reference genome. The preferred approach for contact tracing to identify transmission chains, or other situations that require quick results, is the amplicon-based Nanopore approach which results in full genomes up to Ct 29-33 and can be performed within one day with relatively low costs per sample. However, for this approach the target of interest should be known *a priori*. Contact tracing with amplicon sequencing on the Illumina platform is more sensitive, but also more expensive and slower. Compared to metagenomic sequencing the analysis of amplicon data can be performed much quicker, and is therefore cheaper, because it can be automated in a workflow specifically tailored to finding a virus of interest. If amplicon sequencing is not an option the other approaches can give reliable results, but only at high viral load or at a higher cost and turnaround time.

Table 1 Overview of the usability of the different sequence methods. The sequence method, platform, time to results, costs and suitability for public health decision making is indicated.

Sequence method	Amplicon	Amplicon	Metagenomic	Metagenomic	Capture-based
Platform	Illumina	Nanopore	Illumina	Nanopore	Illumina
Time to result	3 days	<24 hours	3 days	<24 hours	8 days
Costs of the instrument*	€200.000	€1.000	€200.000	€1.000	€200.000
Cost of the analysis**	€100	€100	€600	€600	€600
Costs per sample	€200	€50	€200	€50	€300
Virus specific PCR required	Yes	Yes	No	No	No
Ct threshold for WGS	Ct29-33	Ct25-29	Ct29-33	Ct25-29	Ct29-33
Ct threshold for detection	Ct33	Ct33	Ct33	Ct29-33	Ct33

* Estimated costs are laboratory and/or country dependent

** Estimated costs are estimated based on a €100 per hour and are laboratory and/or country dependent

5.4 Discussion

Recent studies have shown the value of real-time WGS in public health decision making during a pandemic[213] and how WGS provides a tool to close the gaps in our knowledge about the global diversity of animal infecting viruses[214]. With the increasing demand for timely generation of sequence data, choises have to be made between short and long read sequencing and several different available protocols. We compared five different sequencing approaches for the detection and WGS of four different arboviruses in three dilutions and assess the potential use of the different sequence platforms and protocols for public health decision making in different stages of an outbreak.

We show that for the initial detection of an unknown viral pathogen the metagenomic approaches and the capture-based approach can be used in samples with a Ct value up to 33. Looking at sensitivity, capture-based Illumina sequencing is slightly superior to metagenomic Illumina sequencing, which is reflected in the total number of recovered viral reads across all dilutions and viruses. This is different from what was seen before, where capture was shown to result in a higher increase of genome coverage with an increase up to 20-fold for instance in a blood sample spiked with WNV, suggesting that capture is perhaps more suited for those samples with a very high amount of background host DNA or RNA or that for low Ct-value samples individual capture experiments have to be performed. Capture-based Illumina sequencing was also the most expensive and time consuming approach. Between metagenomic Nanopore and Illumina sequencing there is a tradeoff between turnaround time and sensitivity where Nanopore sequencing was shown to be 3 times faster, but is the least sensitive for initial virus detection as it generates only a few reads at Ct 33 in contrast to Illumina. However, given that for initial virus detection a few reads are sufficient and speed of detection is of importance[98], Nanopore metagenomic sequencing would be preferable.

For detailed outbreak investigation using phylogenetic analyses the focus is to generate an as complete as possible genome. From a sheer data volume point of view this seems to strongly favor Illumina sequencing which resulted in 50 and 61 million sequence reads for metagenomic and capture-based sequencing while Nanopore sequencing resulted in 5.2 million sequencing reads for the metagenomic approach. However, to achieve 5x coverage of the complete viral genome not that much data is necessary and for a limited number of samples the amount of reads generated with an Illumina run may be excessive. Previously, a coverage of 400x was determined to be sufficient to perform minor variant analysis[215] showing the excess of coverage with Illumina sequencing in our experiment, which resulted in up to 200 times more coverage. Nanopore has the benefit of generating much longer reads, which means that with the same number of reads more nucleotides can be covered. This can be seen by comparing the difference in mapped read counts with the difference in mapped nucleotides between Illumina and Nanopore. At Ct25, metagenomic Illumina sequencing on a Miseq produced 15 times more reads on average than metagenomic Nanopore, but the total number of sequenced nucleotides was only 5 times larger.

To make Illumina sequencing more cost effective more samples can be pooled in a single run, however the different sequencing approaches are also influenced by the pooling strategy. Even though samples were pooled equimolarly, there was a discrepancy in the number of sequence reads generated at different viral loads in the metagenomic Nanopore and capture-based Illumina sequencing approaches. The difference between the highest and the lowest viral load samples for the two approaches were up to 2- and 35-fold respectively. For metagenomic Nanopore the difference were the result of pooling, which is challenging due to the varying sequence lengths in the sequencing library and small pipetting errors, and therefore a 2-fold difference is generally seen as acceptable. With capture-based Illumina sequencing, the difference is mainly an effect of capturing viral DNA after pooling of the sequencing libraries. In lower load samples the fraction of viral DNA is lower compared to background or host DNA resulting in less captured DNA and therefore an underrepresentation of the sample in the final sequencing library since the pooling is performed based on the total amount of DNA present in the sample. This issue of balancing samples can be overcome by using individual capture reactions but this will also drastically increase the price of sequencing. It has been shown previously that multiple samples can be pooled and sequenced simultaneously[216], however, to do so, the nucleic acid concentration has to be measured accurately, which is especially difficult in the field, or in samples with very low viral loads.

When the target virus is known *a priori* and the focus is on in-depth characterization of the complete viral genome the amplicon approaches have a clear advantage and allow generating high quality genomes up to Ct 33. Nanopore is shown to do so at a modest cost and in a timely manner when 12 samples are multiplexed in one Nanopore sequence run, while Illumina sequencing gives similar results but at a higher turnaround time and cost. The benefit of longer reads is less pronounced with amplicon sequencing because the amplicons have a set size. Although larger amplicons would be possible with Nanopore sequencing, the reason behing using smaller amplicons is the increase in sensitivity[207]. Updating the amplicon primer pool and finetuning the primer concentrations will potentially increase the sensitivity of this approach even further and lead to more even coverage.

5.5 Conclusions

In this work we compare three sequencing approaches on two sequencing platforms using four arboviruses in three different dilutions and show the performance with respect to sensitivity and WGS completeness. The amplicon-based approach performed best for WGS in almost all cases given the assumption that the target virus is known upfront. Capture-based Illumina sequencing performed best at agnostic virus detection, although at a higher cost and lower turnaround time compared to metagenomic seqecing. Choosing a sequencing platform and approach is important for labs adopting genome sequencing, but depends on the stage of an outbreak and the to be answered questions in public health decision making. The data presented in this work offers a deep insight in the characteristics of each approach and help making this choice.

5.6 Material and methods

5.6.1 Sample preparation

All the cell cultured passaged viruses were obtained from the Erasmus Medical Center. Viruses were cultured in Vero cells and cell culture supernatant was diluted in USUV, WNV, YFV and ZIKV negative serum to Ct values 25, 29 and 33 as determined by specific real-time PCRs[217–220]. RNA was extracted and aliquoted in different batches to prevent additional freeze-thaw steps. For USUV the AS201700077 strain was used (MN122189.1), for WNV the B956 strain (AY532665.1), for YFV the YFV_t146a212_Jan19_ur strain (MK760665.1) and for ZIKV the SL1602 strain (KY348640.1).

5.6.2 Library preparation for amplicon sequencing

RNA was transcribed into cDNA using random hexamers (Thermo Fisher) and ProtoScript II (NEB) after which a multiplex PCR was performed in two different reactions as described previously[207]. For USUV and YFV the same primer set was used as previously described[23,217], for ZIKV and WNV new primer sets were developed. The primer sequences and concentrations are displayed in **Supplementary Table 1**. For Illumina sequencing the KAPA HyperPlus Kit (Roche) was used, while for Nanopore sequencing the native barcoding genomic DNA Kit was used (Nanopore).

5.6.3 Library preparation for metagenomic sequencing

RNA was transcribed into cDNA using random hexamers (Thermo Fisher) and SuperScript IV (Thermo Fisher) and dsDNA was generated using Klenow (NEB). For Illumina sequencing the KAPA HyperPlus Kit (Roche) was used with the following modifications. The adapters were 1:10 diluted and an extra AMPure beads (Beckman Coulter) wash step was performed after adapter ligation. For Nanopore sequencing the “Low input genomic DNA” with PCR kit was used following the manufacturer’s instructions, (SQK-PBK004, Nanopore) apart from an additional wash step that was performed after adapter ligation.

5.6.4 Library preparation for VirCapSeq-VERT sequencing

After metagenomic sequence library preparation, as described above, all samples were pooled to a final concentration of 1000ng and a specific capture for all vertebrate viruses was performed using VirCapSeq capture probes which were previously described[27]. All 12 samples were multiplexed in one capture reaction. The capture was performed according to the manufacturer's instruction (Roche) and the hybridization reaction was incubated for 72 hours.

5.6.5 Illumina and Nanopore sequencing

For Nanopore sequencing the DNA concentration was quantified using the Qubit (Thermo Fisher) while for Illumina sequencing the DNA concentration was quantified using the KAPA Library Quantification Kit (Roche). The size of the library was determined on a Agilent Bioanalyzer using the Agilent High Sensitivity DNA kit. For all 5 different sequencing approaches, samples were pooled equimolarly and run on a single flow cell. Illumina sequencing was performed on an Illumina MiSeq to generate 2x300nt paired end sequences and Nanopore sequencing was performed on a GridION using R9.4 FLO-MIN106 flowcells with a run time of 16 hours.

5.6.6 Bioinformatic analysis

Nanopore sequences were demultiplexed using Porechop (https://github.com/rrwick/Porechop) after which the reads were trimmed to a median PHRED score of 10 and a minimal length of 150nt using fastp[162]. Illumina sequences were trimmed from the 3' end with a windowed approach and a mean PHRED score threshold of 20 using fastp[162]. Minimap2[221] and BWA-MEM[222] were used to map the Nanopore and Illumina sequence reads respectively to MN122189.1 (USUV), AY532665.1 (WNV), MK760665.1 (YFV) and KY348640.1 (ZIKV). After mapping primer sequences were clipped with python script using PySam[211] and reads with more than 10% softclipped nucleotides were removed from the alignment. The coverage statistics were determined using samtools's depth method[223]. A custom R script was used to generate the figures and determine the percentage of genome coverage above the coverage thresholds. For Nanopore sequencing the coverage threshold for reliable read coverage was set to 100x, as previously described[23], while for Illumina sequencing, because of its much higher read quality, the threshold was set to 5x read coverage. The complete workflow was written as a Snakemake[224] workflow which is, together with the custom python and R scripts, available at https://github.com/dnieuw/platform-comparison-arbovirus.

Author affiliations

[1] Viroscience Department, Erasmus Medical Center, Rotterdam, the Netherlands

[2] Departement of Virology of the Vaccination Programme, RIVM, Bilthoven, the Netherlands

* Corresponding author

Abbreviations

Ct: Cycle threshold

USUV: Usutu virus

WGS: Whole genome sequencing

WNV: West Nile virus

YFV: Yellow Fever virus

ZIKV: Zika virus

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: All read data are publically available under PRJEB47177, scripts to reproduce the results of this study are available at https://github.com/dnieuw/platform-comparison-arbovirus.

Competing interests: The authors declare that they have no competing interests.

Funding: COMPARE, VEO. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions: DFN, RSS, BBOM and MPGK conceptualized the study. AL and RK performed the sequencing. BBOM and DFN analyzed the data. BBOM and DFN wrote the manuscript. RSS, MPGK, BBM, AL and RK reviewed the manuscript.

Acknowledgements: Not Applicable

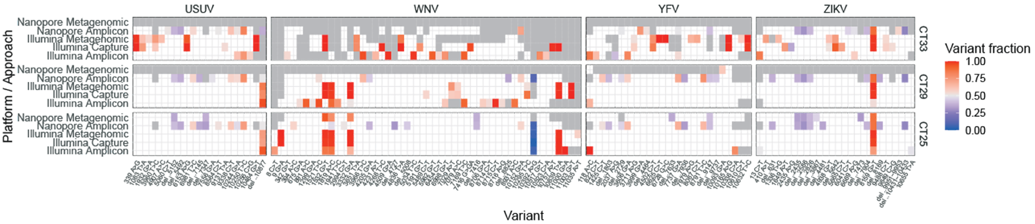


Figure S1 Overview of variants across the genomes before cleanup of read mapping. The x axis shows the variants found across the 4 genomes. The color scale represents the fraction of mapped reads containing the indicated variant in the read alignment. A darkgrey tile color indicates coverage below the coverage threshold (5x Illumina, 100x Nanopore).

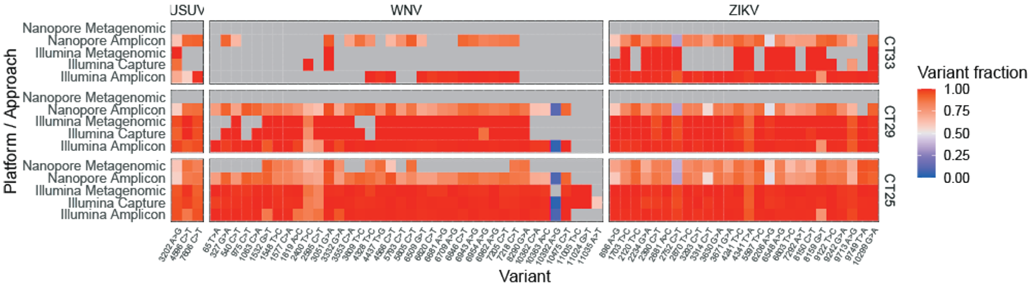
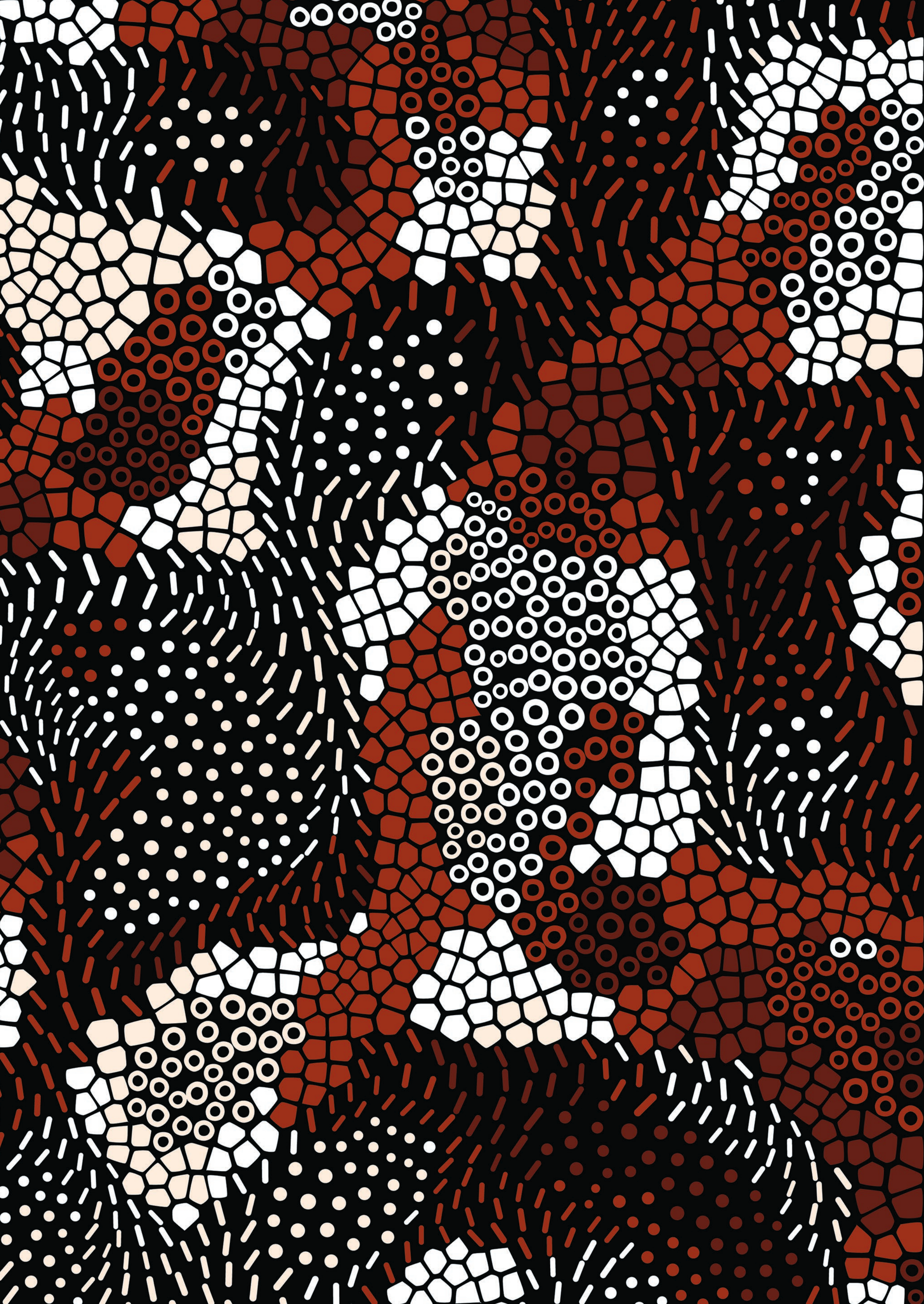


Figure S2 Overview of true variants across the genomes. The x axis shows the variants found across the 4 genomes. The color scale represents the fraction of mapped reads containing the indicated variant in the read alignment. A darkgrey tile color indicates coverage below the coverage threshold (5x Illumina, 100x Nanopore).



Validating whole genome nanopore sequencing, using Usutu virus as an example

Bas B. Oude Munnink[1], **David F. Nieuwenhuijse**[1], Reina S. Sikkema[1], Marion Koopmans[1]

Jove, e60906, 12, 2020

6

6.1 Abstract

Whole genome sequencing can be used to characterize and to trace viral outbreaks. Nanopore-based whole genome sequencing protocols have been described for several different viruses. These approaches utilize an overlapping amplicon-based approach which can be used to target a specific virus or group of genetically related viruses. In addition to confirmation of the virus presence, sequencing can be used for genomic epidemiology studies, to track viruses and unravel origins, reservoirs and modes of transmission. For such applications, it is crucial to understand possible effects of the error rate associated with the platform used. Routine application in clinical and public health settings require that this is documented with every important change in the protocol. Previously, a protocol for whole genome Usutu virus sequencing on the nanopore sequencing platform was validated (R9.4 flowcell) by direct comparison to Illumina sequencing. Here, we describe the method used to determine the required read coverage, using the comparison between the R10 flow cell and Illumina sequencing as an example.

Video link

The video component of this article can be found at https://www.jove.com/video/60906/

6.2 Introduction

Fast developments in third generation sequence technologies allows us to move forward towards close to real-time sequencing during viral outbreaks. This timely availability of genetic information can be useful to determine the origin and evolution of viral pathogens. Gold standards in the fields of next generation sequencing however, are still the second-generation sequencers. These techniques rely on specific and time- consuming techniques like clonal amplification during an emulsion PCR or clonal bridge amplification. The third-generation sequencers are cheaper, hand-held and come with simplified library preparation methodologies. Especially the small size of the sequence device and the low purchase price makes it an interesting candidate for deployable, fieldable sequencing. This could for instance be seen during the Ebola virus outbreak in Sierra Leone and during the ongoing arbovirus outbreak investigations in Brazil[194,217,225]. However, the reported high error rate[209] might limit the applications for which nanopore sequencing can be used. Nanopore sequencing is evolving quickly. New products are available in the market on a regular basis. Examples of this are for instance the 1D squared kits which enables sequencing of both strands of the DNA molecule, thereby boosting the accuracy of the called bases[226] and the development of the R10 flow cell which measures the change in current at two different instances in the pore[227]. In addition, improved bio informatic tools like improvements in basecalling will improve the accuracy of basecalling[228]. One of the most frequently used basecallers, (e.g., Albacore), has been updated at least 12 times in a 9-month time period[226]. Recently, the manufacturer also released a novel basecaller called flip- flop, which is implemented in the default nanopore software[229]. Together, all of these improvements will lead to more accurate sequences and will decrease the error rate of the nanopore sequencer.

Usutu virus (USUV) is a mosquito-borne arbovirus of the family Flaviviridae and it has a positive-stranded RNA genome of around 11,000 nucleotides. USUV mainly affects great grey owls and blackbirds[195,230], although other bird species are also susceptible to USUV infection[231]. Recently, USUV was also identified in rodents and shrews although their potential role in transmission of the virus remains unknown[232]. In humans, asymptomatic infections have been described in blood donors[233–236] while USUV infections also have been reported to be associated with encephalitis or meningo-encephalitis[237,238]. In the Netherlands, USUV was first detected in wild birds in 2016[195] and in asymptomatic blood donors in 2018[234]. Since the initial detection of USUV, outbreaks have been reported during the subsequent years and surveillance, including whole genome sequencing, is currently ongoing to monitor the emerge and spread of an arbovirus in a previously naïve population.

Similar to what has been described for other viruses, such as Ebola virus, Zika virus and yellow fever virus[98,207,217], we have developed a primer set to sequence full length USUV[23]. This polymerase chain reaction (PCR)-based approach allows for the recovery of full length USUV genomes from highly host-contaminated sample types like brain samples in samples up to a Ct value of around 32. Benefits of an amplicon-based sequencing approach are a higher sensitivity compared to metagenomic sequencing and a higher specificity. Limitations of using an amplicon-based approach are that the sequences should be similar in order to design primers fitting all strains and that primers are designed on our current knowledge about the virus diversity.

Given the constant developments and improvements in third generation sequencing, there is a need to evaluate the error rate of the sequencer on a regular basis. Here, we describe a method to evaluate the performance of nanopore directly against Illumina sequencing using USUV as an example. This method is applied to sequences generated with the latest R10 flow cell and basecalling is performed with the latest version of the flip-flop basecaller.

6.3 Protocol

NOTE: List of software tools to be used: usearch v11.0.667; muscle v3.8.1551; porechop 0.2.4; cutadapt 2.5; minimap2 2.16-r922; samtools 1.9; trimmomatic 0.39; bbmap 38.33; spades v3.13.1; kma-1.2.8

Primer design

1. Start with downloading or retrieving a set of relevant reference whole genome sequences from public or private data collections. For instance, retrieve all full length USUV genomes (taxid64286) from the NCBI database[166]. USUV encodes a genome of around 11,000 nucleotides so only retrieve the sequences with a sequence length of 8,000-12,000 nucleotides. Do this using the following search entry:
   - *taxid64286[Organism:noexp] AND 8000[SLEN]:12000[SLEN].*

   Click on **Send to | Complete Record | File**; use Format = FASTA and create the File.
2. To downsize the set of reference sequences, remove duplicate sequences or sequences with over 99% nucleotide identity from the dataset. Do this using the cluster fast option from usearch[106]. On the command line enter:
   - *usearch -cluster_fast All_USUV.fasta -id 0.99 -centroids All_USUV_dedup.fasta*

3. To generate the primers, sequences need to be aligned. This is done using MUSCLE[239]. On the command line enter:
   - *muscle -in All_USUV_dedup.fasta -out All_USUV_dedup_aligned.fasta -log log_muscle.txt*

   NOTE: It is essential to manually inspect the alignment to check for discrepancies. These can be manually corrected if needed and the ends can be trimmed according to the length of most whole genome sequences.
4. Primal is used to make a draft selection of the primers which can be used for full length amplicon sequencing[207]. Upload the alignment to the primal website (http://primal.zibraproject.org/) and select the preferred amplicon length and overlap length between the different amplicons. Go to primal.zibraproject.org, fill in the **Scheme name**, upload the aligned fasta file, select the amplicon length, overlap size, and generate the scheme.
5. Align the complete set of available complete USUV sequences (not the downsized or deduplicated set). On the command line enter:
   - *muscle -in All_USUV.fasta -out All_USUV_aligned.fasta -log log_muscle.txt*

   NOTE: Map the generated primers against the complete alignment (do not use the deduplicated alignment), manually correct errors and include a maximum of 5 degenerative primer positions.

Multiplex PCR

2. Perform the multiplex PCR using the designed primers and nanopore and Illumina sequencing. The multiplex PCR for USUV was performed as previous described[23,207].
3. Perform basecalling with flip-flop version 3.0.6.6+9999d81.

Data analysis to generate consensus sequences from nanopore data

4. Several samples can be multiplexed on a single nanopore sequencing run. After performing the sequence run, demultiplex the nanopore data. Use Porechop[240] for this. To prevent contamination and enhance accuracy, use the *require_two_barcodes* flag. On the command line enter:
   - *porechop -i Run_USUV.fastq -o Run_USUV_demultiplex --require_two_barcodes*
5. After demultiplexing, remove primer sequences (indicated in the file Primers_Usutu.fasta in both orientations) using cutadapt[241]. In addition, remove sequences with a length shorter than 75 nucleotides. The primers have to be removed since they can introduce artificial biases in the consensus sequence. On the command line enter:
   - *cutadapt -b file:Primers_USUV.fasta -o BC01_trimmed.fastq BC01.fastq -m 75*

6. Demultiplexed sequence reads can be mapped against a panel of distinct reference strains using minimap2[221] and a consensus sequence can be generated using samtools[223]. Follow the example below which shows the procedure of a reference-based alignment and the consensus sequence generation of one sample: BC01. On the command line enter:
   - *minimap2 -ax map-ont Random_Refs_USUV.fasta BC01_trimmed.fastq > BC01.bam*
   - *samtools sort BC01.bam > BC01_sorted.bam*
   - *bcftools mpileup -Ou -f Random_Refs_USUV.fasta BC01_sorted.bam | bcftools call -mv -Oz -o BC01.vcf.gz*
   - *bcftools index BC01.vcf.gz*
   - *cat Random_Refs_USUV.fasta | bcftools consensus BC01.vcf.gz > BC01_consensus.fasta*
7. For reference-based alignments it is essential that a closely related reference sequence is used. Therefore, perform a BlastN search with the generated consensus sequence to identify the closest reference strain. After that, repeat the reference-based alignment with the closest reference strain as reference (step 3.3 and 3.4). On the command line enter:
   - *minimap2 -ax map-ont Ref_USUV_BC01.fasta BC01_trimmed.fastq > BC01_ref.bam*
   - *samtools sort BC01_ref.bam > BC01_sorted_ref.bam*
   - *bcftools mpileup -Ou -f Ref_USUV_BC01.fasta BC01_sorted_ref.bam | bcftools call -mv -Oz -o BC01_ref.vcf.gz*
   - *bcftools index BC01_ref.vcf.gz*
   - *cat Ref_USUV_BC01.fasta | bcftools consensus BC01_ref.vcf.gz > BC01_ref_consensus.fasta*

Analysis of the Illumina data

8. These sequences are automatically demultiplexed after sequencing. Reads can be quality controlled using trimmomatic[242]. For paired-end Illumina sequences, use the commonly used cut-off median PHRED score of 33 and a minimal read length of 75 to get accurate, high quality reads. On the command line enter:
   - *trimmomatic PE -phred33 9_S9_L001_R1_001.fastq.gz 9_S9_L001_R2_001.fastq.gz 9_1P.fastq 9_1U.fastq 9_2P.fastq 9_2U.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:3:15 MINLEN:75*
9. Remove primers (indicated in the file Primers_Usutu.fasta in both orientations), since they can introduce artificial biases, using cutadapt[241]. In addition, remove sequences with a length shorter than 75 nucleotides using the commands below. On the command line enter:
   - *cutadapt -b o 9_1P_trimmed.fastq -p 9_2P_trimmed.fastq 9_1P.fastq 9_2P.fastq -m 75*

10. Before de novo assembly, the sequence reads can be normalized for an even coverage across the genome. This is essential since de novo assemblers like SPAdes take the read coverage into account when assembling sequence reads. Normalize reads to a read coverage of 50 using BBNorm from the BBMap package[243]. On the command line enter:
    - *bbmap/bbnorm.sh target=50 in=9_1P_trimmed.fastq in2=9_2P_trimmed.fastq out=Sample9_FW_norm.fastq out2=Sample9_RE_norm.fastq*
11. The normalized reads are de novo assembled using SPAdes[244]. Default settings are used for the assembly using all different kmers (21, 33, 55, 77, 99 and 127). On the command line enter:
    - *spades.py -k 21,33,55,77,99,127 -o Sample9 -1 Sample9.qc.f.fq -2 Sample9.qc.r.fq*
12. Map the QC reads against the obtained consensus sequence using minimap2 and programs like Geneious, Bioedit or Ugene to curate the alignment. It is important to check the beginning and the end of the contig.

1. Align the QC reads against the obtained consensus sequencing using minimap2.
2. Import the alignment in Geneious/Bioedit/UGene.
3. Manually inspect, correct and curate especially the beginning and the end of the genome.

Determining the required read coverage to compensate for the error profile in nanopore sequencing using Illumina data as gold standard

13. Select sequence reads mapping to one amplicon, in this case amplicon 26. Subsequently, map the nanopore reads against this amplicon using minimap2. Use Samtools to select only the reads mapping to amplicon 26 and to convert the bam file into fastq. On the command line enter:
    - *minimap2 -ax map-ont -m 150 Amplicon26.fasta BC01_trimmed.fastq > BC01.bam*
    - *samtools view -b -F 4 BC01.bam > BC01_mapped.bam*
    - *samtools bam2fq BC01_mapped.bam | seqtk seq - -> BC01_mapped.fastq*
14. Randomly select subsets of for instance 200 sequence reads one thousand times. For example, changing it to 10 will result in the random selection of one thousand times a subset of 10 sequence reads. The script is provided as **Supplementary File 1**. On the command line enter:
    - *python Random_selection.py*

15. All randomly selected sequence reads are aligned to amplicon 26. Use KMA[172] to map the sequence reads and to immediately generate a consensus sequence. Use optimized settings for nanopore sequencing, indicated by the -bcNano flag. On the command line enter:

```
– kma index -i Amplicon26.fasta
– for file in random_sample*; do
– sampleID=${file%.fastq}
– kma -i ${sampleID}.fastq -o ${sampleID} -t_db Amplicon26.fasta -mem_mode -mp 5 -mrs
  0.0 -bcNano
– done
```

16. Inspect the generated consensus sequences on the command line using:

```
– cat *.fsa > All_genomes.fsa
– minimap2 -ax map-ont Amplicon26.fasta All_genomes.fsa > All_genomes.bam
– samtools sort All_genomes.bam > All_genomes_sorted.bam
– samtools stats All_genomes_sorted.bam > stats.txt
```

1. The error rate is displayed in the stats.txt under the heading **error rate #mismatches / bases mapped**. Display it on the screen with the following command:

```
– grep ^SN stats.txt | cut -f 2-
```

2. The amount of indels is displayed under the heading **#Indels per cycle**. Display it on the screen with the following command:

```
– grep ^IC stats.txt | cut -f 2-
```

6.4 Representative Results

Recently, a new version of the flow cell version (R10) was released and offered improvements to the basecaller used to convert the electronic current signal to DNA sequences (so-called flip-flop basecaller). Therefore, we have re-sequenced USUV from brain tissue of an USUV-positive owl which was previously sequenced on a R9.4 flow cell and on an Illumina Miseq instrument[23]. Here, we described the method used to determine the required read coverage for reliable consensus calling by direct comparison to Illumina sequencing.

Using the newer flow cell in combination with the basecaller flip-flop we show that a read coverage of 40x results in identical results as compared to Illumina sequencing. A read coverage of 30x results in an error rate of 0.0002% which corresponds to one error in every 585,000 nucleotides sequenced, while a read coverage of 20x results in one error in every 63,529 nucleotides sequenced. A read coverage of 10x results in one error in every 3,312 nucleotides sequenced, meaning that over three nucleotides per full USUV genome are being called wrong. With a read coverage above 30x, no indels were observed. A read coverage of 20x resulted in the detection of one indel position while a read coverage of 10x resulted in indels in 29 positions. An overview of the error rate using different read coverage cut-offs is shown in **Table 1**.

Table 1 Overview of the error rate of nanopore sequencing. Each iteration represents one thousand random samples.

Coverage	Errors iteration 1	Error rate iteration 1	Indels:	Errors iteration 2	Error rate iteration 2	Indels:	Errors iteration 3	Error rate iteration 3	Indels:
10×	100	0.0274%	4	116	0.0297%	18	110	0.0282%	7
20×	4	0.0010%	0	6	0.0015%	1	7	0.0018%	0
30x	2	0.0005%	0	0	0.0000%	0	0	0.0000%	0
40x	0	0.0000%	0	0	0.0000%	0	0	0.0000%	0
50×	0	0.0000%	0	0	0.0000%	0	0	0.0000%	0

Supplementary File 1: Random selection.

6.5 Discussion

Nanopore sequencing is constantly evolving and therefore there is a need for methods to monitor the error rate. Here, we describe a workflow to monitor the error rate of the nanopore sequencer. This can be useful after the release of a new flow cell, or if new releases of the basecalling are released. However, this can also be useful for users who want to set-up and validate their own sequencing protocol.

Different software and alignment tools can yield different results[179]. In this manuscript, we aimed to use freely available software packages which are commonly used, and which have clear documentation. In some cases, preference might be given to commercial tools, which generally have a more user-friendly interfaces but have to be paid for. In the future, this method can be applied to the same sample in case big modifications in sequence technology or basecalling software are introduced Preferentially this should be done after each update of the basecaller or flowcell, however given the speed of the current developments this can be also been done only after major updates.

The reduction in the error rate in sequencing allows for a higher number of samples to be multiplexed. Thereby, nanopore sequencing is getting closer to replacing conventional real time PCRs for diagnostic assays, which is already the case for influenza virus diagnostics. In addition, the reduction of the error rate increases the usability of this technique sequencing, for instance for the determination of minor variants and for high- throughput unbiased metagenomic sequencing.

A critical step in the protocol is that close, reliable reference sequences need to be available. The primers are based on the current knowledge about virus diversity and might need to be updated every once in a while. Another critical point when setting up an amplicon-based sequencing approach is the balancing of the primer concentration to get an even balance in amplicon depth. This enables the multiplexing of more samples on a sequence run and results in a significant cost reduction.

Author affiliations

[1] ErasmusMC, Department of Viroscience, WHO Collaborating Centre for Arbovirus and Viral Hemorrhagic Fever Reference and Research, Erasmus University Medical Center

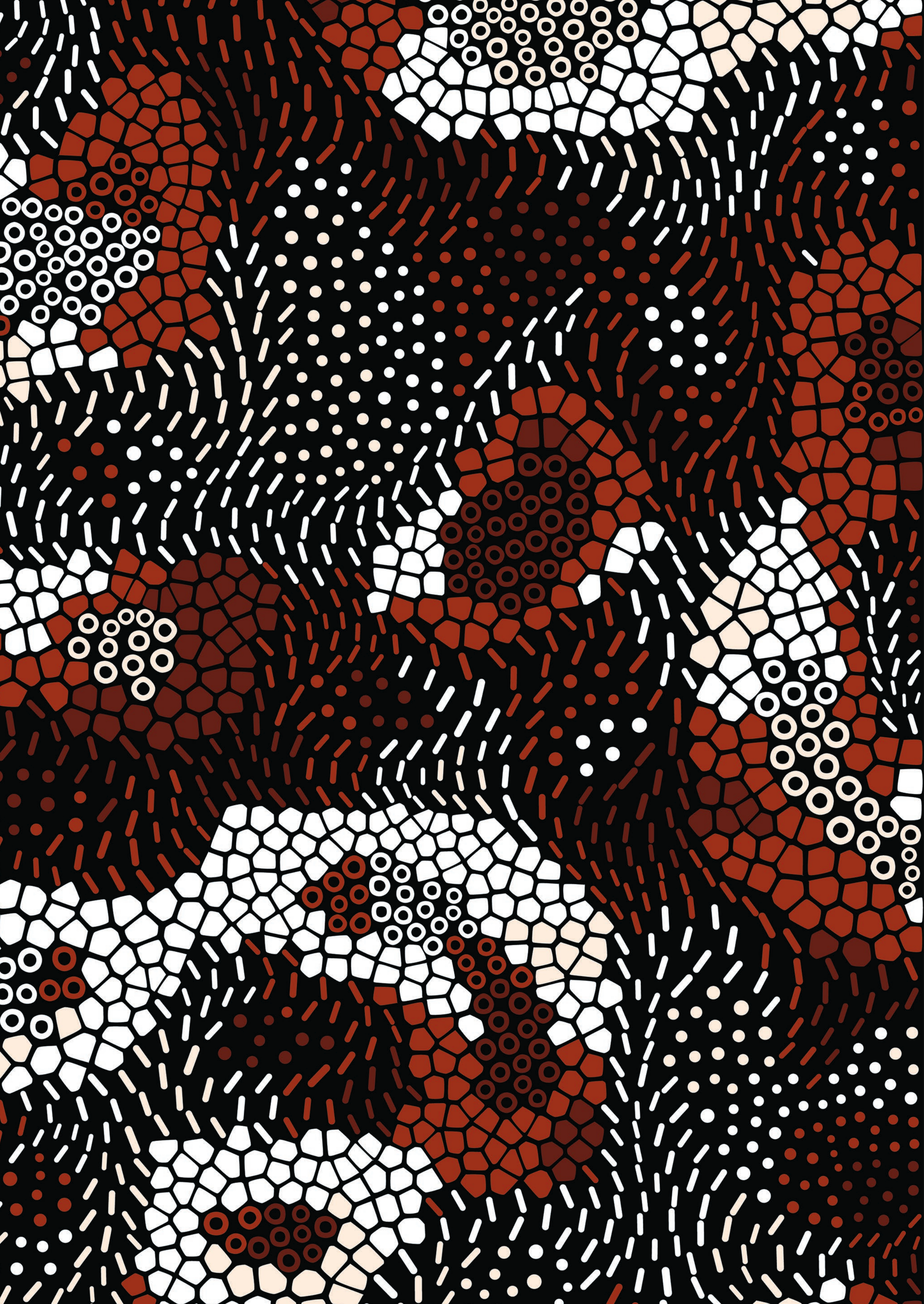
Correspondence to: Bas B. Oude Munnink at b.oudemunnink@erasmusmc.nl

Disclosures

The authors have nothing to disclose.

Acknowledgements

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 643476 (COMPARE).



Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands

Bas B. Oude Munnink[1], **David F. Nieuwenhuijse**1, Mart Stein[2], Áine O'Toole3, Manon Haverkate[2], Madelief Mollers[2], Sandra K. Kamga[2], Claudia Schapendonk[1], Mark Pronk[1], Pascal Lexmond[1], Anne van der Linden[1], Theo Bestebroer[1], Irina Chestakova[1], Ronald J. Overmars[1], Stefan van Nieuwkoop[1], Richard Molenkamp[1], Annemiek A. van der Eijk[1], Corine GeurtsvanKessel[1], Harry Vennema[2], Adam Meijer2, Andrew Rambaut[3], Jaap van Dissel[2], Reina S. Sikkema[1], Aura Timen[2],*, Marion Koopmans[1],* and The Dutch-Covid-19 response team

Nature medicine 26 (9), 1405-1410, 263, 2020

7

7.1 Abstract

In late December 2019, a cluster of cases of pneumonia of unknown etiology were reported linked to a market in Wuhan, China[245]. The causative agent was identified as the species *Severe acute respiratory syndrome-related coronavirus* and was named SARS-CoV-2[246]. By 16 April the virus had spread to 185 different countries, infected over 2,000,000 people and resulted in over 130,000 deaths[247]. In the Netherlands, the first case of SARS-CoV-2 was notified on 27 February. The outbreak started with several different introductory events from Italy, Austria, Germany and France followed by local amplification in, and later also outside, the south of the Netherlands. The combination of near to real-time whole-genome sequence analysis and epidemiology resulted in reliable assessments of the extent of SARS-CoV-2 transmission in the community, facilitating early decision-making to control local transmission of SARS-CoV-2 in the Netherlands. We demonstrate how these data were generated and analyzed, and how SARS-CoV-2 whole-genome sequencing, in combination with epidemiological data, was used to inform public health decision-making in the Netherlands.

7.2 Introduction

Whole-genome sequencing (WGS) is a powerful tool to understand the transmission dynamics of outbreaks and inform outbreak control decisions[14,98,217,248]. Evidence of this was seen during the 2014-2016 West African Ebola outbreak when real-time WGS was used to help public health decision-making, a strategy dubbed 'precision public health pathogen genomics'[249,250]. Immediate sharing and analysis of data during outbreaks is now recommended as an integral part of outbreak response[97,251,252]. Feasibility of real-time WGS requires access to sequence platforms that provide reliable sequences, access to metadata for interpretation, and data analysis at high speed and low cost. Therefore, WGS for outbreak support is an active area of research. Nanopore sequencing has been employed in recent outbreaks of Usutu, Ebola, Zika and yellow fever virus owing to the ease of use and relatively low start-up cost4–7. The robustness of this method has recently been validated using Usutu virus[23,210]. In the Netherlands, the first COVID-19 case was confirmed on 27 February and WGS was performed in near to real-time using an amplicon-based sequencing approach.

From 22 January, symptomatic travelers from countries where SARS-CoV-2 was known to circulate were routinely tested. The first case of SARS-CoV-2 infection in the Netherlands was identified on 27 February in a person with recent travel history to Italy and an additional case was identified one day later, also in a person with recent travel history to Italy. The genomes of these first two positive samples were generated and analyzed by 29 February. These two viruses clustered differently in the phylogenetic tree, confirming separate introductions (**Figure 1a**).

The advice to test hospitalized patients with serious respiratory infections was issued on 24 February and subsequent attempts to identify possible local transmission chains triggered testing for SARS-CoV-2 on a large scale in hospitals. By 9 March local clusters of epidemiologically related cases of SARS-CoV-2 started to appear in the province of Noord-Brabant. The increase in cases was caused by several co-circulating viruses, and is likely to have been triggered by multiple introductions of the virus following the spring holidays (from 13 to 23 February) with travel to ski resorts in Northern Italy (**Figure 1b**). The first intervention was put in place on 9 and 10 March when the prime minister advised people to stop shaking hands and events attended by more than 1,000 visitors were banned in the province of Noord-Brabant. Subsequent analysis identified clusters with local amplification of viruses from patients without any travel history, also outside Noord-Brabant (**Figure 2**). This information, combined with the increase in the total number

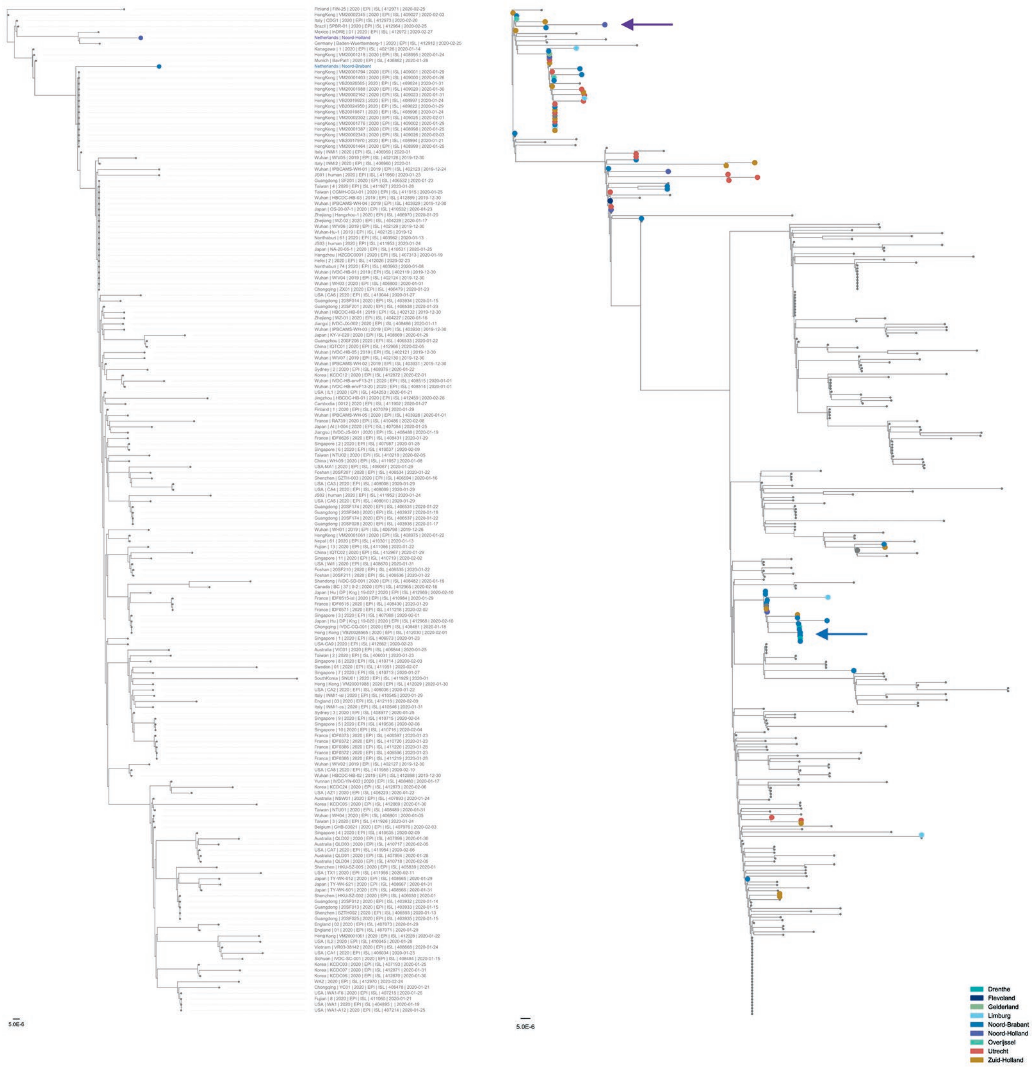


Figure 1 Phylogenetic analysis of the first two Dutch SARS-CoV-2 sequences. All sequences that were publicly available on 29 February (a) or 9 March (b) are included in the analysis. The sequences are colored on the basis of the province of detection. The scale bar represent the amount of nucleotide substitutions per site. Red indicates the Dutch isolates and blue represents SARS-CoV-2 sequences from other countries with recent travel history to Italy.

of infections in the Netherlands, led to the decision to implement stricter measures for the whole country to prevent further spread of SARS-CoV-2 on 12 March. All events with more than 100 people attending were canceled, people were requested to work from home as much as possible and people with symptoms such as a fever or cough had to stay at home. On 15 March, this was followed by the closure of schools, catering industries and sport clubs.

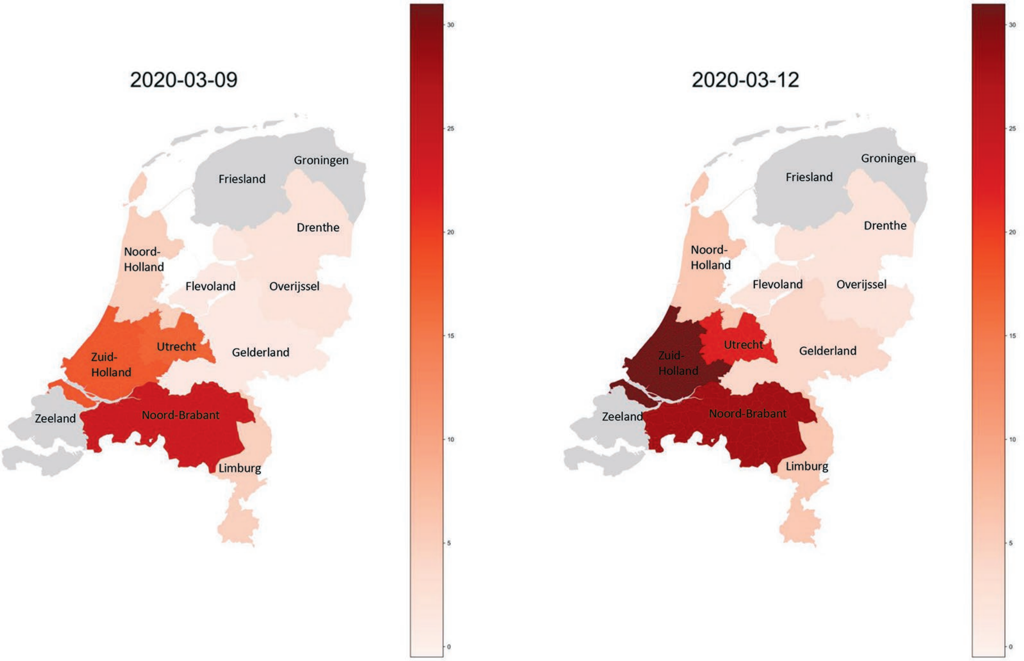


Figure 2 Distribution of SARS-CoV-2 sequences from the Netherlands on 9 and 12 March. The shapefile for the map is derived from https://gadm.org. The color scale represents the location and the number of whole-genome sequences generated at the indicated time points.

In the third phase, sequencing of new cases with emphasis on health-care workers (HCWs) and hospitalized cases was continued. By 15 March, 189 SARS-CoV-2 viruses from the Netherlands were sequenced, at that moment representing 27.1% of the total number of full genome sequences produced worldwide. The sequences detected in the Netherlands continued to be diverse and revealed the presence of multiple co-circulating sequence types, found in several different clusters in the phylogenetic tree (**Figure 3** and **Extended Data Figure 1**). This diversity was also observed in cases with similar travel histories, reflecting that sequence diversity was already present in the originating county, primarily Italy (**Figure 4**). In addition to travel-associated cases, an increasing number of local cases was detected through severe acute respiratory infection surveillance; this was not limited to the province Noord-Brabant but SARS-CoV-2 was also increasing in the provinces Zuid-Holland, Noord-Holland and Utrecht, confirming substantial under-ascertainment of the epidemic. The increase in the number of patients with COVID-19 as well as increasing affected geographic areas and occurrence of local clusters provided further support for the increased movement restrictions.

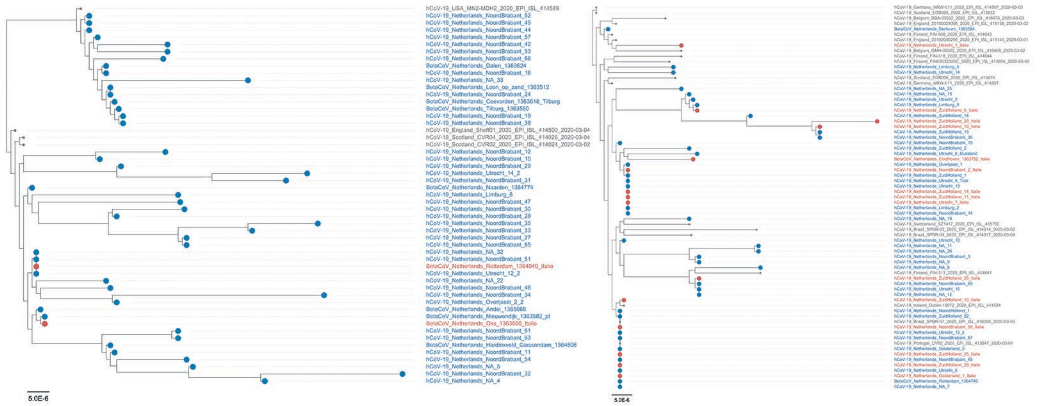

hCoV-19_USA_MN2-MDH2_2020_EPI_ISL_414589
hCoV-19_Netherlands_NoordBrabant_52
hCoV-19_Netherlands_NoordBrabant_49
hCoV-19_Netherlands_NoordBrabant_44
hCoV-19_Netherlands_NoordBrabant_57
hCoV-19_Netherlands_NoordBrabant_42
hCoV-19_Netherlands_NoordBrabant_53
hCoV-19_Netherlands_NoordBrabant_66
BetaCoV_Netherlands_Dalen_1363624
hCoV-19_Netherlands_NoordBrabant_18
hCoV-19_Netherlands_NA_33
BetaCoV_Netherlands_Loon_op_zand_1363512
hCoV-19_Netherlands_NoordBrabant_24
BetaCoV_Netherlands_Coevorden_1363618_Tilburg
BetaCoV_Netherlands_Tilburg_1363550
hCoV-19_Netherlands_NoordBrabant_19
hCoV-19_Netherlands_NoordBrabant_38
hCoV-19_England_Sheff01_2020_EPI_ISL_414500_2020-03-04
hCoV-19_Scotland_CVR04_2020_EPI_ISL_414026_2020-03-04
hCoV-19_Scotland_CVR02_2020_EPI_ISL_414024_2020-03-02
hCoV-19_Netherlands_NoordBrabant_12
hCoV-19_Netherlands_NoordBrabant_10
hCoV-19_Netherlands_NoordBrabant_29
hCoV-19_Netherlands_Utrecht_14_2
hCoV-19_Netherlands_NoordBrabant_31
BetaCoV_Netherlands_Naarden_1364774
hCoV-19_Netherlands_Limburg_6
hCoV-19_Netherlands_NoordBrabant_47
hCoV-19_Netherlands_NoordBrabant_30
hCoV-19_Netherlands_NoordBrabant_28
hCoV-19_Netherlands_NoordBrabant_35
hCoV-19_Netherlands_NoordBrabant_33
hCoV-19_Netherlands_NoordBrabant_27
hCoV-19_Netherlands_NoordBrabant_65
hCoV-19_Netherlands_NA_32
hCoV-19_Netherlands_NoordBrabant_51
BetaCoV_Netherlands_Rotterdam_1364040_Italie
hCoV-19_Netherlands_Utrecht_12_2
hCoV-19_Netherlands_NA_22
hCoV-19_Netherlands_NoordBrabant_48
hCoV-19_Netherlands_NoordBrabant_34
hCoV-19_Netherlands_Overijssel_2_2
BetaCoV_Netherlands_Andel_1365066
BetaCoV_Netherlands_Nieuwendijk_1363582_pt
BetaCoV_Netherlands_Oss_1363500_Italie
hCoV-19_Netherlands_NoordBrabant_61
hCoV-19_Netherlands_NoordBrabant_63
BetaCoV_Netherlands_Hardinxveld_Giessendam_1364806
hCoV-19_Netherlands_NoordBrabant_11
hCoV-19_Netherlands_NoordBrabant_54
hCoV-19_Netherlands_NA_5
hCoV-19_Netherlands_NoordBrabant_32
hCoV-19_Netherlands_NA_4
5.0E-6
hCoV-19_Germany_NRW-011_2020_EPI_ISL_414507_2020-03-03
hCoV-19_Scotland_EDB005_2020_EPI_ISL_415632
hCoV-19_Belgium_DBA-03032_2020_EPI_ISL_416475_2020-03-03
hCoV-19_England_20100024006_2020_EPI_ISL_415138_2020-03-02
BetaCoV_Netherlands_Berlicum_1363564
hCoV-19_Finland_FIN-508_2020_EPI_ISL_414643
hCoV-19_England_20102000206_2020_EPI_ISL_415145_2020-03-01
hCoV-19_Netherlands_Utrecht_1_Italie
hCoV-19_Belgium_GMH-03022_2020_EPI_ISL_416468_2020-03-02
hCoV-19_Finland_FIN-318_2020_EPI_ISL_414644
hCoV-19_Finland_FIN03032020C_2020_EPI_ISL_413604_2020-03-03
hCoV-19_Netherlands_Limburg_5
hCoV-19_Netherlands_Utrecht_14
hCoV-19_Scotland_EDB006_2020_EPI_ISL_415633
hCoV-19_Germany_NRW-011_2020_EPI_ISL_414507
hCoV-19_Netherlands_NA_25
hCoV-19_Netherlands_NA_15
hCoV-19_Netherlands_Utrecht_2
hCoV-19_Netherlands_Limburg_3
hCoV-19_Netherlands_ZuidHolland_6_Italie
hCoV-19_Netherlands_ZuidHolland_18
hCoV-19_Netherlands_ZuidHolland_20_Italie
hCoV-19_Netherlands_ZuidHolland_19_Italie
hCoV-19_Netherlands_ZuidHolland_15
hCoV-19_Netherlands_NoordBrabant_36
hCoV-19_Netherlands_NoordBrabant_15
hCoV-19_Netherlands_ZuidHolland_2
hCoV-19_Netherlands_Utrecht_6_Duitsland
BetaCoV_Netherlands_Eindhoven_1363782_Italie
hCoV-19_Netherlands_Overijssel_1
hCoV-19_Netherlands_NoordBrabant_2_Italie
hCoV-19_Netherlands_ZuidHolland_1
hCoV-19_Netherlands_Utrecht_4_Tirol
hCoV-19_Netherlands_Utrecht_13
hCoV-19_Netherlands_ZuidHolland_14_Italie
hCoV-19_Netherlands_ZuidHolland_11_Italie
hCoV-19_Netherlands_Utrecht_7_Italie
hCoV-19_Netherlands_Limburg_2
hCoV-19_Netherlands_NoordBrabant_14
hCoV-19_Netherlands_NA_19
hCoV-19_Switzerland_SZ1417_2020_EPI_ISL_415702
hCoV-19_Brazil_SPBR-03_2020_EPI_ISL_414014_2020-03-02
hCoV-19_Brazil_SPBR-04_2020_EPI_ISL_414017_2020-03-04
hCoV-19_Netherlands_Utrecht_10
hCoV-19_Netherlands_NA_11
hCoV-19_Netherlands_NA_26
hCoV-19_Netherlands_NoordBrabant_3
hCoV-19_Netherlands_NA_9
hCoV-19_Netherlands_NA_8
hCoV-19_Finland_FIN-313_2020_EPI_ISL_414641
hCoV-19_Netherlands_ZuidHolland_26_Italie
hCoV-19_Netherlands_NoordBrabant_55
hCoV-19_Netherlands_Utrecht_15
hCoV-19_Netherlands_NA_12
hCoV-19_Netherlands_ZuidHolland_10_Italie
hCoV-19_Ireland_Dublin-19072_2020_EPI_ISL_414584
hCoV-19_Netherlands_NoordHolland_1
hCoV-19_Netherlands_ZuidHolland_22
hCoV-19_Brazil_SPBR-07_2020_EPI_ISL_416028_2020-03-03
hCoV-19_Netherlands_NoordBrabant_58_Italie
hCoV-19_Netherlands_Utrecht_13_2
hCoV-19_Netherlands_NoordBrabant_67
hCoV-19_Portugal_CV62_2020_EPI_ISL_413647_2020-03-01
hCoV-19_Netherlands_Gelderland_3
hCoV-19_Netherlands_ZuidHolland_25_Italie
hCoV-19_Netherlands_NoordBrabant_45
hCoV-19_Netherlands_ZuidHolland_29_Italie
hCoV-19_Netherlands_Utrecht_5
hCoV-19_Netherlands_Gelderland_1_Italie
BetaCoV_Netherlands_Rotterdam_1364740
hCoV-19_Netherlands_NA_7
5.0E-6

Figure 3 Phylogenetic analysis of SARS-CoV-2 emergence in the Netherlands. Zoom-ins of two clusters circulating in the Netherlands. The sequences are colored on the basis of travel history; Dutch patients without travel history are indicated in blue while Dutch patients with travel history to Italy are indicated in red. The scale bars represent the number of substitutions per site.

BEAST analysis revealed that the most recent ancestor of the viruses circulating in the Netherlands dates back to the end of January and the beginning of February (**Figure 4**). This is in line with the amplification that occurred in the region (notably Italy and Austria) from which most of the epidemic in the Netherlands was seeded. Most incursions likely occurred during spring break, which is a popular time for winter sports vacations. Retrospective testing showed the presence of the virus in a sample collected on 24 February in a patient with known travel history to Italy.

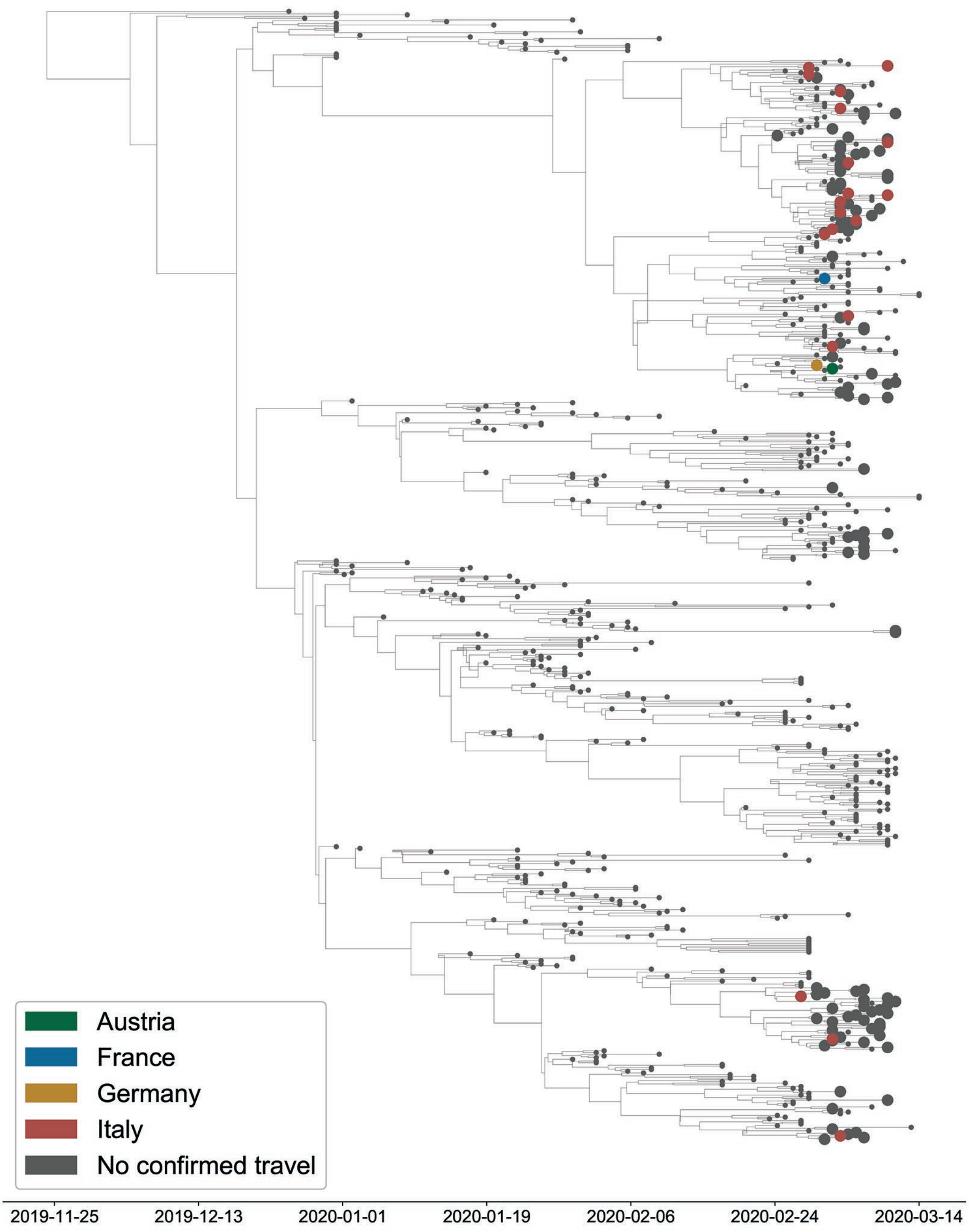


Figure 4 BEAST analysis with travel history. Time-resolved visualization of the emergence of SARS-CoV-2 in the Netherlands. Sequences from the Netherlands are depicted with big circles. Green indicates recent travel history to Austria, blue to France, yellow to Germany, and red to Italy.

In this study, we show that WGS in combination with epidemiological data strengthened the evidence base for public health decision-making in the Netherlands as it enabled a more precise understanding of the transmission patterns in various initial phases of the outbreaks. As such, we were able to understand the genetic diversity of the multiple introduction events in phase 1, the extent of local and regional clusters in phase 2 and the transmission patterns within the HCW groups in phase 3 (among which the absence or occurrence of very limited nosocomial transmission). This information complemented the data obtained from more traditional methods such as contact investigation.

At the time of the study, sequences from the Netherlands made up a substantial part of the total collection of SARS-CoV-2 genomes. Although implementation of WGS in the Dutch disease prevention and control strategy has shown its added value, there were limitations due to the paucity of genomic information available from certain parts of the world, including Italy. The information available from Iran, another major country where the virus was presumably spreading exponentially in the week before the take-off of the epidemic in the Netherlands, was also limited. This sampling bias needs to be considered when drawing conclusions based on genomic data during early stages of an emerging disease outbreak. Without a representative and sizable selection of reference sequences, reliable phylogenetic analysis is difficult. Clustering and conclusions on the origin of viruses may change substantially when virus sequences of other geographical regions are added to the analysis. Moreover, global monitoring of the genetic diversity of the virus is essential to reliably model and predict the spread of the virus. Since early March, the number of publicly available genomes has grown considerably, and the geographic signature in the dataset is becoming increasingly clear. Since its emergence, the global spread of SARS-CoV-2 led to diversification into lineages that reflect ongoing chains of transmission in specific geographic regions globally, in Europe, and – during the second and third phases – in the Netherlands. The average single nucleotide polymorphism distance between the sequenced viruses in our study was 7.39 and this diversification provided the basis for the use of WGS to investigate possible transmission chains locally (for instance, in health-care settings, where it can be used to inform infection control and prevention when combined with background data on contact histories among others). Moreover, the continued effort will lay the foundation for the enhanced surveillance that will be paramount during the next phase of the pandemic, when confinement measures will gradually be lifted and testing of people with mild symptoms is increased. Given the widespread circulation, the most likely scenario is that SARS-CoV-2 will (sporadically) re-emerge, and discrimination between novel introductions versus prolonged local circulation is important to inform appropriate public health decisions. In addition, owing to genomic mutations, the phenotype and the transmission dynamics of the virus might change over time. Therefore, close monitoring of the behavior of the virus in combination with genetic information is essential as well. We have used an amplicon-based sequencing approach to monitor the emergence of SARS-CoV-2 in the Netherlands. A critical step in using amplicon-based sequencing is that close, reliable reference sequences need to be available. The primers are designed on the basis of our current knowledge about SARS-CoV-2 diversity and therefore need regular updating. In the future, this may be overcome using metagenomic sequencing. However, at the moment, conventional metagenomic sequencing (Illumina) takes too long for near to real-time sequencing, and nanopore-based metagenomic sequencing is not sensitive enough to allow recovery of whole-genome sequences in a similar fashion and with similar costs compared to amplicon-based nanopore sequencing. We provide a description of the incursion of SARS-CoV-2 into the Netherlands. The combination of real-time WGS with the data from the National Public Health response team has provided information that helped decide on the next steps in the decision-making. Sharing of metadata is needed within a country but also on a global level. We urge countries to share sequence information to combine our efforts in understanding the spread of SARS-CoV-2. The Global Initiative on Sharing All Influenza Data (GISAID)[253,254] made sharing of sequence information coupled to limited metadata possible in a manner that protects the intellectual property and acknowledges the data providers. However, to fully capitalize on the potential added value of WGS for public health decision-making, systems for combined analysis of data are needed that are in agreement with general data protection rules. We previously developed a model for collaborative exploration of WGS and metadata in a protected sharing environment[193,255]. For truly global collaboration, such systems would need to be further developed and hosted under the auspices of the WHO (World Health Organization).

Author affiliations

[1]ErasmusMC, Department of Viroscience, WHO Collaborating Centre for Arbovirus and Viral Hemorrhagic Fever Reference and Research, Rotterdam, the Netherlands. [2]Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, the Netherlands. [3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.
*These authors contributed equally: Aura Timen, Marion Koopmans.

The Dutch-Covid-19 response team

G. J. A. P. M. Oudehuis[4], Janke Schinkel[5], Jan Kluytmans[6,7], Marjolein Kluytmans-van den Bergh[6,7], Wouter van den Bijllaardt[6], Robbert G. Berntvelsen[6], Miranda M. L. van Rijen[6], Peter Schneeberger[8], Suzan Pas[9], Bram M. Diederen[9], Anneke M. C. Bergmans[9], P. A. Verspui van der Eijk[10], Jaco J. Verweij[7], Anton G. N. Buiting[7], Roel Streefkerk[11], A. P. Aldenkamp[12], P. de Man[13], J. G. M. Koelemal[13], D. Ong[13], S. Paltansing[13], N. Veassen[13], Jacqueline Sleven[14], Leendert Bakker[15], Heinrich Brockhoff[15], Ariene Rietveld[16], Fred Slijkerman Megelink[17], James Cohen Stuart[17], Anne de Vries[18], Wil van der Reijden[18], A. Ros[18], Esther Lodder[19], Ellen Verspui-van der Eijk[20], Inge Huijskens[20], E. M. Kraan[21], M. P. M. van der Linden[21], S. B. Debast[22], N. Al Naiemi[23], A. C. M. Kroes[24], Marjolein Damen[25], Sander Dinant[25], Sybren Lekkerkerk[25], Oscar Pontesilli[25], Pieter Smit[25], Carla van Tienen[25], P. C. R. Godschalk[26], Jorien van Pelt[27], Alewijn Ott[27], Charlie van der Weijden[28], Heiman Wertheim[29], Janette Rahamat-Langendoen[29], Johan Reimerink[30], Rogier Bodewes[30], Erwin Duizer[30], Bas van der Veer[30], Chantal Reusken[30], Suzanne Lutgens[31], Peter Schneeberger[31], Mirjam Hermans[31], P. Wever[31], A. Leenders[31], Henriette ter Waarbeek[32] and Christian Hoebe[32]

[4]Academic Hospital Maastricht, Maastricht, the Netherlands. [5]Amsterdam Medical Centre, Amsterdam, the Netherlands. [6]Amphia Hospital, Breda, the Netherlands. [7]Bernhoven Hospital, Uden, the Netherlands. [8]Bravis Ziekenhuis, Bergen op Zoom & Roosendaal, Roosendaal, the Netherlands. [9]Dienst Gezondheid & Jeugd Zuid-Holland Zuid, Dordrecht, the Netherlands. [10]Elisabeth-Tweesteden Hospital, Tilburg, the Netherlands. [11]RLM, Dordecht, the Netherlands. [12]Foundation PAMM, Eindhoven, the Netherlands. [13]Franciscus Gasthuis & Vlietland, Rotterdam, the Netherlands. [14]MHC Gooi & Vechtstreek, Bussum, the Netherlands. [15]MHC Haaglanden, The Hague, the Netherlands. [16]MHC Hart voor Brabant, Tilburg, the Netherlands. [17]MHC Holland Noorden, Alkmaar, the Netherlands. [18]MHC Kennemerland, Haarlem, the Netherlands. [19]MHC West-Brabant, Breda, the Netherlands. [20]MHC Zuid-Holland Zuid, Dordrecht, the Netherlands. [21]Ijsselland Hospital, Capelle aan den IJssel, the Netherlands. [22]Isala Hospital, Zwolle, the Netherlands. [23]Laboratorium Microbiologie Twente Achterhoek, Hengelo, the Netherlands. [24]Leids Universitair Medisch Centrum, Leiden, the Netherlands. [25]Maasstad Hospital, Rotterdam, the Netherlands. [26]Meander Medical Centre, Amersfoort, the Netherlands. [27]MHC Drente, Meppel, the Netherlands. [28]MHC Flevoland, Lelystad, the Netherlands. [29]Radboud University Medical Center, Nijmegen, the Netherlands. [30]Centre for Infectious Disease Control, Bilthoven, the Netherlands. [31]Foundation Jeroen Bosch Hospital's, Hertogenbosch, the Netherlands. [32]Nursing Home Maastricht, Maastricht, the Netherlands.

7.3 Methods

7.3.1 COVID-19 response

This study was carried out in liaison with the national outbreak response team. This team develops guidance on case-finding and containment, based on WHO and European Centre for Disease Prevention and Control recommendations and expert advice, as defined by the crisis and emergency response structure[256,257]. Diagnostics were initially performed on suspected cases with a recent travel history to China, but between 25 and 28 February also suspected cases with travel history to affected municipalities in Northern Italy were tested. Between 1 and 11 March, all suspected cases with travel history to all four provinces in Northern Italy were tested and after 11 March all suspected cases with travel history to Italy were tested. The sequencing effort was embedded in the stepwise response to the outbreak (**Extended Data Figure 2**), which evolved from the initial testing of symptomatic travelers including the testing of symptomatic contacts (phase 1), followed by inclusion of routine testing of patients hospitalized with severe respiratory infections (phase 2), to inclusion of HCWs with a low-threshold case definition and testing to define the extent of suspected clusters (phase 3). Depending on the phase and clinical severity, initial contact with patients was established through public health physicians or nurses from the municipal health service (for travel-related cases, contacts of (hospitalized) cases, and patients belonging to risk groups). The different phases in this study were based on observations described in this manuscript. Ethical approval was not required for this study as only anonymous aggregated data were used, and no medical interventions were made on human individuals.

7.3.2 Contact tracing

On 29 January, COVID-19 was classified as a notifiable disease in group A in the Netherlands, with physicians and laboratories having to report any suspected and confirmed case to the Dutch public health services (PHS) by phone. On notification, the PHS initiates source identification and contact tracing, and performs risk assessments. In the early outbreak phase (containment), the PHS traced and informed all high- and low-risk contacts of cases with the aim to stop further transmission. For each case, epidemiological information such as demographic information, symptoms, date of onset of symptoms, travel history, contact information, suspected source, underlying disease and occupation were registered. People were asked to report their travel history for the past 14 days, including potential travel to several countries. Owing to the magnitude of the COVID-19 outbreak, this quickly became impracticable in severely affected regions, and the strategy shifted to registering only data on confirmed cases and informing their high-risk contacts (phase 2) with continued active case-finding in less affected regions. The PHS informed the national public health authority of the Netherlands (RIVM) about all laboratory-confirmed cases. There, a national case registry was kept in which a contact matrix was kept for the first 250 cases.

7.3.3 Sample selection

In the first phase, all samples were selected for sequencing, reflecting travel-associated cases and their contacts. In the second phase, priority was given to patients identified through enhanced case-finding by testing of hospitalized patients with severe acute respiratory infections and continued sequencing of new incursions. In the third phase, the epidemic started to expand exponentially, and sequencing was performed to continue to monitor the evolution of the outbreak. In line with the national testing policy, a substantial proportion of new cases sequenced were HCWs (20%).

7.3.4 SARS-CoV-2 diagnostics

Clinical specimens were collected and phocine distemper virus was added as an internal nucleic acid (NA) extraction control to the supernatant. Clinical specimens included oropharyngeal and nasopharyngeal swabs, bronchoalveolar lavage and sputum. Total NA was extracted from the supernatant using Roche MagNA Pure systems. The NA was screened for the presence of SARS-CoV-2 using real-time single-plex PCRs with reverse transcription for phocine distemper virus, for the SARS-CoV-2 RdRp gene and for the SARS-CoV-2 E gene as described by Corman et al.[258].

SARS-CoV-2 WGS. A SARS-CoV-2-specific multiplex PCR for nanopore sequencing was performed, similar to amplicon-based approaches as previously described22. In short, primers for 89 overlapping amplicons spanning the entire genome were designed using primal (http://primal.zibraproject.org/)[207]. The amplicon length was set to 500 base pairs with a 75-base-pair overlap between the different amplicons. The used concentrations and primer sequences are shown in **Supplementary Table 1**. The libraries were generated using the native barcode kits from Nanopore (EXP-NBD104, EXP-NBD114 and SQK-LSK109) and sequenced on a R9.4 flow cell multiplexing up to 24 samples per sequence run.

7.3.5 Sequence data analysis

The resulting raw sequence data were demultiplexed using qcat (https://github.com/nanoporetech/qcat) or Porechop (https://github. com/rrwick/Porechop). Primers were trimmed using cutadapt[241], after which a reference-based alignment was performed using minimap2[221] to the GISAID sequence EPI_ISL_412973. The run was monitored using RAMPART (https:// artic-network.github.io/rampart/) and the analysis process was automated using snakemake[224], which was used to perform near to real-time analysis with new data every 10 min. The consensus genome was extracted and positions with a coverage <30 were replaced with an 'N' with a custom script using biopython and pysam (https://github.com/dnieuw/ENA_SARS_Cov2_nanopore). An overview of the success rate of the sequencing is shown in **Supplementary Table 2**. Mutations in the genome as compared to the GISAID sequence EPI_ISL_412973 were confirmed by manually checking the alignment. In addition, homopolymeric regions were manually checked and resolved by consulting reference genomes. The average single nucleotide polymorphism difference was determined using snp-dists (https://github.com/tseemann/snp-dists). Human reads were removed by mapping against the human genome (GCF_000001405.26), after which the demultiplexed sequence reads were uploaded to the COVID-19 data portal under the accession numbers ERR4164763–ERR4164952.

7.3.6 Phylogenetic analysis

All available full-length SARS-CoV-2 genomes were retrieved from GISAID on 22 March 2020 (**Supplementary Table 3**) and aligned with the Dutch SARS-CoV-2 sequences from this study using MUSCLE. Sequences with >10% 'N's were excluded. The alignment was manually checked for discrepancies, after which IQ-TREE[259] was used to perform a maximum-likelihood phylogenetic analysis under the GTR + F + I + G4 model as the best predicted model using the ultrafast bootstrap option with 1,000 replicates. The phylogenetic trees were visualized using custom python and baltic scripts (https://github.com/evogytis/baltic).

7.3.7 BEAST analysis

All available full-length SARS-CoV-2 genomes were retrieved from GISAID[253,254] on 18 March 2020 and downsampled to include only representative sequences from epidemiologically linked cases. Sequences lacking date information were also removed from the dataset. To assess the temporal signal within the data, a maximum-likelihood phylogeny was performed using IQTREE v1.6.8[260] and the root-to-tip divergence was visualized as a function of sample date using TempEst v1.5.1[261] (**Extended Data Figure 3**). The correlation coefficient for the root-to-tip analysis was 0.53, which is adequate for subsequent Bayesian analysis as much of this noise is accounted for in the Bayesian model. Bayesian phylogenetic trees were estimated using BEAST v1.10.4[262,263] using an HKY nucleotide substitution model and a strict molecular clock[264]. The analysis was run for 100,000,000 states with an exponential growth prior. Every 10,000 states, trees and parameters were sampled. Log files were inspected in Tracer v1.7.1[265] and Tree annotator v1.10.0 was used to remove the burn-in from the tree files and to infer the maximum clade credibility tree. Reported statistics are shown in Supplementary Table 4. Baltic and custom python scripts (https://github.com/evogytis/baltic) were used to visualize the maximum clade credibility tree.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data produced in this study are available on the COVID-19 data portal under the accession numbers ERR4164763–ERR4164952 and on the GISAID portal under the accession numbers EPI_ISL_413564-EPI_ISL_413591, EPI_ISL_414423-EPI_ISL_414471, EPI_ISL_414529-EPI_ISL_414566 and EPI_ISL_415460-EPI_ISL_415535.

Acknowledgements

We thank the following additional members of the Dutch-Covid-19 response team: staff of the Department of Communicable Disease Control, Public Health Service Amsterdam, MHC Amsterdam, Amsterdam; staff of the Department of Communicable Disease Control, Public Health Service Gelderland-Zuid, MHC Gelderland-Zuid, Tiel; the Laboratorium Medische Microbiologie of JBZ, MHC Hart voor Brabant, Tilburg; staff of the Department of Communicable Disease Control, Public Health Service Utrecht, MHC Regio Utrecht, Utrecht; staff of the Department of Communicable Disease Control, Public Health Service Rotterdam-Rijnmond, MHC Rotterdam-Rijnmond, Rotterdam; staff of the Department of Communicable Disease Control, Public Health Service Zaanstreek-Waterland, MHC Zaanstreek Waterland, Zaandam; staff of the Department of Communicable Disease Control, Public Health Service Brabant Zuidoost, MHC Brabant Zuidoost, Eindhoven. We gratefully acknowledge the originating laboratories, where specimens were first obtained, and the submitting laboratories, where sequence data were generated and submitted to the EpiFlu Database of the GISAID, on which this research is based. All contributors of data may be contacted directly via the GISAID website (http://platform.gisaid.org). We also acknowledge M. de Graaf for her expertise in the interpretation of the early phylogenetic analysis. B.B.O.M., R.S.S., D.F.N., R.M. and M.K. received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 874735 (VEO), 848096 (SHARP JA) and 101003589 (RECoVER) and C.S., B.B.O.M. and M.K. from the European Joint Programme One Health EJP under the grant agreement number 773830 (METASTAVA). Part of this work was funded by the NIAID/NIH contract HHSN272201400008C for T.B., P.L. and M.P.

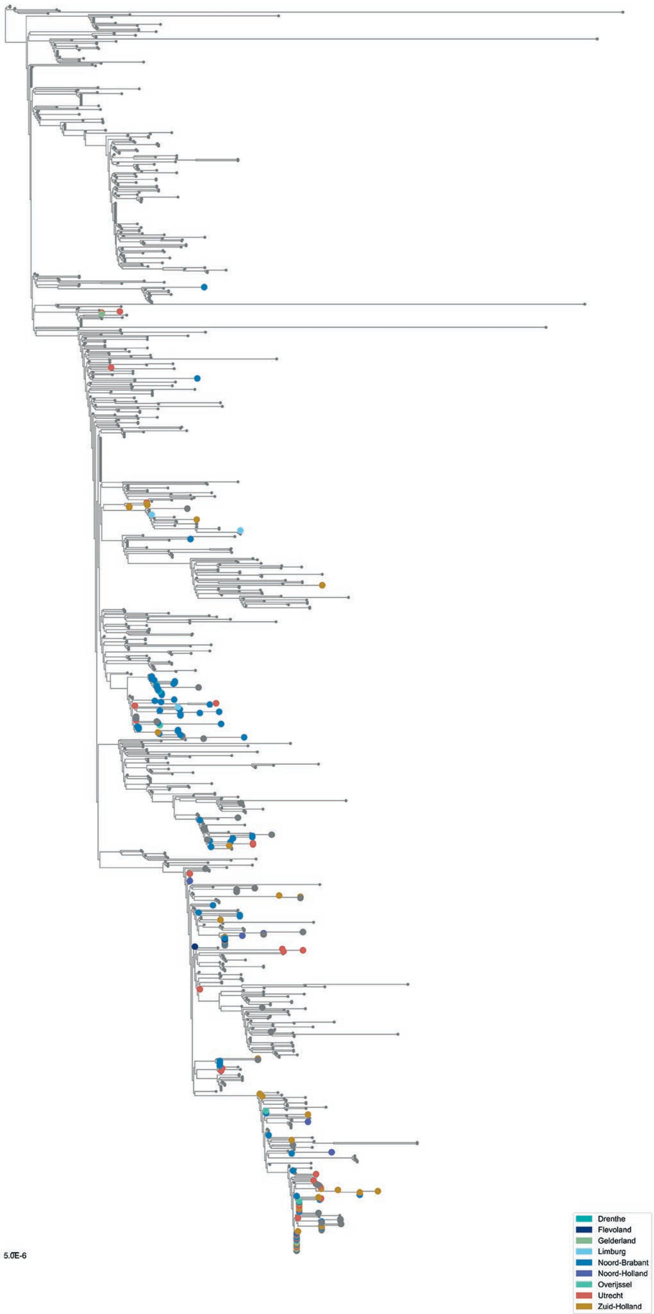
Author contributions

B.B.O.M., R.S.S., Á.O'T. and M.K. wrote the manuscript, Á.O'T., M.S., M.H. and M.M. set up sample and data collection, B.B.O.M, A.v.d.L., I.C., M.P., P.L., S.v.N., T.B., C.S. and R.J.O generated sequence data, S.K.K., R.M., A.A.v.d.E. and C.G. were involved in sample and data collection, B.B.O.M., R.R.S., D.F.N., A.R., A.M., H.V., A.O., Á.O'T., J.v.D. and M.K. were involved in data analysis and interpretation, B.B.O.M., M.S., M.H., M.M., R.R.S., Á.O'T. and M.K. designed the study. All authors provided critical feedback.

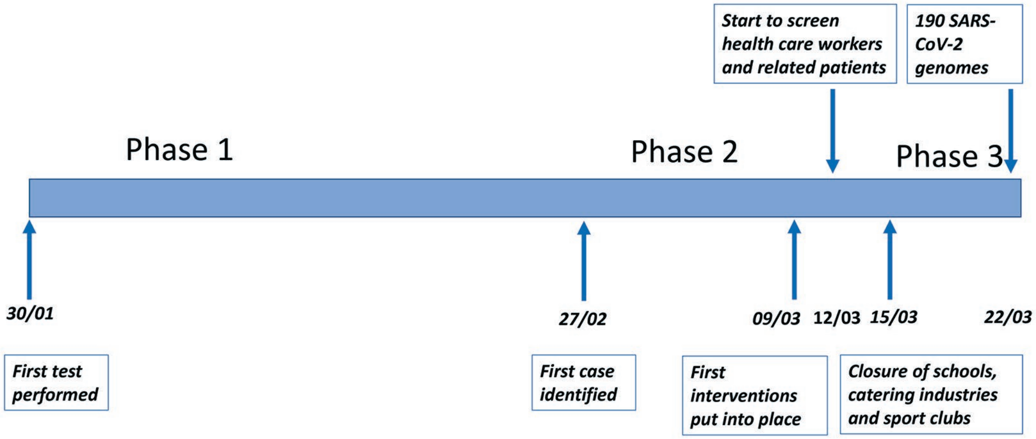
Competing interests

The authors declare no competing interests.

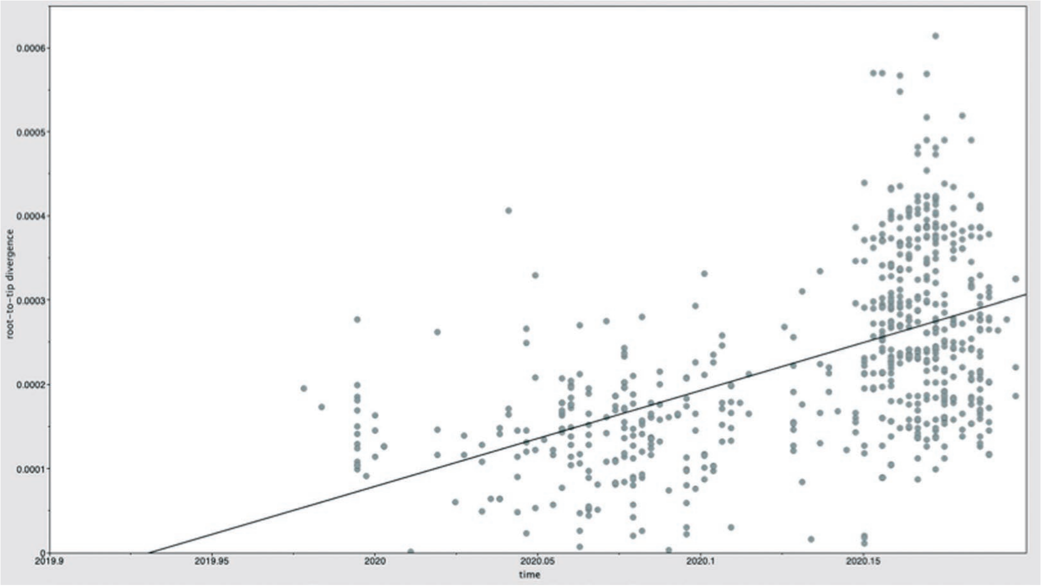
Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-0997-y.



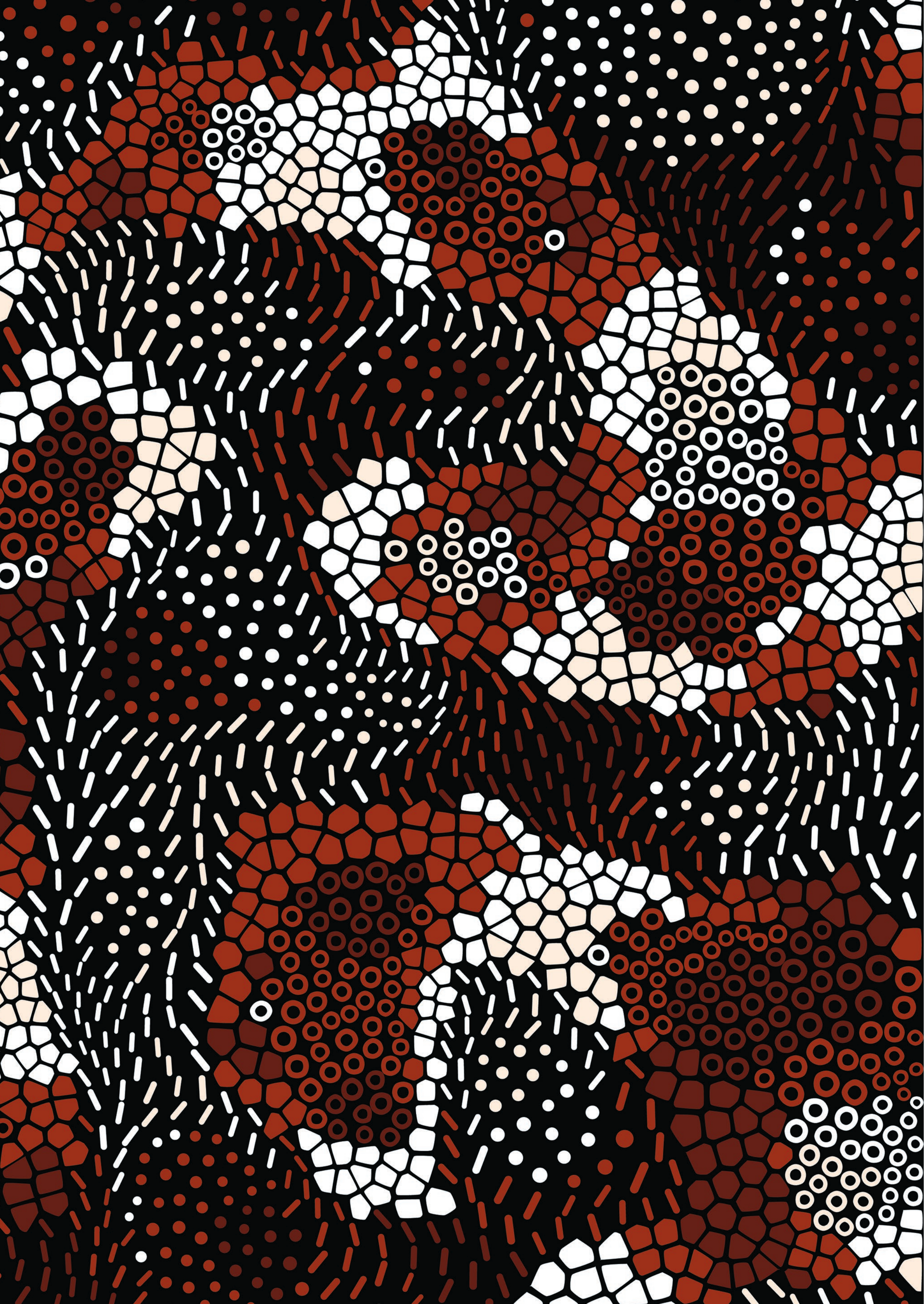
Extended Data Figure 1 Full maximum likelihood tree. Sequences are colored based on province of detection. Scale bar represent the number of substitutions per site.


Start to screen health care workers and related patients
190 SARS-CoV-2 genomes
Phase 1
Phase 2
Phase 3
30/01
27/02
09/03
12/03
15/03
22/03
First test performed
First case identified
First interventions put into place
Closure of schools, catering industries and sport clubs

Extended Data Figure 2 Timeline of the different phases. Graphical overview of the timeline of the of the different phases in the response to the SARS-CoV-2 outbreak in the Netherlands.



Extended Data Figure 3 Root-to-tip analysis. Report of the correlation coefficient for the root-to-tip divergence as a function of sample date.



COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study

Reina S. Sikkema[1]*, Suzan D. Pas[3,7]*, **David F. Nieuwenhuijse[1]**, Áine O'Toole[4], Jaco J. Verweij[5], Anne van der Linden[1], Irina Chestakova[1], Claudia Schapendonk[1], Mark Pronk[1], Pascal Lexmond[1], Theo Bestebroer[1], Ronald J. Overmars[1], Stefan van Nieuwkoop[1], Wouter van den Bijllaardt[7], Robbert G. Bentvelsen[7,9], Miranda M. L. van Rijen[7], Anton G. M. Buiting[5,6], Anne J. G. van Oudheusden[6], Bram M. Diederen[3], Anneke M. C. Bergmans[3], Annemiek van der Eijk[1], Richard Molenkamp[1], Andrew Rambaut[4], Aura Timen[10,11], Jan A. J. W. Kluytmans[2,7,8], Bas B. Oude Munnink[1], Marjolein F. Q. Kluytmans van den Bergh[2,7,8]*, Marion P. G. Koopmans[1]*

The Lancet Infectious Diseases 20 (11), 1273-1280, 234, 2020

8.1 Summary

10 days after the first reported case of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection in the Netherlands (on Feb 27, 2020), 55 (4%) of 1497 health-care workers in nine hospitals located in the south of the Netherlands had tested positive for SARS-CoV-2 RNA. We aimed to gain insight in possible sources of infection in health-care workers.

We did a cross-sectional study at three of the nine hospitals located in the south of the Netherlands. We screened health-care workers at the participating hospitals for SARS-CoV-2 infection, based on clinical symptoms (fever or mild respiratory symptoms) in the 10 days before screening. We obtained epidemiological data through structured interviews with health-care workers and combined this information with data from whole-genome sequencing of SARS-CoV-2 in clinical samples taken from health-care workers and patients. We did an in-depth analysis of sources and modes of transmission of SARS-CoV-2 in health-care workers and patients.

Between March 2 and March 12, 2020, 1796 (15%) of 12 022 health-care workers were screened, of whom 96 (5%) tested positive for SARS-CoV-2. We obtained complete and near-complete genome sequences from 50 health-care workers and ten patients. Most sequences were grouped in three clusters, with two clusters showing local circulation within the region. The noted patterns were consistent with multiple introductions into the hospitals through community-acquired infections and local amplification in the community.

Although direct transmission in the hospitals cannot be ruled out, our data do not support widespread nosocomial transmission as the source of infection in patients or health-care workers.

8.2 Introduction

In January, 2020, a cluster of patients with pneumonia of unknown cause was reported in Wuhan, China;[245] the disease was subsequently named COVID-19, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The clinical spectrum of COVID-19 varies from asymptomatic or mild symptomatic infections to severe respiratory symptoms and death, with older age groups generally presenting with more severe disease and higher death rates[266,267]. Since its identification, SARS-CoV-2 has rapidly spread across the globe. On June 22, 2020, 177 countries had reported cases of COVID-19, adding up to more than 8.9 million reported cases and 468 000 deaths worldwide[247].

Health-care workers are at increased risk of being exposed to SARS-CoV-2 and could potentially have a role in hospital transmission. Nosocomial outbreaks of severe acute respiratory syndrome coronavirus and Middle East respiratory syndrome coronavirus (MERS-CoV) are thought to have played a crucial part in the amplification and spread of these viruses. For MERS-CoV, hospital outbreaks caused approximately 50% of confirmed cases, of which around 40% were in health-care workers[268]. Currently, the extent of SARS-CoV-2 transmission and risk factors associated with infection in health-care settings are unclear. During the WHO–China Joint Mission on COVID-19[267], 2055 laboratory confirmed cases were reported in health-care workers from 476 hospitals in China, mostly (88%) from Hubei province. Most health-care workers were thought to have been infected within household settings rather than in a health-care setting, although conclusive evidence was scant[267].

On Feb 27, 2020, the first patient in the Netherlands tested positive for SARS-CoV-2 RNA after returning from a holiday to Lombardy, Italy[269]. In the following week, the number of infections in the country grew to 128, with an increasing proportion of cases without a known source of infection. These cases included nine health-care workers from two hospitals in the province of North Brabant, in the south of the Netherlands[270,271]. The Dutch national outbreak management team advised to extend screening of health-care workers to other hospitals in North Brabant, to assess possible community transmission. From March 6 to March 8, 2020, 1097 employees of nine hospitals were tested, of whom 45 (4%) were positive for SARS-CoV-2[271]. A follow-up study was done at three hospitals to assess the clinical presentations of COVID-19 of these health-care workers[270]. The impending shortage of personal protective equipment (PPE) and the proposed changes in its use in later phases of the outbreak response also triggered a debate on possible risks to health-care workers[272].

To understand sources and modes of transmission of SARS-CoV-2 in health-care workers and patients in the same hospitals, we did an in-depth analysis combining epidemiological data with whole-genome sequencing (WGS) of SARS-CoV-2 from clinical samples obtained from health-care workers and patients in three different hospitals

8.2.1 Evidence before this study

We searched Google Scholar on April 27, 2020, for articles published since 2020, with the keywords "SARS-CoV-2" AND "healthcare workers" AND "whole genome sequencing". We did not restrict our search to a publication language. Our search retrieved 13 results. Two reports presented original research of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); no reports were of the role of health-care workers in SARS-CoV-2 transmission or used whole-genome sequencing (WGS). Hospital transmission had an important role in previous outbreaks of Middle East respiratory syndrome and severe acute respiratory syndrome. The scarcity of personal protective equipment led to changes in policy during the initial phases of the SARS-CoV-2 outbreak response, also triggering a debate on possible risks to health-care workers. Up to now, possible SARS-CoV-2 outbreaks in health-care facilities have only been described using traditional molecular diagnostic tools combined with epidemiological data. However, previous studies implementing WGS have shown that hypotheses on virus transmission routes can be incorrect based solely on these data. Moreover, screening of health-care workers can be used to assess the level of local community transmission, but this can only be done if patient-to-health-care worker transmission can be reliably excluded.

8.2.2 Added value of this study

Our study aimed to gain insight in possible sources of infection of health-care workers at three hospitals in the Netherlands. All health-care workers with respiratory symptoms or fever in the previous 10 days were screened for SARS-CoV-2 infection. WGS was done of samples obtained from health-care workers and patients at these hospitals and this information was combined with epidemiological data.

8.2.3 Implications of all the available evidence

At the beginning of the SARS-CoV-2 outbreak in the Netherlands, health-care workers were probably infected in the community rather than at the hospitals. Possible nosocomial outbreaks should be carefully investigated using both epidemiological data and WGS to exclude or confirm transmission in health-care facilities.

8.3 Methods

8.3.1 Study design and participants

We did a cross-sectional study at two teaching hospitals (Amphia Hospital, Breda, Netherlands [700 beds], and Elisabeth-TweeSteden Hospital, Tilburg, Netherlands [800 beds]) and one regional hospital (Bravis Hospital, Roosendaal and Bergen op Zoom, Netherlands [600 beds]), at which 12 022 health-care workers in total were employed. PPE was used according to national guidelines that applied during this period of the outbreak[273,274]. Patients with suspected COVID-19 were nursed under strict isolation precautions and health-care workers applied additional PPE (gowns, gloves, goggles, hair cover, and type IIR surgical masks) on entering the isolation room. When aerosol-generating procedures were done, an FFP2 mask was used.

All health-care workers at these three hospitals who had fever or mild respiratory symptoms in the 10 days before screening for SARS-CoV-2 infection were eligible for testing, which was voluntary. All patients testing positive for SARS-CoV-2 and who had been admitted 2 days or more before the last date of onset of symptoms of health-care workers per hospital were included. All health-care workers with confirmed SARS-CoV-2 infection underwent a structured interview to obtain epidemiological data and to record any history of foreign travel and attendance at public events with more than 50 people, such as the yearly carnival in February, 2020 (appendix 1 p 1).

Ethics approval was obtained from the Medical Ethics Committee Brabant, with a waiver of written informed consent (METC Brabant/20.134/NW2020-26). Verbal informed consent was obtained from all health-care workers for SARS-CoV-2 testing, sequencing, and data collection. Data were deidentified before analysis. For patients, location and sequence data were obtained as part of the routine infection control policy in outbreak situations. All patients are notified of this policy on hospital admission and can actively dissent (opt out).

8.3.2 Procedures

We tested for SARS-CoV-2 infection using oropharyngeal or nasopharyngeal swabs in universal transport medium (Copan, Brescia, Italy) or E-swab medium (Amies; Copan), following local infection control policy during outbreaks. At Amphia Hospital and Bravis Hospital, total nucleic acids were extracted for RT-PCR after an external lysis step (1:1 with lysis binding buffer; Roche Diagnostics, Almere, Netherlands), using MagnaPure96 (Roche) with an input volume of 500 µL and output volume of 100 µL. The extraction was internally controlled by addition of a known concentration of phocine distemper virus (PDV)[275]. Subsequently, 10 µL extracted nucleic acids was amplified in three singleplex reactions in 25 µL final volume, using TaqMan Fast Virus 1-Step Master Mix (Thermofisher, Nieuwerkerk aan den IJssel, Netherlands), and 1 µL of primers and probe mixture for envelope (*E*) gene, RNA- dependent RNA-polymerase gene, and PDV[258]. Amplifi- cation was done in a 7500SDS (Thermofisher) with a cycling profile of 5 min at 50°C, 20 s at 95°C, 45 cycles of 3 s at 95°C, and 30 s at 58°C. At Elisabeth-TweeSteden Hospital, total nucleic acids were extracted, with a known concentration of PDV as internal control, using the QIAsymphony DSP virus pathogen midi kit and pathogen complex 400 protocol of the QIAsymphony Sample Processing system (Qiagen, Hilden, Germany), with an input volume of 400 µL and output volume of 110 µL. The amplification reaction was done in a volume of 25 µL with TaqMan Fast Virus 1-Step Master Mix (Thermofisher) and 10 µL extracted nucleic acids. A duplex PCR for *E* gene and PDV[258,276] with optimised primer and probe concentrations were done. Amplification with Rotorgene (QIAgen) consisted of 5 min at 50°C and 15 min at 95°C followed by 45 cycles of 15 s at 95°C, 30 s at 60°C, and 15 s at 72°C. Validations of RT-PCR procedures were done according to International Standards Organization guidelines (15189)[277].

For WGS, samples were selected based on a cycle threshold value less than 32. A SARS-CoV-2-specific multiplex PCR for nanopore sequencing was done, as previously described[201]. The resulting raw sequence data were demultiplexed using qcat. Primers were trimmed using cutadapt,[241] after which a reference-based alignment to the GISAID (Global Initiative on Sharing All Influenza Data) sequence EPI_ISL_412973 was done using minimap2[221]. The consensus genome was extracted and positions with a coverage less than 30 reads were replaced with N using a custom script using biopython software (version 1.74) and the python module pysam (version 0.15.3), as previously described[210]. Mutations in the genome were confirmed by manually checking the alignment, and homopolymeric regions were manually checked and resolved, consulting the reference genome. Genomes were included when having greater than 90% genome coverage. All available full-length SARS-CoV-2 genomes were retrieved from GISAID[254] on March 20, 2020 (appendix 1 pp 8–65), and aligned with the newly obtained SARS-CoV-2 sequences in this study using the multiple sequence alignment software MUSCLE (version 3.8.1551)[239]. Sequences with more than 10% of N position replacements were excluded. The alignment was manually checked for discrepancies, after which the phylogenomic software IQ-TREE (version 1.6.8)[260] was used to do a maximum-likelihood phylogenetic analysis, with the generalised time reversible substitution model GTR+F+I+G4 as best predicted model. The ultrafast bootstrap option was used with 1000 replicates. Clusters were ascertained based on visual clustering and lineage designations[278].

The code to generate the minimum spanning phylo- genetic tree was written in the *R* programming language. Ape[279] and igraph software packages were used to write the code to generate the minimum spanning tree, and the visNetwork software package was used to generate the visualisation. Pairwise sequence distance (used to generate the network) was calculated by adding up the absolute nucleotide distance and indel-block distance. Unambiguous positions were dealt with in a pairwise manner. Sequences that were mistakenly identified as identical, because of transient connections with sequences containing missing data, were resolved.

The multiple sequence alignment was curated and any error-rich sequences or sequences without a date were removed. The alignment was manually inspected and trimmed of the 5’ and 3’ untranslated regions in the bioinformatics software Geneious (version 11.1.3) to include only coding regions. The final length of the alignment was 29 408 nucleotides. Bayesian phylogenetic trees were estimated using BEAST version 1.10.4[263], with a Hasegawa-Kishino-Yano nucleotide substitution model[264] and a strict molecular clock. Two independent chains were run for 100 million states, with a Skygrid coalescent prior (appendix 1 p 3)[280,281] and parameters were sampled every 10 000 states. The LogCombiner program was used to combine the independent chains and to remove the burn-in from the tree file, and Tracer[265] was used to assess convergence. The maximum clade credibility tree was inferred using the TreeAnnotator program and visualised using baltic code and custom python scripts.

8.3.3 Statistical analysis

Epidemiological data obtained at structured interviews were entered in Castor Electronic Data Capture, version 2019. Continuous variables were expressed as medians and ranges and categorical variables were summarised as numbers and percentages. All analyses were done with SPSS version 25.0 (IBM, Armonk, NY, USA). Because of the descriptive nature of our study, sample size calculations and analyses of significance were not done. Results were reported following STROBE guidelines for observational studies.

8.3.4 Role of the funding source

The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all data in the study and had final responsibility for the decision to submit for publication.

8.4 Results

Between March 2 and March 12, 2020, 1796 (15%) of 12 022 health-care workers were voluntarily screened at the three participating hospitals (appendix 1 p 5). At Amphia Hospital, 42 (5%) of 783 health-care workers tested positive for SARS-CoV-2 RNA; at Bravis Hospital, ten (2%) of 443 health-care workers tested positive; and at Elisabeth-TweeSteden Hospital, 44 (8%) of 570 health-care workers tested positive. Characteristics of these 96 health-care workers who tested positive for SARS-CoV-2 RNA are shown in the table. The health-care workers were employed in 58 different departments, including on 42 medical wards. The median age of affected health-care workers was 49 years (range 22–66), and 80 (83%) of 96 were female, reflecting the proportion of female health-care workers among the total population employed in the participating hospitals (ie, 9784 of 12 022 [81%]). 20 staff members who did not have direct contact with patients tested positive for SARS-CoV-2 RNA, of whom six (30%) reported contact with colleagues who had also tested positive. Ten health-care workers reported a history of foreign travel in the 14 days before onset of symptoms, three (30%) of whom had travelled to northern Italy. 60 (63%) health-care workers had celebrated carnival in the 14 days before onset of symptoms, mostly in Breda, Prinsenbeek, and Tilburg. One health-care worker (who reported first symptoms on Feb 21, 2020) attended several carnival events while symptomatic but unaware of having COVID-19. 31 (32%) health-care workers reported close contact with an individual with confirmed COVID-19 in the 14 days before onset of symptoms, either a patient (n=3), colleague (n=18), household member (n=1), or another person outside the hospital (n=9).

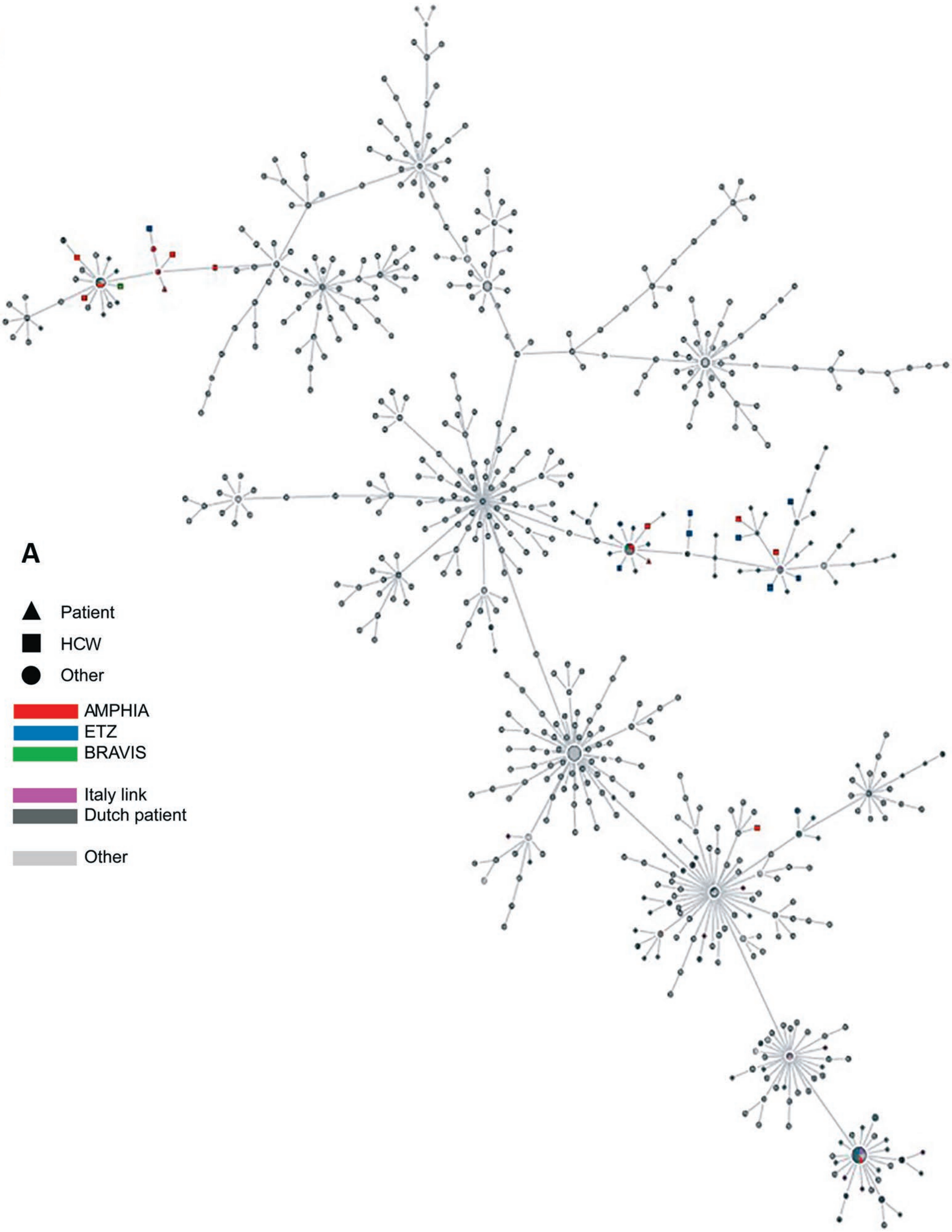
Table Descriptive characteristics of 96 health-care workers testing positive for severe acute respiratory syndrome coronavirus 2 RNA at three hospitals in the south of the Netherlands in March, 2020

	Health-care workers (n=96)
Sex	
Male	16 (17%)
Female	80 (83%)
Age, years	49 (22–66)
Residence	
Breda	11 (11%)
Prinsenbeek	11 (11%)
Tilburg	24 (25%)
Other city	50 (52%)
Department	
Medical	76 (79%)
Staff without direct patient contact	20 (21%)
Foreign travel, 14 days before onset of symptoms	10 (10%)
Northern Italy	3 (3%)
Austria	3 (3%)
UK	1 (1%)
Spain	1 (1%)
Portugal	1 (1%)
Switzerland	1 (1%)
Attendance at carnival with 50 people or more, 14 days before onset of symptoms	60 (63%)
Breda	7 (7%)
Prinsenbeek	11 (11%)
Tilburg	20 (21%)
Other city	22 (23%)
Attendance at other event with 50 people or more, 14 days before onset of symptoms	31 (32%)

	Health-care workers (n=96)
Close contact with individual with confirmed COVID-19, 14 days before onset of symptoms	31 (32%)
Patient	3 (3%)
Colleague	18 (19%)
Household member	1 (1%)
Other, outside hospital	9 (9%)

Data are n (%) or median (range).

Between March 2 and March 7, 2020 (Amphia Hospital), March 2 and March 10, 2020 (Bravis Hospital), and Feb 29 and March 9, 2020 (Elisabeth-TweeSteden Hospital), 856 patients were tested for SARS-CoV-2 RNA, of whom 345 were at Amphia Hospital, 228 were at Bravis Hospital, and 283 were at Elisabeth-TweeSteden Hospital (appendix 1 p 5). 23 (3%) patients tested positive for SARS-CoV-2 RNA, nine at Amphia Hospital and 14 at Elisabeth-TweeSteden Hospital. We obtained complete and near-complete SARS-CoV-2 genomes from 50 of 96 health-care workers (appendix 1 pp 4–5). 30 health-care workers were from Amphia Hospital, six were from Bravis Hospital, and 14 were from Elisabeth-TweeSteden Hospital. We obtained near-complete SARS-CoV-2 sequences from seven patients at Amphia Hospital and three patients at Elisabeth-TweeSteden Hospital. 46 (92%) of 50 sequences from health-care workers in this study grouped in three clusters (**figure, A**; appendix 1 p 4; appendix 2). Ten (100%) of ten sequences from patients in the study grouped into the same three clusters: seven were in cluster 1, two were in cluster 2, and one was in cluster 3. Cluster 1 contained 29 sequences (of which 12 were identical) of SARS-CoV-2 in samples taken from health-care workers and patients at all three hospitals (appendix 1 p 5). 13 (45%) sequences were from Amphia Hospital, three (10%) were from Bravis Hospital, and 13 (45%) were from Elisabeth-TweeSteden Hospital (**figure, C**). 11 (79%) of 14 health-care workers and two (67%) of three patients at Elisabeth-TweeSteden Hospital were in cluster 1. Cluster 2 contained 20 sequences (of which ten were identical) of SARS-CoV-2 in samples taken from health-care workers and patients at all three hospitals (appendix 1 p 5). 17 (85%) sequences originated from Amphia Hospital, two (10%) were from Bravis Hospital, and one (5%) was from Elisabeth-TweeSteden Hospital (**figure, B**). Health-care workers in cluster 2 were associated with Prinsenbeek and Breda, either by attendance at the carnival or by residence, more frequently compared with the other clusters (appendix 1 p 4).


A
Patient
HCW
Other
AMPHIA
ETZ
BRAVIS
Italy link
Dutch patient
Other

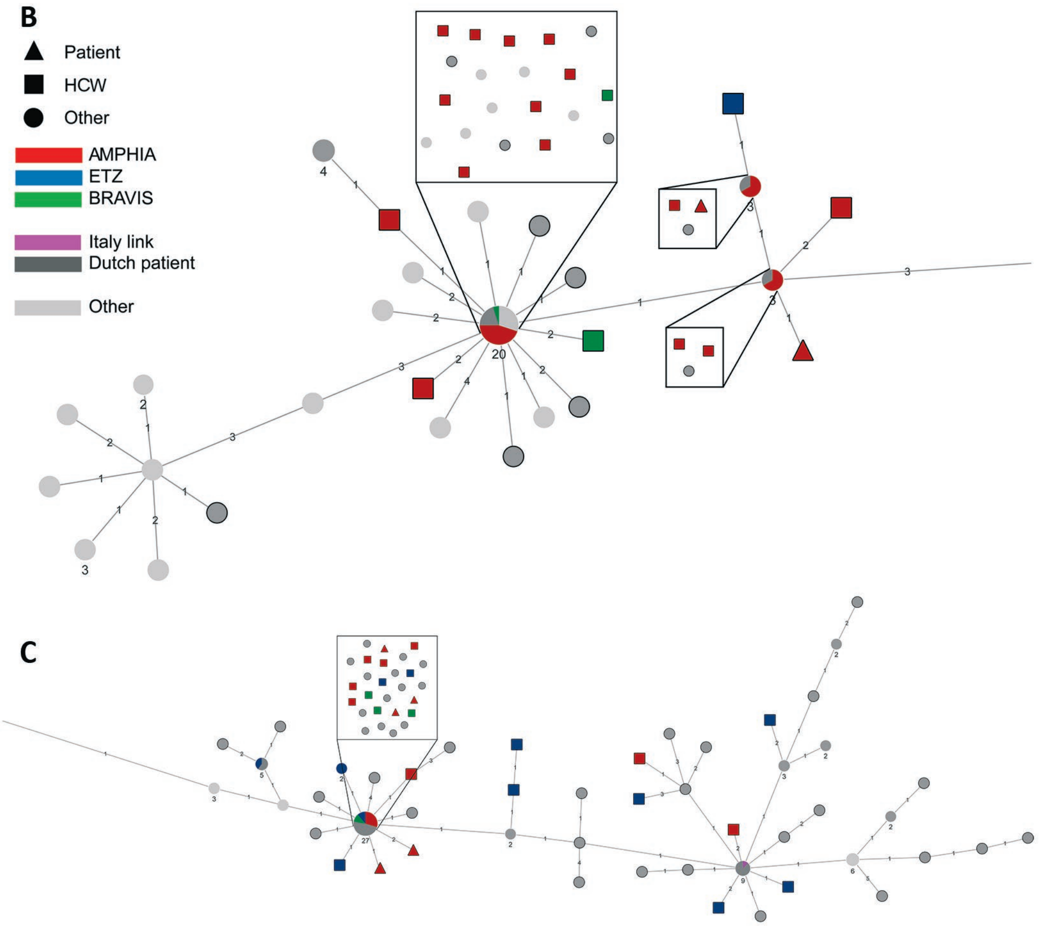


Figure Minimum spanning tree of available full-length SARS-CoV-2 genomes obtained from GISAID on March 20, 2020. The full tree (A) shows three clusters of SARS-CoV-2 genomes, obtained from sequencing samples from health-care workers and patients in the south of the Netherlands in March, 2020. An interactive version of the full tree can be found in appendix 2; it can be accessed by unzipping and opening the visNetwork.html file. Clusters 2 (B) and 1 (C) are shown in more detail. Numbers next to nodes indicate the number of sequences included. Numbers on branches indicate the difference in number of nucleotides between sequences. SARS-CoV-2=severe acute respiratory syndrome coronavirus 2. GISAID=Global Initiative on Sharing All Influenza Data.

Cluster 3 contained seven sequences (of which four were identical) of SARS-CoV-2 in samples taken from health-care workers and patients at all three hospitals. Four sequences were from health-care workers at Amphia hospital and one each was from Bravis Hospital and Elisabeth-TweeSteden Hospital. One sequence from a patient at Elisabeth-TweeSteden Hospital was also included in this cluster. A relatively large proportion of sequences in cluster 3 were from people with a travel history to northern Italy, as described elsewhere.[16] However, only two of six health-care workers in this cluster reported recent travel to either Italy or Austria (appendix 1 pp 4, 6–7).

Within each cluster, identical or near-identical sequences in health-care workers at the same hospital, and between patients and health-care workers at the same hospital, were found, but no consistent link was noted among health-care workers on the same ward or between health-care workers and patients on the same ward. Most (81–100%) health-care workers testing positive for SARS-CoV-2 at the three hospitals did not work on a ward with patients with confirmed COVID-19 (appendix 1 p 2). In wards with patients and health-care workers infected with SARS-CoV-2, direct transmission could be excluded in most cases, based on available WGS data (appendix 1 p 2). Notably, in Bravis Hospital, no patients with confirmed SARS-CoV-2 infection were hospitalised within 2 days before health-care workers at that hospital reported onset of symptoms. Additionally, no clusters were reported of more than three health-care workers on the same ward with identical or near-identical (two nucleotide difference or less) sequences. However, we cannot exclude health-care workers being infected in common hospital areas such as staff restaurants.

8.5 Discussion

In the present study, we combined epidemiological data with WGS to obtain a deeper understanding of the sources and modes of transmission of SARS-CoV-2 at three hospitals in the south of the Netherlands, which were the first hospitals to identify patients with COVID-19 in the Netherlands. Although possible hospital trans- mission of SARS-CoV-2 and health-care workers with COVID-19 have been reported,[267,282,283] to our knowledge, our study is the first to use WGS to analyse possible SARS-CoV-2 nosocomial transmission. Infection of health-care workers could have occurred through foreign travel, community contacts, or nosocomial transmission. The epidemiological data we obtained, combined with the presence of identical viruses in all three hospitals, and with non-hospitalised cases in other locations, indicates widespread community transmission in a very early phase of the outbreak. Mass gatherings, such as carnivals, in which just under two-thirds of health-care workers testing positive for SARS-CoV-2 participated, possibly acted as local super-spreading events.

Health-care workers are at increased risk of being exposed to viruses within hospitals but can also be a source of transmission by introducing a virus into their hospital. SARS-CoV-2 infections in health-care workers can have a substantial effect, because pathogens are introduced into settings with high numbers of individuals with comorbidities, potentially causing high morbidity and mortality among patients. The current study did not find evidence of large-scale nosocomial transmission in the early phase of the Dutch outbreak, and prevailing use of PPE and other infectious disease prevention measures were considered sufficient based on these early analyses and results[284].

Outbreaks in health-care settings are traditionally investigated by molecular diagnostic methods combined with epidemiological data. However, previous studies using WGS for hospital outbreak investigations have shown that hypotheses on virus transmission routes can be incorrect based solely on these data. By adding WGS data, particularly if results can be generated in a timely manner, and as long as sufficient reference sequences are available to allow a high resolution of the findings, the sequence analysis can provide essential information and inform subsequent infection control measures[285].

The mutation rate of SARS-CoV-2 is estimated to be around $1{\cdot}16 \times 10^{-3}$ substitutions per site per year, which corresponds to around one mutation every 2 weeks[286]. Therefore, finding identical or near-identical sequences in several locations and hospitals makes it difficult to draw definite conclusions on individual direct health-care worker-to-health-care worker or health-care worker-to-patient transmissions based on sequence data alone in this early stage of the SARS-CoV-2 outbreak, when genetic diversity of the circulating pathogen was negligible. Moreover, we did not obtain WGS of all health-care workers and patients testing positive for SARS-CoV-2 and, because of the small sample size, our analyses should be interpreted with caution. However, the finding of diverse clusters does exclude infection from one source. Moreover, the sequence-based analysis could be biased when sampling and sequencing is not done systematically and when sequence data in some areas are scarce, as is the case for COVID-19 internationally. For the Netherlands, we sequenced a substantial proportion of SARS-CoV-2 genomes as part of the national public health response[201], which was used as a reference set.

In conclusion, the genomic diversity recorded in our study is consistent with multiple introductions through community-acquired infections, and some local amplification related to specific social events in the community, rather than widespread within-hospital transmission. Although direct transmission in hospitals cannot be ruled out, our data do not support widespread nosocomial transmission as the source of infection in patients or health-care workers in our study. Because of the near-real-time sequence generation and analysis, our information was rapidly shared within the Dutch outbreak management team. Partly based on these data, SARS-CoV-2 was concluded to have already spread in the population in the province of North Brabant, which led to a change of policy, in which containment measures were complemented by targeted physical distance measures, starting in the south of the Netherlands initially and later comprising the whole country[201].

Author affiliations

**Contributed equally*

[1] Viroscience, Erasmus MC, Rotterdam, Netherlands

[2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands

[3] Microvida Laboratory for Microbiology, Bravis Hospital, Roosendaal, Netherlands

[4] University of Edinburgh, Edinburgh, UK

[5] Laboratory for Medical Microbiology and Immunology

[6] Laboratory for Medical Microbiology and Immunology and Department of Infection Control

[7] Elisabeth-TweeSteden Hospital, Tilburg, Netherlands, Microvida Laboratory for Microbiology

[8] Elisabeth-TweeSteden Hospital, Tilburg, Netherlands, Microvida Department of Infection Control

[9] Amphia Hospital, Breda, Netherlands; Department of Medical Microbiology, Leiden University Medical Center, Leiden, Netherlands

[10] Landelijke Coördinatie Infectieziektebestrijding, Rijksinstituut voor Volksgezondheid en Milieu, Bilthoven, Netherlands

[11] VU University Amsterdam, Amsterdam, Netherlands

Correspondence to

Dr Reina S Sikkema, Viroscience, Erasmus MC, 3015 CA Rotterdam, Netherlands
r.sikkema@erasmusmc.nl

Contributors

RSS, SDP, MFQKvdB, and MPGK wrote the report. SDP, MFQKvdB, and JAJWK set up sample and data collection. BBOM, AvdL, IC, MP, PL, SvN, TB, CS, and RJO generated sequence data. WvdB, RGB, MMLvR, AGMB, AJGvO, BMD, AMCB, AvdE, AT, JJV, and RM were involved in sample and data collection. RSS, DFN, AO'T, AR, BBOM, MPGK, and MFQKvdB were involved in data analysis and interpretation. RSS, SDP, JAJWK, MFQKvdB, and MPGK designed the study. All authors provided critical feedback.

Declaration of interests

We declare no competing interests.

Acknowledgments

We thank David van der Vijver and Miranda de Graaf (Erasmus MC, Rotterdam, Netherlands) for technical support. This study has been partly funded by EU Horizon 2020 projects RECoVer (no 101003589), VEO (no 874735), and the European Joint Programme One Health METASTAVA (no 773830), and by the National Institute of Allergy and Infectious Diseases, National Institutes of Health (contract HHSN272201400008C).

Funding EU Horizon 2020 (RECoVer, VEO, and the European Joint Programme One Health METASTAVA), and the National Institute of Allergy and Infectious Diseases, National Institutes of Health.

Rapidly analyzing and visualizing mismatches of specific RT-PCR primers/probes applied during the ongoing SARS-CoV-2 outbreak

David. F. Nieuwenhuijse1, Judit Szarvas2, Zsofia Igloi[1], Reina S. Sikkema[1], Richard Molenkamp[1], Ole Lund[2], Bas B. Oude Munnink[1], Marion P. G. Koopmans[1]

Manuscript in preparation

9

9.1 Abstract

Several RT-PCR-based SARS-CoV-2 diagnostic assays have been published based on the first available whole-genome sequences. With ongoing global spread of the virus, it is important to monitor mutations and the potential effect on the performance of these RT-PCR assays. We have developed a user-friendly web tool to *in silico* analyze and visualize primer mismatches that can potentially influence the performance of RT-PCR assays and demonstrate the use with SARS-CoV-2 as an example.

The aim of this study is to create a tool that quickly visualizes mismatches between primer sets and viral genome sequence alignments. Using the tool we show the mismatches in currently circulating strains of publicly shared SARS-CoV-2 genomes.

A tool was developed for the visualization of the primer and probe mismatches. The sequences to be matched were downloaded from GISAID and aligned using AliView.

A user friendly tool was created that shows easy to interpret summaries of primer and probe mismatches. A dataset of 4,239 sequences was downloaded and used as an example alignment.

The tool can aid clinicians to quickly check their primer and probe sequences against an up-to-date viral genome alignment.

9.2 Background

SARS-CoV-2, the causative infectious agent of Corona Virus Disease 19 (COVID-19), was first reported on the 31st of December 2019[287]. The first human cases of this novel coronavirus infection[288] were detected in Wuhan, China, and subsequent large scale human-to-human transmission was responsible for the current national and international spread of the virus[287,289–291]. By August 25th the virus had spread over more than 220 countries and at least 71,919,725 cases of COVID-19 were confirmed, including 1,623,064 deaths[292]. An initiative was started on GISAID[253] to facilitate the rapid sharing of SARS-CoV-2 genomic sequence data. To date (December 16th 2020) 267,626 complete genome sequences have been deposited in this database. The first RT-PCR assay, to detect SARS-CoV-2 RNA, was shared online on the 13th of January and validated and published ten days later[258].

Reliable and timely detection of infection is essential for proper clinical management and public health decision making. As well as for many other viral diseases, also for SARS-CoV-2 RT-PCR-based diagnostic assays are essential for diagnosis. Several institutes worldwide have developed RT-PCR-based assays to detect SARS-CoV-2 and have shared the details of their assays with the scientific and public health community. An up-to-date list of assays for which some level of validation has been provided on the WHO website[293] and is summarized in **Table 1**. Following these examples, also additional commercial assays have been developed. To monitor the sensitivity and specificity of RT-PCR-based assays an important first step is to find mismatches between primer and probes sequences and their target template, which is part of a routine procedure in clinical microbiology. Although difficult to predict the impact on performance, because this is based on the mismatch location and type, it is clear that more mismatches will generally have a bigger impact on the assay performance. It is evident that the effect of mismatches on assay performance should be evaluated by an expert user combined with experimental analysis after which primers or probes with an effect on assay performance should be optimized or revised. This is particularly important in an emerging infectious disease outbreak, where seeding of the strain and continued circulation in different regions may lead to evolution of lineages that differ genetically. Mutations in the RT-PCR target region have already been shown to reduce the sensitivity of the used test in clinical diagnostics[294,295]. Therefore, there may be a need for updating of primers and probes.

To rapidly match publicly available and laboratory-developed diagnostic primer and probe sets for any virus, with newly released genome sequences, we have developed a tool that aligns primers and probes to a multiple sequence alignment that can be obtained from public databases such as the GISAID database. Since the tool was designed for end-users without experience in computer programming, we have developed it as a web-based user-friendly interface which can be easily used.

9.3 Methods

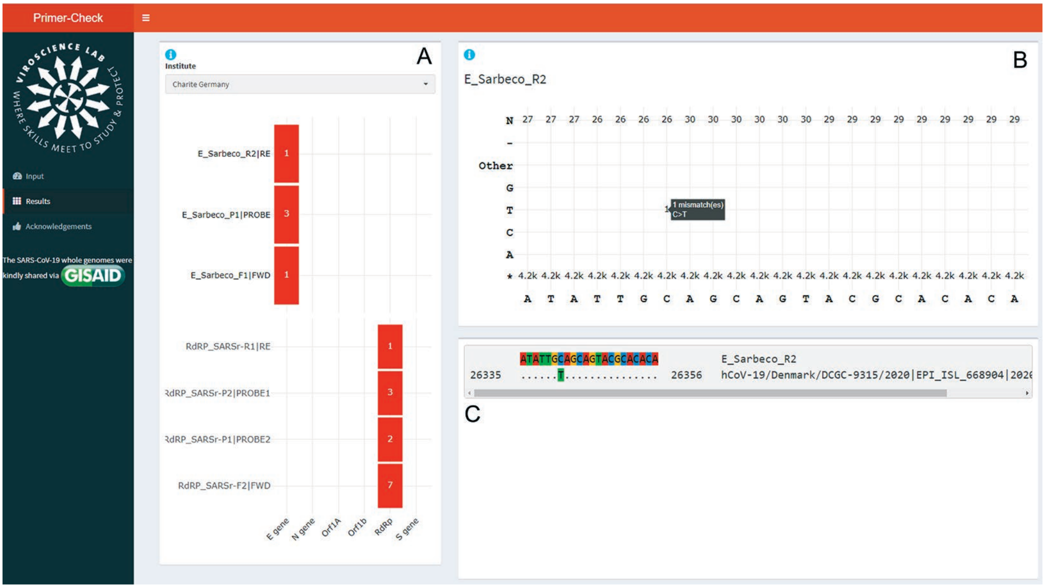
To visualize the mismatches with potential influence on the performance of primers and probes from RT-PCR based diagnostic assays, a data dashboard was developed using open-source R Shiny software[187]. Shiny is a framework to create interactive web-apps that can be shared online using the ShinyApps.io platform on a private server or on a local machine. The primer check dashboard is currently hosted online at https://viroscience-emc.shinyapps.io/primer-check/. The code repository is available at https://github.com/dnieuw/primer-check.

9.4 Results

The primer checking is performed by comparing a provided set of primers and probes with a multiple sequence alignment of genome sequences of interest in FASTA format. The first sequence in the multiple alignment is assumed to perfectly match all primers and probes and acts as a reference template for a single local alignment. In practice small mismatches and inserts are allowed, but large discrepancies will result in incorrectly aligned primers/probes and faulty results. The alignment position of the primers and probes to all other sequences in the multiple alignment file is inferred based on the alignment position of all primers and probes to the reference. The statistics of matches and mismatches are impromptu calculated in the results part of the application. As example dataset, the primers and probes of the 20 published diagnostic assays on the WHO website were checked against a selection of 4,239 whole genome SARS-CoV-2 sequences from the 16thth of October until the 16th of December (**Supplementary Table 1**, available in online version).

The selection was made by clustering the GISAID provided multiple alignment file by day, country and a minimal 5 nucleotide difference threshold and extracting the sequences from the last two months using a custom R script (**Supplementary File 1**). This resulted in 233,145 possible alignments that can be further investigated using the interactive panels of the result section of the dashboard (**Figure 1**):

- In the overview window, the primer and probe sets can be selected by institute/country in a dropdown menu (**Figure 1a**).
- Primer and probe alignments can be selected based on their name and target gene using a clickable tile diagram. In this diagram red tiles indicate one or more mismatches (number) with the currently included genome sequences. Gray tiles in the diagram represent primers and probes with no mismatches.
- In the box in **Figure 1b**, alignments are displayed in a matrix like diagram showing the primer or probe sequence on the x axis and the number and type (y axis) of matched (*) and mismatched nucleotides with the target genomes.
- By clicking on one of the numbers indicating a match or a mismatch all selected alignments are shown (**Figure 1c**), together with the start and stop position of the alignment, and the sequence names.


Primer-Check
≡
VIROSCIENCE LAB
WHERE SKILLS MEET TO STUDY & PROTECT
Input
Results
Acknowledgements
The SARS-CoV-19 whole genomes were kindly shared via GISAID
A
E_Sarbeco_R2|RE
1
E_Sarbeco_P1|PROBE
3
E_Sarbeco_F1|FWD
1
RdRP_SARSr-R1|RE
1
RdRP_SARSr-P2|PROBE1
3
RdRP_SARSr-P1|PROBE2
2
RdRP_SARSr-F2|FWD
7
E gene
N gene
Orf1A
Orf1b
RdRp
S gene
B
E_Sarbeco_R2
N 27 27 27 26 26 26 26 30 30 30 30 30 30 29 29 29 29 29 29 29 29 29
-
Other
G
T
C
A
* 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k 4.2k
A T A T T G C A G C A G T A C G C A C A C A
1
1 mismatch(es)
C>T
C

Institute: Charite Germany

```
         ATATTGCAGCAGTACGCACACA          E_Sarbeco_R2
26335    ......T...............   26356  hCoV-19/Denmark/DCGC-9315/2020|EPI_ISL_668904|202
```

Figure 1 Overview of the primer-check web visualization interface. (A) Separated by institute, diagnostic primer sets are displayed in the form of a clickable heatmap. Red blocks indicate one or more mismatches with that specific primer or probe. (B) Based on the selection in A the number of mismatches and type of mismatches with all available reference sequences are shown. (C) Based on the selected mismatch in B an alignment is shown of a primer or probe with the reference in which matches are replaced with a "." symbol. The numbers on either side of the alignment indicate the alignment position in the reference genome.

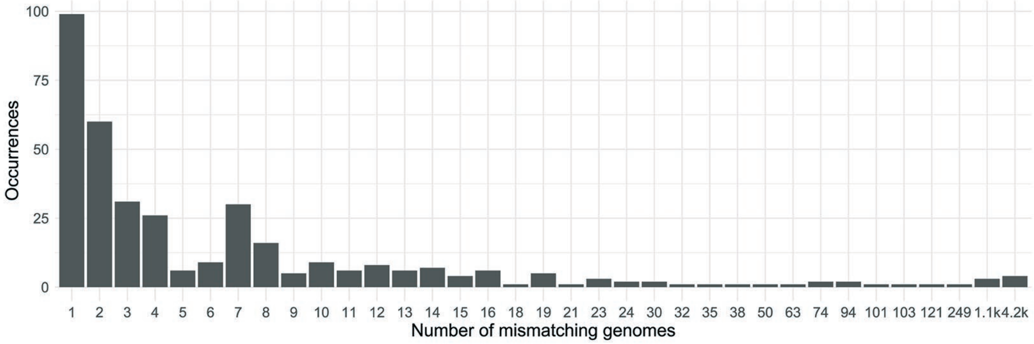


Figure 2 Summary of all mismatches and their occurrences in the example dataset. The x-axis displays the number of genomes with the same mismatch and the y-axis displays how often such a number of mismatches occurs. The total number of genomes in the example dataset is 4,239.

Following these four simple steps, users can browse through many alignments and quickly find mismatch profiles of their primers and probes of interest. A summary of all mismatches and their occurrences in the example dataset is displayed in **Figure 2**. Most of the mismatches occur in only 1 or 2 genomes, which could indicate a rare mutation or a sequencing error. There are, however, also mismatches that occur in more than 10 genomes or even in all the genomes in the example dataset.

9.5 Discussion

After the initial identification and release of the novel SARS-CoV-2 sequence the outbreak has continued and nucleotide mutations in the genome have been observed. With the release of every novel sequence the question arises whether the current RT-PCR-based assays can detect that specific variant. We show that these mutations can occur in target sites of diagnostic primers and probes and provide a tool for end-users without programming experience to analyze their own primer and probe sets against a set of genome sequences of their interest.

To analyze other PCR assays, users can upload and analyze their own primer sets and multiple genome alignments to the application hosted online or download and install the tool locally and perform the assessment locally. For SARS-CoV-2, the online application will be periodically updated with novel primer and probe sets if they are made available on the WHO website and with new sequences that are deposited in the GISAID database. This will allow continuous assessment of the performance of the different SARS-CoV-2 RT-PCR-based assays. Note that it is not allowed to share the GISAID alignment as an example dataset and therefore the option to use example data will not work locally.

9.6 Conclusions

We created a user-friendly primer-check application to aid users in choosing the right primer and probe sets for their RT-PCR-based assays. The application is hosted online and can analyze any user specified primer and probe set and sequence alignment. It also provides a periodically updated example dataset with any newly released SARS-CoV-2 genomes, allowing for continuous prediction of the *in silico* performance of the published primer/probe sets.

The data used as an example in this study reveal that not all SARS-CoV-2 primer and probe sets listed on the WHO website perfectly match with all SARS-CoV-2 genomes, ranging from mismatches with single individual genomes to mismatches with all published genomes. Although these mismatches do not necessarily indicate decreased performance it is a warning that the assay should be validated *in-vitro*. The primer-check visualization dashboard gives users the opportunity to quickly check whether current diagnostic primers still align to the novel genome variants that are uploaded to the GISAID platform.

Supplementary data

Supplementary file 1: Available online, at https://github.com/dnieuw/primer-check

Author affiliations

[1] Erasmus Medical Centre, Rotterdam, the Netherlands

[2] National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark

Acknowledgments

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 874735 (VEO) and from ZonMw (grant agreement No. 10150062010005).

We gratefully acknowledge the originating laboratories, where specimens were first obtained, and the submitting laboratories, where sequence data were generated and submitted to the EpiFlu Database of the Global Initiative on Sharing All Influenza Data (GISAID),

on which this research is based. All contributors of data may be contacted directly via the GISAID website (http://platform.gisaid.org). The accession numbers of all genetic sequences used in this study are provided in supplementary table 1 and are accessible from the website of GISAID (http://platform.gisaid.org).

Conflict of interest

The authors declare no conflict of interest.

Address for correspondence: prof. M.P.G. Koopmans, Viroscience department, Erasmus MC, Dr. Molewaterplein 40, Rotterdam, the Netherlands; email: m.koopmans@erasmusmc.nl

Table 1 Overview of available SARS-CoV-2 diagnostic assays as listed by the WHO[293]. (FWD) Forward primer. (RE) Reverse primer. All sequences are in 5' - 3' orientation.

Name	Sequence	Type	Origin	Target
TARGET 1 (ORF1AB) F	CCCTGTGGGTTTTACACTTAA	FWD	CDC China	Orf1ab
TARGET 1 (ORF1AB) R	ACGATTGTGCATCAGCTGA	RE	CDC China	Orf1ab
TARGET 1 (ORF1AB) P	FAM -CCGTCTGCGGTATGTGGAAAGGT-TATGG-BHQ1	PROBE	CDC China	Orf1ab
Target 2 (N) F	GGGGAACTTCTCCTGCTAGAAT	FWD	CDC China	N
TARGET 2 (N) R	CAGACATTTTGCTCTCAAGCTG	RE	CDC China	N
TARGET 2 (N) P	FAM-TTGCTGCTGCTTGACAGATT-TAMRA	PROBE	CDC China	N
RDRP_SARSR-F2	GTGARATGGTCATGTGTGGCGG	FWD	Charite Germany	RdRP
RdRP_SARSr-R1	CARATGTTAAASACACTATTAGCATA	RE	Charite Germany	RdRP
RDRP_SARSR-P2	FAM-CAGGTGGAACCTCATCAGGAGAT-GC-BBQ	PROBE1	Charite Germany	RdRP
RDRP_SARSR-P1	FAM-CCAGGTGGWACRTCATCMGGT-GATGC-BBQ	PROBE2	Charite Germany	RdRP
E_SARBECO_F1	ACAGGTACGTTAATAGTTAATAGCGT	FWD	Charite Germany	E
E_Sarbeco_R2	ATATTGCAGCAGTACGCACACA	RE	Charite Germany	E
E_SARBECO_P1	FAM-ACACTAGCCATCCTTACT-GCGCTTCG-BBQ	PROBE	Charite Germany	E
HKU-ORF1B-NSP14F	TGGGGYTTTACRGGTAACCT	FWD	HKU HongKong	Orf1b
HKU- ORF1B-NSP14R	AACRCGCTTAACAAAGCACTC	RE	HKU HongKong	Orf1b
HKU-ORF1b-nsp141P	FAM-TAGTTGTGATGCWATCATGACTAG-TAMRA	PROBE	HKU HongKong	Orf1b
HKU-NF	TAATCAGACAAGGAACTGATTA	FWD	HKU HongKong	N
HKU-NR	CGAAGGTGTGACTTCCATG	RE	HKU HongKong	N
HKU-NP	FAM-CCGCAAATTGCACAATTTGC-TAMRA	PROBE	HKU HongKong	N
WH-NIC N-F	CGTTTGGTGGACCCTCAGAT	FWD	NIH Thailand	N
WH-NIC N-R	CCCCACTGCGTTCTCCATT	RE	NIH Thailand	N
WH-NIC N-P	FAM-CAACTGGCAGTAACCA-BQH1	PROBE	NIH Thailand	N
NIID_2019-NCOV_N_F2	AAATTTTGGGGACCAGGAAC	FWD	NIID Japan	N
NIID_2019-nCOV_N_R2	TGGCAGCTGTGTAGGTCAAC	RE	NIID Japan	N
NIID_2019-NCOV_N_P2 F	FAM-ATGTCGCGCATTGGCATGGA-BHQ	PROBE	NIID Japan	N

Name	Sequence	Type	Origin	Target
NIID_WH-1_F501	TTCGGATGCTCGAACTGCACC	FWD	NIID Japan	Orf1a
NIID_WH-1_R913	CTTTACCAGCACGTGCTAGAAGG	RE	NIID Japan	Orf1a
NIID_WH-1_F509	CTCGAACTGCACCTCATGG	FWD	NIID Japan	Orf1a
NIID_WH-1_R854	CAGAAGTTGTTATCGACATAGC	RE	NIID Japan	Orf1a
NIID_WH-1_SEQ_F519	ACCTCATGGTCATGTTATGG	FWD	NIID Japan	Orf1a
NIID_WH-1_SEQ_R840	GACATAGCGAGTGTATGCC	RE	NIID Japan	Orf1a
WuhanCoV-spk1-f	TTGGCAAAATTCAAGACTCACTTT	FWD	NIID Japan	S
WUHANCOV-SPK2-R	TGTGGTTCATAAAAATTCCTTTGTG	RE	NIID Japan	S
NIID_WH-1_F24381	TCAAGACTCACTTTCTTCCAC	FWD	NIID Japan	S
NIID_WH-1_R24873	ATTTGAAACAAAGACACCTTCAC	RE	NIID Japan	S
NIID_WH-1_Seq_F24383	AAGACTCACTTTCTTCCACAG	FWD	NIID Japan	S
NIID_WH-1_SEQ_R24865	CAAAGACACCTTCACGAGG	RE	NIID Japan	S
2019-NCOV_N1-F	GACCCCAAAATCAGCGAAAT	FWD	CDC US	N
2019-NCOV_N1-R	TCTGGTTACTGCCAGTTGAATCTG	RE	CDC US	N
2019-nCoV_N1-P	FAM-ACCCCGCATTACGTTTGGTGGACC-BHQ1	PROBE	CDC US	N
2019-NCOV_N2-F	TTACAAACATTGGCCGCAAA	FWD	CDC US	N
2019-NCOV_N2-R	GCGCGACATTCCGAAGAA	RE	CDC US	N
2019-NCOV_N2-P	FAM-ACAATTTGCCCCCAGCGCTTCAG-BHQ1	PROBE	CDC US	N
2019-NCOV_N3-F	GGGAGCCTTGAATACACCAAAA	FWD	CDC US	N
2019-NCOV_N3-R	TGTAGCACGATTGCAGCATTG	RE	CDC US	N
2019-NCOV_N3-P	FAM-AYCACATTGGCACCCGCAATCCTG-BHQ1	PROBE	CDC US	N
NCOV_IP2-12669FW	ATGAGCTTAGTCCTGTTG	FWD	Institut Pasteur	RdRp
NCOV_IP2-12759RV	CTCCCTTTGTTGTGTTGT	RE	Institut Pasteur	RdRp
NCOV_IP2-12696BPROBE	AGATGTCTTGTGCTGCCGGTA	PROBE	Institut Pasteur	RdRp
NCOV_IP4-14059FW	GGTAACTGGTATGATTTCG	FWD	Institut Pasteur	RdRp
NCOV_IP4-14146RV	CTGGTCAAGGTTAATATAGG	RE	Institut Pasteur	RdRp
NCOV_IP4-14084PROBE	TCATACAAACCACGCCAGG	PROBE	Institut Pasteur	RdRp
E_SARBECO_F1-PASTEUR	ACAGGTACGTTAATAGTTAATAGCGT	FWD	Institut Pasteur	E gene
E_SARBECO_R2-PASTEUR	ATATTGCAGCAGTACGCACACA	RE	Institut Pasteur	E gene
E_SARBECO_P1-PASTEUR	ACACTAGCCATCCTTACTGCGCTTCG	PROBE	Institut Pasteur	E gene

Summarizing discussion

10

Viral metagenomic sequencing is an exciting new approach that has been put forward as a solution for a variety of applications. This has been supported by the development of better and new sequencing technologies. With the rapid development of second and third generation sequencing platforms, it is getting cheaper to produce increasingly more data. Especially nanopore-based portable long-read sequencing enables affordable sequencing in the field as demonstrated in a campaign to track and trace the Ebola virus outbreak in 2014 in Guinea[98]. Sequencing has been developed as a tool for various diagnostic applications[296], for outbreak investigations[297–299] and for the discovery of novel viruses in specific hosts or environments[300,301]. The idea of metagenomic sequencing is to have a single protocol to capture all pathogens of interest, a "catch-all" approach, for all these different applications. The potential and challenges of this approach for food- and water borne viruses are reviewed in **chapter 2** of this thesis. At that time, we concluded that several challenges still had to be overcome to make viral metagenomic sequencing an effective tool. Here, the challenges of applying viral metagenomics as tool for a variety of applications, the complexity of metagenomic sequencing data analysis, and the challenges of applying virus sequencing in a real-world scenario are discussed.

10.1 The application of viral metagenomic sequencing in diagnostics, public health virology and virus discovery

The term “metagenomic sequencing” covers several different approaches which are mainly distinguished by the goal for which they are used. On the one hand there is sequencing without a predefined target and on the other hand there is sequencing to determine the diversity of viruses, the so-called “virome”, in a sample. To avoid confusion, we will distinguish between “agnostic” single pathogen sequencing and “virome” sequencing to determine the entire diversity of viruses in a complex sample. Yet, there are applications where the difference between agnostic and virome sequencing is less clear, but in general this subdivision of metagenomic sequencing holds. Depending on the application there are several challenges that have to be addressed.

10.2 Virome sequencing for public health surveillance

Virome sequencing can play an important role in detecting and tracing of “Disease X”. Since it is not known what Disease X will be, it is important for virome surveillance to know what viruses can be detected in normal situations and how stable the virome is to know when something extraordinary is being found. The challenge is that, without a link to a patient, it is complex to predict if the detected virus is pathogenic[302]. It may therefore seem futile to catalogue environmental or animal viruses that are not (yet) associated with human illness. Still, the majority of currently known human viral pathogens have originated from a non-human host and therefore keeping track of which viruses reside in which animal or what environment could give important clues as to how the virus ended up in the human population and could guide efforts to prevent these spillover events.

To show the potential of global virome surveillance in sewage samples in **chapter 3** we sequenced a cross-sectional “baseline” of sewage samples from major cities around the world to get a better insight in the sewage virome. We found interesting differences in patterns of virus diversity around the globe that seem to match known virus seasonality. Even though we did not find strong indicators of viral abundance that could signify an ongoing outbreak, it is not the case that such signals cannot be found, as has been shown recently during the COVID19 pandemic[303]. In this study, although using targeted amplicon sequencing, SARS-CoV-2 sequences could be recovered from sewage and used to track the epidemic “waves” of the outbreak by standardized read counts. The next step would be to use virome sequencing to keep track of a broad set of known pathogens and to be vigilant in detecting novel ones.

10.2.1 Sensitivity

One of the main challenges of virome sequencing is the sensitivity and interpretability of sequencing of complex samples. A promising solution to increase the sensitivity and interpretability of virome sequencing is the use of capture probe sets which enrich a specific set of viruses of interest, as first showcased by Biese et al.[27] and Wylie et al.[304] with a probe set based on sequences of all vertebrate viruses available in 2014 and recently shown with a respiratory virus enrichment protocol for SARS-CoV-2 surveillance in sewage[305,306]. A virus capture protocol can help to enrich the sequence output for viruses of interest and thus increase the sensitivity. In our study in **chapter 5** the capture approach did not provide a large benefit over agnostic sequencing, but that may be explained by the relatively "clean" samples used in the study. The capture approach may work best in "dirty" samples such as sewage, where the issue is not the abundance of the virus genetic material, but the high abundance of other genetic material. Also, there is room to improve the targets of the capture enrichment approach since the capture probe set can be tailor made to enrich for specific targets of interest.

10.2.2 Cost

The cost of virome sequencing is another major obstacle to its use in routine pathogen surveillance. In a research setting, virome sequencing is a great tool for virus discovery, virus diversity characterization and surveillance and the cost of sequencing determines the scale of the experiment, but not whether the technology is cost-effective and whether it can be used routinely. For implementation, however, it is important to include the costs of the sequencing, as described in **chapter 5** when comparing different approaches especially if the aim is to eventually use it in a routine surveillance setting. It is difficult to estimate how much surveillance is sufficient and whether it warrants the costs to perform it routinely, but the chance to pick up an early signal of virus spillover to the human population with a possibility of mitigating an epidemic could be worth the investment. One example of a successful sewage surveillance guided approach was described for a poliovirus outbreak in Israel in 2013-2014[307], in that study qPCR based surveillance detected poliovirus circulation in sewage before it was detected in clinical samples which triggered a vaccination campaign. The next step would be to use virome sequencing for the same purpose. On a larger scale, the global SARS-CoV-2 pandemic has shown the financial damages a pandemic can cause and if there is any chance of mitigating such pandemic by routine virome sequencing surveillance it may be worth the necessary investments and costs.

10.2.3 Complexity

A last challenge is the complexity of virome sequencing data analysis. Because of its sensitivity and untargeted nature virome sequencing is very prone to contamination and erroneous results can be difficult to recognize. Contamination can, for instance, occur during library preparation[308] or come from lab reagents[309,310] making it difficult to determine the validity of the signal. This is mainly problematic in samples with low concentrations of viruses, and in cases where only few viral reads are detected. To address these issues, positive and negative controls can be included, such as samples for which the viral content is already known, or samples that are spiked with a known virus, although this can be problematic with low viral load samples in which the spiked virus can quickly become the sole sequence result.

Beside quality control in the experimental setup, it is of importance to include extensive quality assessment steps in the analysis process, especially when comparing samples or experiments. As shown in **chapter 3**, data quality is determined not only by the number of generated reads, but also by other factors such as the amount of background sequences, and the read replication rate. These measures can indicate how successful the virome was sequenced. Preprocessing of the data also strongly depends on the sequencing approach and if viral enrichment was used as shown in **chapter 5**. It is therefore important to report details on how the virome data was generated for it to be interpretable and reused in other studies.

To help interpreting the complex data resulting from virome sequencing, data interpretation tools can be made more user friendly and have better ways of visualizing the data as described in **chapter 4** of this thesis. Another way of improving the ease of use of bioinformatics software is by involving commercial parties to build user friendly interfaces to the software, and keep the software up-to-date. The difficulty with commercial bioinformatics software is that bioinformatics is a very dynamic field and before being developed into a fully mature software a better version has been developed in a research setting. Custom software-as-a-service (SAAS) providers could be a solution[192], but hiring such companies can be expensive. Moreover, professionalization of bioinformatics software will reduce its flexibility and openness, which is highly valued within the bioinformatics community. Quality of bioinformatics analysis software is mainly tested based on quality assessment ring trails and in silico methods comparisons. Closed source and pay-to-use software are therefore difficult to compare with newly developed methods, because they are essentially a “black box”, especially when performing an in-depth bioinformatic comparison where not only the input and output are considered, but also the reason behind the differences in results are investigated. The best solution would therefore be to encourage improvement of data visualization functionality and maintenance of bioinformatics tools from within the academic community. The difficulty, though, is that while they can be very time-consuming, these activities currently do not accumulate much academic credit, leading to a discrepancy between the number of tools developed and the number that actually is taken up in a sustained manner.

10.3 Agnostic sequencing for diagnostics, public health surveillance, and outbreak investigation

Beside the sequencing of a whole community of viruses to monitor and describe their diversity, metagenomic sequencing has potential as a pathogen detection tool as shown in several studies[311–313]. In these cases metagenomic sequencing is used to detect and characterize a single or several disease causing viruses without knowing what virus is present up front: an approach referred to as agnostic sequencing. The main advantage here is that agnostic sequencing can detect any virus, where other molecular diagnostic approaches are targeted to specific (panels of) viruses. The challenge here is that if agnostic sequencing is to replace these other approaches it has to be able to compete on sensitivity, cost and data interpretability. Here we discuss the challenges that still have to be addressed.

10.3.1 Sensitivity

As with virome sequencing the sensitivity of agnostic sequencing to detect a pathogen is still a issue, in **chapter 5** we compare the ability of several different platforms and sequencing approaches to perform whole genome sequencing of four different arboviruses with typical Ct values. We show that agnostic sequencing can detect the virus up to a Ct value of 33 using both Illumina sequencing and Nanopore sequencing showing the ability of agnostic sequencing to compete with other pathogen detection approaches. However, a higher concentration of the virus was necessary to generate a complete genome of the virus, which is necessary to fully characterize the virus and perform outbreak investigations. In those cases it is necessary to perform follow up experiments to complete the genome.

10.34.2 Cost

The benefit of agnostic sequencing as a clinical diagnostic tool is that it has the potential to detect all pathogens in a sample in a single test. However, it is more expensive than performing an RT-PCR test, which is why it is usually used as a last resort and not as a replacement. However, a recent analysis has highlighted the relatively little investment necessary compared to the large benefit gained from the additional information that genome sequencing provides if it can be used in an outbreak investigation[30]. Here, if put into monetary value, the losses caused by an outbreak quickly outweigh costs of genomic surveillance. However, time and cost effectiveness does depend on how the genomic surveillance is organized, how many samples are processed to benefit from the economies of scale, and the morbidity associated with the investigated pathogen[30]. Also, the study compared conventional diagnostics to whole genome sequencing using an amplicon based approach, therefore agnostic sequencing would be a next step.

10.3.3 Complexity

The lack of standardization is a main bottleneck for the implementation of agnostic sequencing in a routine setting, and several challenges have to be overcome to reduce the lab-to-lab variation of agnostic sequencing results[181,314,315]. The results are strongly impacted by the sample quality, library preparation methodology, sequence technology and data analysis, especially in complex sample matrixes. Besides, agnostic sequencing differs from other assays such as qPCR in that there is no single Ct value by which the similarity of the result can be measured, but there is the absence or presence of a virus, the number of reads, and the coverage of the genome which should be considered. It is therefore important to continue performing ring-trials to compare sequencing performance between laboratories, and to repeat these trials as sequencing technologies develop like we mention in **chapter 6**. Of importance, though, is that with these ring-trials, apart from virus detection, other parameters are also taken into account, such as cost effectivity, ability to standardize sequencing and hands-on time. To get a better understanding of how realistic the performance of a laboratory is it is as important to test if a laboratory can routinely and cost effectively perform the compared sequencing procedure as it is to determine if they can pick up the target pathogen.

To enable routine agnostic sequencing it may be beneficial to automate sample preparation, which is mainly available in the form of general-purpose pipetting workstations, which require large investments. A more affordable alternative would be small scale microfluidic devices for sample and library preparation, but these have not been successfully commercialized and are therefore only available on research level[316]. Nevertheless, research in on this topic is still ongoing, and new insights and developments in this field may result in library preparation microfluidics devices in the future[317].

10.4 The bioinformatics of viral metagenomic sequencing data

10.4.1 Bioinformatics workflow development

A paradigm shift has taken place when it comes to bioinformatics software development. In the last decade, driven by the FAIR principles[318] and facilitated by software distribution platforms such as Anaconda and CRAN, bioinformatics software has become increasingly easy to install. The infamous "dependency hell"[319] of installing bioinformatics software has been partially if not completely mitigated with the help of these platforms. In addition to the shift towards easily installable of bioinformatics software, there has also been a shift towards the use of bioinformatics workflow management tools such as Snakemake[320] and NextFlow[321]. These tools force a certain structure in the bioinformatics workflow that allows for easier interpretation and comparison of workflows, better scalability, and better reusability of parts of workflows.

A lack of structure and reusability has led to an explosion of workflows developed for virus metagenomics data analysis[31] as it is generally easier to write a new workflow than to continue working on, or adapting a preexisting workflow. This makes it difficult for a potential user to choose a workflow for their specific application or to compare different workflows. In practice this also means that a variety of workflows are being used for the same task, which has been shown lead to different analysis results[179,315], aggravating the problem of deciding which workflow to use. Therefore, the shift to using workflow management tools and software distribution platforms to structure and facilitate installation should make the comparison (in for instance EQAs) of different workflows easier going forwards.

The choice of workflow also depends on the purpose of the analysis, the type of data and the targeted viruses[31]. This has been exemplified in **chapter 5** where, depending on whether an amplicon approach, an agnostic approach or a capture-based approach was used, the bioinformatic data processing had to be adjusted to deal with the approach-specific data characteristics. Similar data processing challenges were encountered while analyzing the global sewage study in **chapter 3**, where the (necessary) high level of library amplification had to be adjusted for during data analysis to perform a proper quantitative comparison of the samples.

Therefore, for a fitting bioinformatic data analysis it is important to have a good understanding of the characteristics of the sequenced sample and the resulting data as they largely dictate the workflow to be used and choices to be made when analyzing the data. Fortunately, adjusting and comparison of different workflows is becoming easier as bioinformaticians start using software distribution platforms and workflow management tools to design them.

10.5 Increasing data volume

10.5.1 Increasing throughput of sequencing machines

Sequencing machines can nowadays easily generate millions of reads and thereby increase the potential of capturing partial and complete genomes of known or novel viruses. During outbreaks, as has been seen with SARS-CoV-2, the global community is now capable generating enormous amounts of viral genomes related to the outbreak[213]. On the other hand, more, and more in-depth virus diversity research allows for the redefinition of viral families, based on sequence data alone[158]. Increased throughput has also made it more cost effective to sequence more samples at the same sequencing depth as before. However, the high throughput of new sequencing machines does not always mean that more information is generated as we demonstrate in **chapter 3 and 5**, where more sequencing only results in more replicates in low-input samples.

10.5.2 Reference databases

One of the main challenges in viral bioinformatics is the choice of reference databases, where the choice is between inclusive databases such as Genbank[166] containing less-well curated sequences and exclusive databases such as RefSeq[322] containing well curated reference sequences. The latter may be more suitable to detect and annotate well-known viruses in a clinical diagnostic or public health surveillance setting, but for metagenomic virus discovery using an as inclusive as possible reference database may be preferred. However, the choice is difficult because of several issues with either inclusive or exclusive databases.

RefSeq contains only single reference genomes for most viral species, which is very limited given the viral diversity that may exist within a viral species. Therefore, while being a good reference database, it does not contain a good set of representatives for every viral species.

Also, especially for viruses, an up-to-date reference is important since viruses naturally evolve relatively rapidly over time and a reference genome of a few decades ago does not accurately represent the currently circulating strain of the same species.

On the other hand, Genbank contains all virus sequences submitted to the International Nucleotide Sequence Database Collaboration (INSDC). The database contains 8,099,669 sequences in the latest release (release 248), which is expected to increase exponentially following the massive attention to sequence based surveillance during the SARS-CoV-2 pandemic. While this is a much better representation of the true diversity of viruses the issue with inclusive databases is that they contain many redundant sequences, miss annotations, and have sequence artefacts, making it difficult to interpret the value of a “hit”.

Besides the officially recognized and characterized species, there are many non-classified viruses from metagenome datasets which expand the known virus diversity even more[323]. Also, the knowledge of bacteria tremendously increased as the amount of Genbank entries almost doubled from 1,235,345 to 2,286,858 between 2015 and 2022. For metagenomic sequence analysis this means that increasingly bigger databases have to be queried, which, with the current approaches, at some point will only be possible using extremely powerful computers.

There are several strategies to approach this database issue, such as restricting the database to a limited set of reference sequences, as has been done in the global sewage analysis in **chapter 3**. Other approaches make use of advanced search algorithms to enable querying of enormous databases[324]. There have also been attempts to “cleanup” and thereby reduce the size of genome databases for viruses[325] or create virus specific databases with better curated sequences[326]. There is however a tradeoff between sensitivity and specificity when choosing the appropriate reference databases and cleaned databases may give reliable results on the sequences they do contain but miss interesting viruses when performing virus discovery.

Thus, future efforts should be aimed towards the generation of curated, yet inclusive, reference databases and quality control should not only be applied to the sequence data, but also to the reference database used for annotation. Brute-force high-speed search algorithms will at some point still fail to cope with the exponential increase in reference sequences, therefore smarter solutions will have to be developed to handle the increase in reference sequences.

10.5.3 Large repositories of metagenomic read data

Beside viral reference databases, there is also a huge increase in unannotated and partially annotated metagenomic read data. Recent endeavors to map the viruses present in these data have shown an enormous wealth of virus diversity[327]. Adding this wealth of virus diversity to reference databases could help to put the findings of future metagenomic studies in perspective and reduce the fraction of so-called “dark matter” by linking these unknown sequences between metagenomic samples. However, current reference databases are already reaching their limits in terms of usability. Therefore, when these “dark matter” sequences are added to the mix, sequence annotation will be even more challenging and a smarter than brute-force solution must be found where database curation and sequence space reduction will play a large role.

10.6 Interpretation of complex metagenomic sample data

10.6.1 Virus taxonomy

The increase in virome sequencing of viruses strongly contributes to the expansion of knowledge about virus diversity. Therefore, it is important to catalogue the viruses in virome data well and thereby contribute to the general knowledge of viral taxonomy for future reference. However, viral taxonomy is very complex because of the lack of universal linker genes, complex evolutionary trajectories, and virus specific rules and thresholds. It is therefore difficult to automate taxonomic assignment. Recently several taxonomic annotation workflows have been developed for viruses that attempt to unify the expert knowledge derived and sequence data driven taxonomy which is a great step towards automated data-driven taxonomy[143,157]. However, these tools focus on annotation of complete genomes, which are often not available in virome sequencing datasets. Therefore, to allow taxonomic annotation of virome data, a solution is needed that enables the placement of partial genomes in this data-driven taxonomic framework as well. One approach would be by separating virus annotation into several quality categories[328,329]. The next step is to shift towards using this newly defined virus taxonomy in virus annotation workflows, which currently often rely on the NCBI determined taxonomic definitions. Transitioning to new sequence driven method of taxonomic annotation, will hopefully lead to more reliable and more informative taxonomic annotations of metagenomic data derived viruses.

10.6.2 Virus typing

Beside taxonomic annotation, for meaningful interpretation of the viral sequence the devil is often in the details. Not just the viral species, but determining the clade, type, subtype and/or lineage is usually crucial to a virologist to fully interpret the viral genome. These types of annotations are usually based on a shared phenotypic trait or an association with a specific phylogenetic or epidemiologic cluster and therefore detailed knowledge is needed on what genomic information is necessary to assign the annotation. Automation of these annotations is complex and will differ greatly from virus to virus and therefore specific viral typing tools are usually developed specifically tailored to the virus in question[28,330]. A specific complexity with metagenomic data is the fact that usually only part of the viral genome is recovered from the data, which means that if the typing region is not present in the data, the genome cannot be typed. A solution may be to look at approaches that use multiple, genome-covering, loci for typing analogous to multi-locus-sequence-typing as is used for typing of bacterial strains. However, if the recovered region is simply not informative for viral typing, or if the virus is known to regularly recombine parts of its genome, full typing is simply not possible based on a partial genome and follow-up experiments, for instance using a specific PCR, are necessary to type the virus. Given the complexity of virus typing it remains to be seen if a general automated typing tool for metagenomic sequencing data can be developed or if individual tools developed by individual expert groups will continue to be the standard. Nevertheless, it is important for viral metagenomic data analysis to consider that species determination if often not sufficient and detailed genome annotation is often needed to interpret the characteristics of the viruses that are present in the data.

10.6.3 Variant annotation

Another step in metagenomic sequence annotation is the reliable detection of variants and minor variants in virus genomes[331]. In **chapter 5** we show that variants can be reliably determined using a variety of approaches. However, we also show that, especially at lower virus concentrations, the reliability of variant determination diminishes. With the appropriate data curation methods, it is possible to increase the reliability and determine the complete genome up to Ct 33. For minor variant detection an amplicon approach, high number of genome copies per sample, replicate samples, deep sequencing, conservative SNP calling, and careful primer trimming is recommended for reliable intra-host diversity measurements[215]. Those criteria limit the analysis of this type of diversity to well-controlled experimental setups, although the described data quality criteria can be a useful guideline to reduce the risk of miss annotation.

10.6.4 Virus mixture separation

For environmental samples or other samples where a mix of similar viruses is expected to be present it is important to realize that assembly of these data can potentially generate mixture consensus viruses. One potential solution could be to use a haplotype recovery tool such as "Hansel and Gretel"[332] to attempt to recover "pure" viruses as has been done in a recent description of crAssphage diversity in metagenome samples[333]. Another promising approach, used to recover SARS-CoV-2 variants from sewage sample data, makes use of a tool that was originally used to determine RNA isoforms in Eukaryotes[334]. Especially in complex samples mixtures of similar viruses can be expected when using metagenomic sequencing and therefore recovering the genomes of individual viruses can be difficult and requires extra attention when analyzing the data in detail.

10.7 Real-world application of genomic surveillance during the SARS-CoV-2 pandemic

10.7.1 The added benefit of viral sequencing data for detection and following of an outbreak

By using agnostic sequencing and quick data sharing, the genome of the novel SARS-CoV-2 was available within weeks after the beginning of the pandemic[335], which greatly improved the pandemic preparedness of other countries. An incredible volume of sequencing data has been produced following the start of the SARS-CoV-2 pandemic which has shown the potential of large scale near real time sequence surveillance during an infectious disease outbreak to guide policy making (**chapter 7**) and to perform viral tracking and tracing in a hospital environment (**chapter 8**). For local pandemic response, contact tracing, and public health decision making, local dense sequencing information is very helpful. However, this pandemic in particular has shown how quickly a novel virus can spread across the globe and therefore global surveillance is necessary to predict and prevent the spread of novel viruses and continuous updates have to be executed to keep up with novel variants (**chapter 9**). It has become evident that there are major gaps in the global sequence surveillance coverage, especially in middle- and low-income countries[336]. Therefore, sequencing capacity building using low-cost equipment should be a main target for global infectious disease preparedness, as well as the establishment of a sustained cold-chain delivery scheme for reagents at compatible costs. A better coverage of sequence diversity research worldwide may also help to discover viral origins, which in case of SARS-CoV-2 is currently still under debate[337]. More prospective viral diversity research using metagenomic sequencing of animal, sewage and environmental samples may make it possible to find a better connection between outbreaks and these virus reservoirs in the future.

10.8.2 Massive scale genome sequencing during outbreaks

With the massive amounts of genome sequences new challenges have arisen to keep track of and extract information from this data. GISAID[253] has been very successful in collecting, curating, and distributing the generated genome sequences. From the research community various tools to interpret this data have been made such as Pangolin[330], to subdivide the sequences into lineages. There are still some technical challenges to be addressed, such as methods to distill good representative sequences from the bulk of (largely) similar sequences, and the annotation and interpretation of (minor) variants from read data. On the other hand, it is good to question if this massive amount of concentrated genomic surveillance is useful without a better picture of the rest of the world. There is a strong imbalance in sequencing capacity between high and middle- and low-income countries which gives an incomplete image of disease global disease spread during a pandemic[338]. The SARS-CoV-2 pandemic has shown how fast viral variants can spread across the globe, and therefore local genomic surveillance must be supported by better genomic surveillance capacity in Africa and other low- and middle income countries to get a better view of global infectious disease emergence and spread[339]. Nevertheless, this pandemic has turbocharged the worlds knowledge and capacity to perform genomic surveillance, which will undoubtedly influence genomic surveillance in the future.

10.8.3 Organization of sequence surveillance during a pandemic

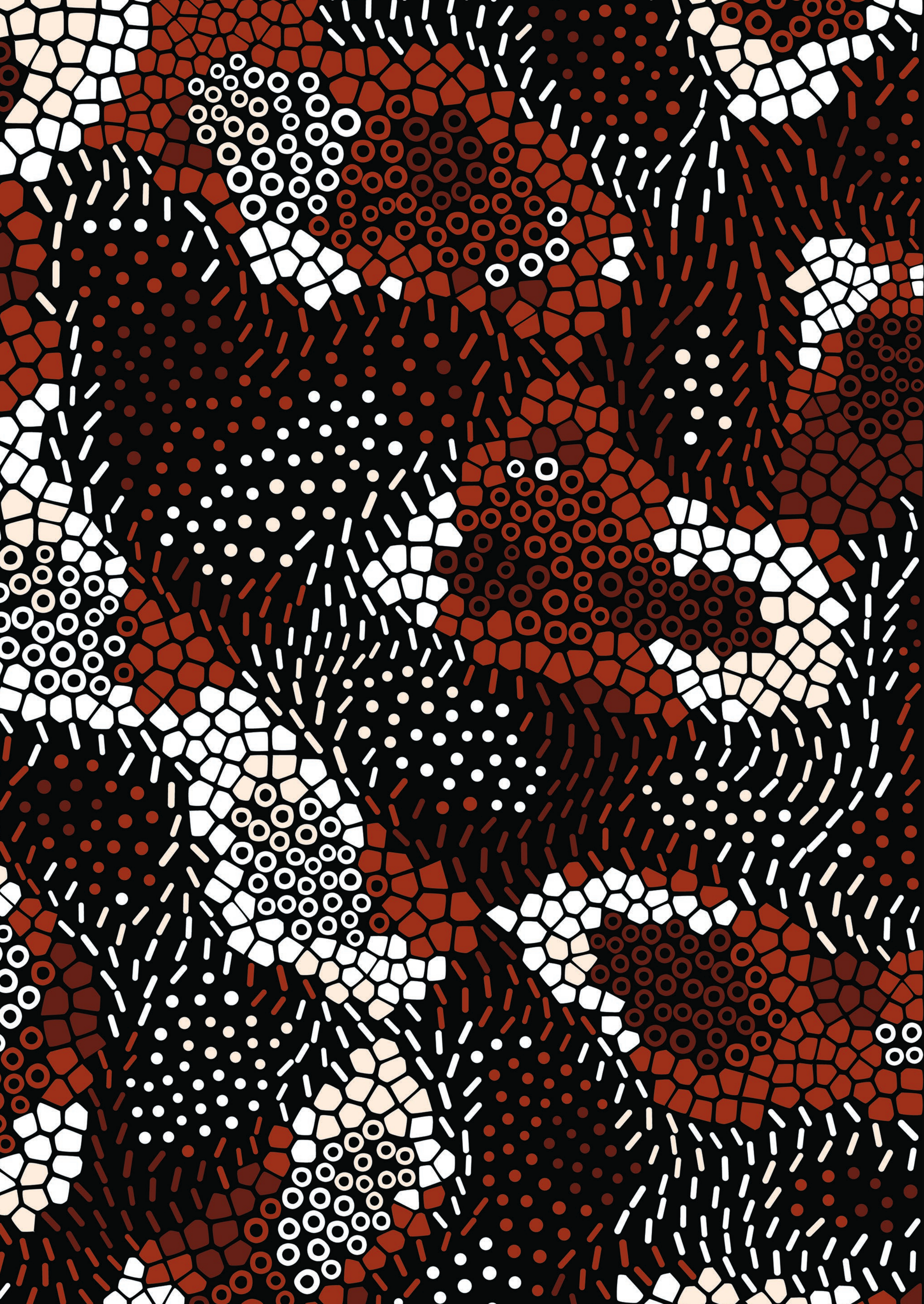
Following the discovery of a novel virus, the speed and number of complete genomes shared on the GISAID platform[253] showed the potential of genome sequence sharing for worldwide infectious disease outbreak surveillance. However, besides the aforementioned technical challenges of large-scale sequence-based surveillance, the COVID-19 pandemic has also revealed the challenges regarding data sharing and (inter)national collaboration, in particular when it comes to metadata. Genome sequences provide an epidemiological link between patients, but the metadata associated with the genomic data provide the extra information necessary to figure out why the sequences are linked. This information, however, often lacks necessary detail or is absent due to several metadata sharing challenges. The challenges of metadata sharing for proper outbreak response have already been studied in depth before the SARS-CoV-2 pandemic[340] and especially academic data sharing has improved over the years and is required by many academic journals before publication. Data sharing primarily becomes very complex when, apart from academic institutes, (inter)national public health organisations and commercial parties are involved, which have their own political, ethical, economic, administrative, regulatory and legal barriers[340]. The benefit of rapid sharing of sequence data during an outbreak was first shown during the 2009 H1N1 influenza A pandemic, facilitated by the development of the GISAD platform, and later during the Ebola outbreak in 2014 it (again) became apparent that data sharing was hampered by political and cultural barriers[341]. While the political and cultural concerns are understandable, the successful future of infectious disease surveillance and preparedness depends the data sharing solutions developed based on these “lessons learned”.

A first step to a solution would be a re-evaluation of what kind of data can be shared publicly and how to facilitate the sharing of confidential data for research purposes. In principle all metadata can be shared, but some data can only be shared using protected databases, while other data can be shared publicly. The issue here is that regulations surrounding data sharing are complex and it is often unclear what can and cannot be shared, which leads to a default towards data protection instead of attempting to share as much data as possible[340]. Adding to this is the large attention in society and the media towards data safety with an emphasis on the harm of “leaking” of data[342]. On the contrary there is little attention for the potential harm of excessive data privacy[343], which can hamper investigations due to the inability to share information. However, sharing of infectious disease outbreak information can also have severe negative (financial) consequences to the involved individual, company or country, making it difficult to balance between sharing and withholding data, especially when the benefit of data sharing is indirect and unclear, while the benefit of data privacy is direct and clear.

A second step would be to investigate what data is necessary for what investigations and what has to be done to facilitate the sharing of this information. The challenge here is to determine what information is relevant to share because it depends on the type of analysis. The “easy” solution would be to share everything, but there is a balance between the effort of sharing more data and the usefulness of sharing data. It must not be underestimated how much effort goes into the collection of data before it is shared and prioritization will help the willingness to share[344].

The willingness to share is probably the most important and a third step would be to investigate ways to incentivize data sharing by ensuring mutual benefit. For instance, by facilitating the act of data sharing, through user friendly interfaces in public and/or private databases, but more importantly providing services that make sharing data worth the effort and/or risk. To incentivize data sharing it has recently been suggested to use data Digital Object Identifiers (DOIs), allowing scientists to acknowledge eachothers data sharing efforts in their manuscripts, although this idea is not new and has already been proposed before in 2014[345] and has apparently not caught on as of yet.

All together effective data sharing is one of the most complex and important challenges to overcome to build a successful global infectious disease surveillance network. Following the SARS-CoV-2 pandemic, the importance of overcoming these challenges has been emphasized and important steps have been made, but also complex challenges remain.



References

11

1. Coker, R. J., Hunter, B. M., Rudge, J. W., Liverani, M. & Hanvoravongchai, P. Emerging infectious diseases in southeast Asia: Regional challenges to control. *The Lancet* **377**, 599–609 (2011).
2. Filion, P. & Keil, R. Contested Infrastructures: Tension, Inequity and Innovation in the Global Suburb. *Urban Policy Res.* **35**, 7–19 (2017).
3. Wood, C. L., McInturff, A., Young, H. S., Kim, D. & Lafferty, K. D. Human infectious disease burdens decrease with urbanization but not with biodiversity. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, (2017).
4. Curtis, P. G., Slay, C. M., Harris, N. L., Tyukavina, A. & Hansen, M. C. Classifying drivers of global forest loss. *Science (80-.).* **361**, 1108–1111 (2018).
5. Gottdenker, N. L., Streicker, D. G., Faust, C. L. & Carroll, C. R. *Anthropogenic Land Use Change and Infectious Diseases: A Review of the Evidence*. *EcoHealth* **11**, 619–632 (Springer Science and Business Media, LLC, 2014).
6. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science (80-.).* **310**, 676–679 (2005).
7. CDC, SARS Basics Fact Sheet. Available at: https://www.cdc.gov/sars/about/fs-sars.html#outbreak. (Accessed 7th February 2023).
8. Dawood, F. S. *et al.* Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet Infectious Diseases* **12**, e687–e695 (2012).
9. Saéz, A. M. *et al.* Investigating the zoonotic origin of the West African Ebola epidemic. *EMBO Mol. Med.* **7**, 17–23 (2015).
10. Coltart, C. E. M., Lindsey, B., Ghinai, I., Johnson, A. M. & Heymann, D. L. The Ebola outbreak, 2013–2016: old lessons for new epidemics. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 2013–2016 (2017).
11. Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. *Nat. 2016 5387624* **538**, 193–200 (2016).
12. Moore, S. M. *et al.* Leveraging multiple data types to estimate the size of the Zika epidemic in the Americas. *PLoS Negl. Trop. Dis.* **14**, e0008640 (2020).
13. Pan American Health Organization (PAHO). Zika cases and congenital syndrome associated with Zika virus reported by countries and territories in the Americas, 2015-2018 cumulative cases. (2018).
14. Faria, N. R. *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. *Science (80-.).* **352**, 345–349 (2016).
15. Bagherian, H., Farahbakhsh, M., Rabiei, R., Moghaddasi, H. & Asadi, F. National communicable disease surveillance system: A review on information and organizational structures in developed countries. *Acta Informatica Medica* **25**, 271–276 (2017).
16. Wang, Y. *et al.* A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic Acids Res.* **32**, 1197–1207 (2004).

17. Caliendo, A. M. Multiplex PCR and emerging technologies for the detection of respiratory pathogens. *Clin. Infect. Dis.* **52**, 326–330 (2011).

18. Huang, H. S. *et al.* Multiplex PCR system for the rapid diagnosis of respiratory virus infection: systematic review and meta-analysis. *Clinical Microbiology and Infection* **24**, 1055–1063 (2018).

19. Gibbons, C. L. *et al.* Measuring underreporting and under-ascertainment in infectious disease datasets: A comparison of methods. *BMC Public Health* **14**, 1–17 (2014).

20. Tang, J. W. *et al.* Global epidemiology of non-influenza RNA respiratory viruses: data gaps and a growing need for surveillance. *The Lancet Infectious Diseases* **17**, e320–e326 (2017).

21. Smits, S. L. *et al.* New viruses in idiopathic human diarrhea cases, the Netherlands. *Emerg Infect Dis* **20**, 1218–1222 (2014).

22. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E. & Fouchier, R. A. M. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* **367**, 1814–1820 (2012).

23. Oude Munnink, B. B. B. *et al.* Towards high quality real-time whole genome sequencing during outbreaks using Usutu virus as example. *Infect. Genet. Evol.* **73**, 49–54 (2019).

24. Lewandowska, D. W. *et al.* Optimization and validation of sample preparation for metagenomic sequencing of viruses in clinical samples. *Microbiome 2017 51* **5**, 1–13 (2017).

25. Yang, J. *et al.* Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J. Clin. Microbiol.* **49**, 3463–3469 (2011).

26. Lipkin, W. I. & Anthony, S. J. Virus hunting. *Virology* **479**–**480**, 194–9 (2015).

27. Briese, T. *et al.* Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *MBio* **6**, e01491-15 (2015).

28. Kroneman, A. *et al.* An automated genotyping tool for enteroviruses and noroviruses. *J. Clin. Virol.* **51**, 121–125 (2011).

29. Kundu, S. *et al.* Next-Generation Whole Genome Sequencing Identifies the Direction of Norovirus Transmission in Linked Patients. *Clin. Infect. Dis.* **57**, 407–414 (2013).

30. Alleweldt, F. *et al.* Economic evaluation of whole genome sequencing for pathogen identification and surveillance – results of case studies in Europe and the Americas 2016 to 2019. *Euro Surveill.* **26**, 1900606 (2021).

31. Nooij, S., Schmitz, D., Vennema, H., Kroneman, A. & Koopmans, M. P. G. Overview of Virus Metagenomic Classification Methods and Their Biological Applications. *Front. Microbiol.* **9**, 749 (2018).

32. Leipzig, J. A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* **18**, bbw020 (2016).

33. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, 1000605 (2009).

34. Cotten, M. *et al.* Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. *PLoS One* **9**, e93269 (2014).

35. Smits, S. L. *et al.* Recovering full-length viral genomes from metagenomes. *Front. Microbiol.* **6**, 1069 (2015).

36. Tam, C. C. *et al.* Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut* **61**, 69–77 (2012).

37. Sethi, D., Wheeler, J., Rodrigues, L. C., Fox, S. & Roderick, P. Investigation of under-ascertainment in epidemiological studies based in general practice. *Int. J. Epidemiol.* **28**, 106–12 (1999).

38. de Wit, M. A. *et al.* A comparison of gastroenteritis in a general practice-based study and a community-based study. *Epidemiol. Infect.* **127**, 389–97 (2001).

39. Payne, D. C. *et al.* Epidemiologic Association Between FUT2 Secretor Status and Severe Rotavirus Gastroenteritis in Children in the United States. *JAMA Pediatr.* **169**, 1040–5 (2015).

40. de Graaf, M., van Beek, J. & Koopmans, M. P. G. Human norovirus transmission and evolution in a changing world. *Nat. Rev. Microbiol.* **14**, 421–433 (2016).

41. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222 (2012).

42. Cadwell, K. The virome in host health and disease. *Immunity* **42**, 805–13 (2015).

43. Verhoef, L. *et al.* An Integrated Approach to Identifying International Foodborne Norovirus Outbreaks. *Emerg. Infect. Dis.* **17**, 412–418 (2011).

44. Eurosurveillance editorial team. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2011 has been published. *Euro Surveill. Bull. Eur. sur les Mal. Transm. = Eur. Commun. Dis. Bull.* **18**, 20449 (2015).

45. Petrignani, M. *et al.* Underdiagnosis of Foodborne Hepatitis A, the Netherlands, 2008–20101. *Emerg. Infect. Dis.* **20**, 596–602 (2014).

46. Moon, S. *et al.* Emerging Pathogens and Vehicles of Food- and Water-borne Disease Outbreaks in Korea, 2007-2012. *Osong public Heal. Res. Perspect.* **5**, 34–9 (2014).

47. ESFA. Manual for reporting on food-borne outbreaks in accordance with Directive 2003/99/EC for information deriving from the year 2015. *EFSA Support. Publ.* **13**, (2016).

48. Deng, X., den Bakker, H. C. & Hendriksen, R. S. Genomic Epidemiology: Whole-Genome-Sequencing–Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annu. Rev. Food Sci. Technol.* **7**, 353–374 (2016).

49. Rodriguez-Manzano, J. *et al.* Adenovirus and Norovirus Contaminants in Commercially Distributed Shellfish. *Food Environ. Virol.* **6**, 31–41 (2014).

50. Ahmed, S. M. *et al.* Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis. *Lancet. Infect. Dis.* **14**, 725–30 (2014).

51. Havelaar, A. H. *et al.* World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010. *PLoS Med.* **12**, e1001923 (2015).

52. Pires, S. M. *et al.* Attributing the Human Disease Burden of Foodborne Infections to Specific Sources. *Foodborne Pathog. Dis.* **6**, 417–424 (2009).

53. Hald, T., Lo Fo Wong, D. M. A. & Aarestrup, F. M. The Attribution of Human Infections with Antimicrobial Resistant *Salmonella* Bacteria in Denmark to Sources of Animal Origin. *Foodborne Pathog. Dis.* **4**, 313–326 (2007).

54. Verhoef, L. *et al.* Norovirus Genotype Profiles Associated with Foodborne Transmission, 1999–2012. *Emerg. Infect. Dis.* **21**, 592–599 (2015).

55. Kazama, S. *et al.* Temporal dynamics of norovirus determined through monitoring of municipal wastewater by pyrosequencing and virological surveillance of gastroenteritis cases. *Water Res.* **92**, 244–253 (2016).

56. Tao, Z. *et al.* Environmental Surveillance of Genogroup I and II Noroviruses in Shandong Province, China in 2013. *Sci. Rep.* **5**, 17444 (2015).

57. Iritani, N. *et al.* Detection and genetic characterization of human enteric viruses in oyster-associated gastroenteritis outbreaks between 2001 and 2012 in Osaka City, Japan. *J. Med. Virol.* **86**, 2019–2025 (2014).

58. Wang, Y., Zhang, J. & Shen, Z. *The impact of calicivirus mixed infection in an oyster-associated outbreak during a food festival*. *Journal of Clinical Virology* **73**, (2015).

59. Newell, D. G. *et al.* Food-borne diseases - the challenges of 20 years ago still persist while new ones continue to emerge. *Int. J. Food Microbiol.* **139 Suppl**, S3-15 (2010).

60. Gossner, C., Gossner, C. & Severi, E. Three simultaneous, food-borne, multi-country outbreaks of hepatitis A virus infection reported in EPIS-FWD in 2013: what does it mean for the European Union? *Eurosurveillance* **19**, 20941 (2014).

61. Bruni, R. *et al.* Key role of sequencing to trace hepatitis a viruses circulating in Italy during a large multi-country European foodborne outbreak in 2013. *PLoS One* **11**, e0149642 (2016).

62. Tavoschi, L. *et al.* Food-borne diseases associated with frozen berries consumption: a historical perspective, European Union, 1983 to 2013. *Eurosurveillance* **20**, 21193 (2015).

63. Guillois, Y. *et al.* High Proportion of Asymptomatic Infections in an Outbreak of Hepatitis E Associated With a Spit-Roasted Piglet, France, 2013. *Clin. Infect. Dis.* **62**, 351–7 (2016).

64. Lewis, H. C., Wichmann, O. & Duizer, E. Transmission routes and risk factors for autochthonous hepatitis E virus infection in Europe: a systematic review. *Epidemiol. Infect.* **138**, 145 (2010).

65. Tei, S., Kitajima, N., Takahashi, K. & Mishiro, S. Zoonotic transmission of hepatitis E virus from deer to human beings. *Lancet* **362**, 371–373 (2003).

66. Izopet, J. *et al.* Hepatitis E Virus Strains in Rabbits and Evidence of a Closely Related Strain in Humans, France. *Emerg. Infect. Dis.* **18**, (2012).

67. Teixeira, J. *et al.* Prevalence of hepatitis E virus antibodies in workers occupationally exposed to swine in Portugal. *Med. Microbiol. Immunol.* 1–5 (2016). doi:10.1007/s00430-016-0484-8

68. Di Bartolo, I., Angeloni, G., Ponterio, E., Ostanello, F. & Ruggeri, F. M. Detection of hepatitis E virus in pork liver sausages. *Int. J. Food Microbiol.* **193**, 29–33 (2015).

69. Karesh, W. B. & Noble, E. The Bushmeat Trade: Increased Opportunities for Transmission of Zoonotic Disease. *Mt. Sinai J. Med. A J. Transl. Pers. Med.* **76**, 429–434 (2009).

70. Rahman, M. A. *et al.* Date Palm Sap Linked to Nipah Virus Outbreak in Bangladesh, 2008. *Vector-Borne Zoonotic Dis.* **12**, 65–72 (2012).

71. Funk, A. L. *et al.* MERS-CoV at the Animal–Human Interface: Inputs on Exposure Pathways from an Expert-Opinion Elicitation. *Front. Vet. Sci.* **3**, (2016).

72. Islam, M. S. *et al.* Nipah Virus Transmission from Bats to Humans Associated with Drinking Traditional Liquor Made from Date Palm Sap, Bangladesh, 2011-2014. *Emerg. Infect. Dis.* **22**, 664–70 (2016).

73. Wolfe, N. D., Daszak, P., Kilpatrick, A. M. & Burke, D. S. Bushmeat Hunting, Deforestation, and Prediction of Zoonotic Disease. *Emerg. Infect. Dis.* **11**, 1822–1827 (2005).

74. Mann, E. *et al.* A Review of the Role of Food and the Food System in the Transmission and Spread of Ebolavirus. *PLoS Negl. Trop. Dis.* **9**, e0004160 (2015).

75. McCloskey, B., Dar, O., Zumla, A. & Heymann, D. L. Emerging infectious diseases and pandemic potential: status quo and reducing risk of global spread. *Lancet Infect. Dis.* **14**, 1001–1010 (2014).

76. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–17 (2012).

77. Zhang, T. *et al.* RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* **4**, e3 (2006).

78. Colson, P. *et al.* Pepper Mild Mottle Virus, a Plant Virus Associated with Specific Immune Responses, Fever, Abdominal Pains, and Pruritus in Humans. *PLoS One* **5**, e10041 (2010).

79. Honda, K. & Littman, D. R. The microbiota in adaptive immune homeostasis and disease. *Nature* **535**, 75–84 (2016).

80. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669–85 (2004).

81. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).

82. Hellmér, M. *et al.* Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. *Appl. Environ. Microbiol.* **80**, 6771–81 (2014).

83. Ng, T. F. F. *et al.* High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J. Virol.* **86**, 12161–75 (2012).

84. Aw, T. G., Howe, A. & Rose, J. B. Metagenomic approaches for direct and cell culture evaluation of the virological quality of wastewater. *J. Virol. Methods* **210**, 15–21 (2014).

85. Furtak, V. *et al.* Environmental surveillance of viruses by tangential flow filtration and metagenomic reconstruction. *Eurosurveillance* **21**, 30193 (2016).

86. Cantalupo, P. G. *et al.* Raw sewage harbors diverse viral populations. *MBio* **2**, (2011).

87. Severi, E. *et al.* Large and prolonged food-borne multistate hepatitisA outbreak in Europe associated with consumption offrozen berries, 2013 to 2014. (2015).

88. Temmam, S. *et al.* Screening for Viral Pathogens in African Simian Bushmeat Seized at A French Airport. *Transbound. Emerg. Dis.* n/a-n/a (2016). doi:10.1111/tbed.12481

89. Bellou, M., Kokkinos, P. & Vantarakis, A. Shellfish-Borne Viral Outbreaks: A Systematic Review. *Food Environ. Virol.* **5**, 13–23 (2013).

90. Benabbes, L. *et al.* Norovirus and Other Human Enteric Viruses in Moroccan Shellfish. *Food Environ. Virol.* **5**, 35–40 (2013).

91. Alavandi, S. V & Poornima, M. Viral metagenomics: a tool for virus discovery and diversity in aquaculture. *Indian J. Virol.* **23**, 88–98 (2012).

92. Greninger, A. L. *et al.* Rapid metagenomic next-generation sequencing during an investigation of hospital-acquired human parainfluenza virus 3 infections. *J. Clin. Microbiol.* JCM.01881-16 (2016). doi:10.1128/JCM.01881-16

93. Hoenen, T. *et al.* Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerg. Infect. Dis.* **22**, 331–4 (2016).

94. Arias, A. *et al.* Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* **2**, vew016 (2016).

95. Sardi, S. I. *et al.* Coinfections of Zika and Chikungunya Viruses in Bahia, Brazil, Identified by Metagenomic Next-Generation Sequencing. *J. Clin. Microbiol.* **54**, 2348–53 (2016).

96. Greninger, A. L. *et al.* Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* **7**, 99 (2015).

97. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science (80-.).* **345**, 1369–1372 (2014).

98. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).

99. Quick, J. *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* **16**, 114 (2015).

100. Conceição-Neto, N. *et al. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis*. *Scientific reports* **5**, (nature.com, 2015).

101. Oude Munnink, B. B. *et al.* Autologous antibody capture to enrich immunogenic viruses for viral discovery. *PLoS One* **8**, e78454 (2013).

102. Gruber, K. Here, there, and everywhere: From PCRs to next-generation sequencing technologies and sequence databases, DNA contaminants creep in from the most unlikely places. *EMBO Rep.* **16**, 898–901 (2015).

103. Lusk, R. W. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One* **9**, e110808 (2014).

104. Schmieder, R. & Edwards, R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS One* **6**, e17288 (2011).

105. Spjuth, O. *et al.* Recommendations on e-infrastructures for next-generation sequencing. *Gigascience* **5**, 26 (2016).

106. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–1 (2010).

107. Wood, D. E. *et al.* Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

108. Menzel, P. *et al.* Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).

109. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

110. Vázquez-Castellanos, J. F. *et al.* Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* **15**, 37 (2014).

111. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. metaSPAdes: a new versatile de novo metagenomics assembler. *arXiv Prepr. arXiv1604.03071* (2016).

112. Boisvert, S. *et al.* Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13**, R122 (2012).

113. Treangen, T. J. *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* **14**, R2 (2013).

114. Afiahayati *et al.* MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Res.* **22**, 69–77 (2015).

115. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–8 (2012).

116. Afshinnekoo, E. *et al.* Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst.* **1**, 72–87 (2015).

117. La Rosa, G. *et al.* Surveillance of hepatitis A virus in urban sewages and comparison with cases notified in the course of an outbreak, Italy 2013. *BMC Infect. Dis.* **14**, 419 (2014).

118. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).

119. Flygare, S. *et al.* Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.* **17**, 111 (2016).

120. Gomes, M. F. C. *et al.* Assessing the international spreading risk associated with the 2014 west african ebola outbreak. *PLoS Curr.* **6**, (2014).

121. Koopmans, M. *et al.* Familiar barriers still unresolved—a perspective on the Zika virus outbreak research response. *Lancet Infect. Dis.* **19**, e59–e62 (2019).

122. Thézé, J. *et al.* Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico. *Cell Host Microbe* **23**, 855-864.e7 (2018).

123. Glennon, E. E., Jephcott, F. L., Restif, O. & Wood, J. L. N. Estimating undetected Ebola spillovers. *PLoS Negl. Trop. Dis.* **13**, e0007428 (2019).

124. Peeling, R. W., Murtagh, M. & Olliaro, P. L. Epidemic preparedness: why is there a need to accelerate the development of diagnostics? *Lancet Infect. Dis.* **19**, e172–e178 (2019).

125. Nieuwenhuijse, D. F. & Koopmans, M. P. G. G. Metagenomic Sequencing for Surveillance of Food- and Waterborne Viral Diseases. *Front. Microbiol.* **8**, 230 (2017).

126. Chan, E. H. *et al.* Global capacity for emerging infectious disease detection. *Proc. Natl. Acad. Sci.* (2010). doi:10.1073/pnas.1006219107

127. World Health Organization. A research and development Blueprint for action to prevent epidemics. (2019). Available at: https://www.who.int/blueprint/en/.

128. Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10–19 (2019).

129. Aarestrup, F. M. *et al.* Integrating Genome-based Informatics to Modernize Global Disease Monitoring, Information Sharing, and Response. *Emerg. Infect. Dis.* **18**, e1–e1 (2012).

130. Holmes, E. C., Rambaut, A. & Andersen, K. G. Pandemics: spend on surveillance, not prediction. *Nature* **558**, 180–182 (2018).

131. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).

132. Neiderud, C. J. How urbanization affects the epidemiology of emerging infectious diseases. *African J. Disabil.* (2015). doi:10.3402/iee.v5.27060

133. Callender, D. M. Factors contributing to and strategies to combat emerging arboviruses. *Global Public Health* (2018). doi:10.1080/17441692.2018.1464588

134. Van Der Avoort, H. G. A. M., Reimerink, J. H. J., Ras, A., Mulders, M. N. & Van Loon, A. M. Isolation of epidemic poliovirus from sewage during the 1992–3 type 3 outbreak in the Netherlands. *Epidemiol. Infect.* (1995). doi:10.1017/S0950268800052195

135. Asghar, H. *et al.* Environmental surveillance for polioviruses in the global polio eradication initiative. *J. Infect. Dis.* (2014). doi:10.1093/infdis/jiu384

136. Kaliner, E. *et al.* The Israeli public health response to wild poliovirus importation. *Lancet Infect. Dis.* **15**, 1236–1242 (2015).

137. Niedrig, M., Patel, P., El Wahed, A. A., Schädler, R. & Yactayo, S. Find the right sample: A study on the versatility of saliva and urine samples for the diagnosis of emerging viruses. *BMC Infect. Dis.* **18**, 707 (2018).

138. Gourinat, A.-C., O'Connor, O., Calvez, E., Goarant, C. & Dupont-Rouzeyrol, M. Detection of Zika Virus in Urine. *Emerg. Infect. Dis.* **21**, 84–86 (2015).

139. Benschop, K. S. M. *et al.* Polio and Measles Down the Drain: Environmental Enterovirus Surveillance in the Netherlands, 2005 to 2015. *Appl. Environ. Microbiol.* **83**, (2017).

140. Drosten, C. *et al.* Clinical features and virological analysis of a case of Middle East respiratory syndrome coronavirus infection. *Lancet. Infect. Dis.* **13**, 745–51 (2013).

141. Wang, X.-W. *et al.* Concentration and detection of SARS coronavirus in sewage from Xiao Tang Shan Hospital and the 309th Hospital. *J. Virol. Methods* **128**, 156–161 (2005).

142. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, 1124 (2019).

143. Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243 (2017).

144. McMinn, B. R., Ashbolt, N. J. & Korajkic, A. Bacteriophages as indicators of faecal pollution and enteric virus removal. *Lett. Appl. Microbiol.* **65**, 11–26 (2017).

145. Calero-Cáceres, W. & Balcázar, J. L. Antibiotic resistance genes in bacteriophages from diverse marine habitats. *Sci. Total Environ.* **654**, 452–455 (2019).

146. Rosario, K., Symonds, E. M., Sinigalliano, C., Stewart, J. & Breitbart, M. Pepper mild mottle virus as an indicator of fecal pollution. *Appl. Environ. Microbiol.* **75**, 7261–7 (2009).

147. Tijssen, P., Pénzes, J. J., Yu, Q., Pham, H. T. & Bergoin, M. Diversity of small, single-stranded DNA viruses of invertebrates and their chaotic evolutionary past. *J. Invertebr. Pathol.* **140**, 83–96 (2016).

148. Boonnak, K., Suttitheptumrong, A., Jotekratok, U. & Pattanakitsakul, S.-N. Phylogenetic analysis reveals genetic variations of densovirus isolated from field mosquitoes in bangkok and surrounding regions. *Southeast Asian J. Trop. Med. Public Health* **46**, 207–14 (2015).

149. Ng, T. F. F. *et al.* Broad Surveys of DNA Viral Diversity Obtained through Viral Metagenomics of Mosquitoes. *PLoS One* **6**, e20579 (2011).

150. Palinski, R. *et al.* A novel porcine circovirus distantly related to known circoviruses is associated with porcine dermatitis and nephropathy syndrome and reproductive failure. *J VIROL* **91**, (2017).

151. Vu, D.-L. L., Cordey, S., Brito, F. & Kaiser, L. *Novel human astroviruses: Novel human diseases? Journal of Clinical Virology* **82**, 56–63 (Elsevier, 2016).

152. van Beek, J. *et al.* Molecular surveillance of norovirus, 2005–16: an epidemiological analysis of data collected from the NoroNet network. *Lancet Infect. Dis.* **18**, 545–553 (2018).

153. Ahmed, S. M., Lopman, B. A. & Levy, K. A Systematic Review and Meta-Analysis of the Global Seasonality of Norovirus. *PLoS One* **8**, e75922 (2013).

154. Thongprachum, A., Khamrin, P., Maneekarn, N., Hayakawa, S. & Ushijima, H. Epidemiology of gastroenteritis viruses in Japan: Prevalence, seasonality, and outbreak. *J. Med. Virol.* **88**, 551–570 (2016).

155. Pons-Salort, M. *et al.* The seasonality of nonpolio enteroviruses in the United States: Patterns and drivers. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3078–3083 (2018).

156. Hjelmsø, M. H. *et al.* Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing. *PLoS One* **12**, e0170199 (2017).

157. Aiewsakun, P. & Simmonds, P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. *Microbiome* **6**, 38 (2018).

158. Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* (2016). doi:10.1038/nature20167

159. Krishnamurthy, S. R. & Wang, D. Origins and challenges of viral dark matter. *Virus Res.* **239**, 136–142 (2017).

160. Schaeffer, J. *et al.* Improving the efficacy of sewage treatment decreases norovirus contamination in oysters. *Int. J. Food Microbiol.* **286**, 1–5 (2018).

161. Endoh, D. *et al.* Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription. *Nucleic Acids Res* **33**, e65 (2005).

162. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

163. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts565

164. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).

165. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* (2008). doi:10.1038/nrg2323

166. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).

167. Kahlke, T. & Ralph, P. J. BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods Ecol. Evol.* **10**, 100–103 (2019).

168. Solonenko, S. A. *et al.* Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14**, 320 (2013).

169. Gomez-Alvarez, V., Teal, T. K. & Schmidt, T. M. Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* **3**, 1314–1317 (2009).

170. Legendre, P. & Gallagher, E. D. Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**, 271–280 (2001).

171. Oksanen, J. *et al.* vegan: Community Ecology Package. (2019).

172. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).

173. Rubel, F. & Kottek, M. Observed and projected climate shifts 1901-2100 depicted by world maps of the Köppen-Geiger climate classification. **19**, 135–141 (2010).

174. Xu, B. *et al.* Metagenomic analysis of the Rhinopithecus bieti fecal microbiome reveals a broad diversity of bacterial and glycoside hydrolase profiles related to lignocellulose degradation. *BMC Genomics* **16**, 174 (2015).

175. Wang, Q., Jia, P. & Zhao, Z. VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. *PLoS One* **8**, (2013).

176. Naccache, S. N. *et al.* A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* **24**, 1180–1192 (2014).

177. Ho, T. & Tzanetakis, I. E. Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **471**–**473**, 54–60 (2014).

178. Li, Y. *et al.* VIP: An integrated pipeline for metagenomics of virus identification and discovery. *Sci. Rep.* **6**, 1–10 (2016).

179. Brinkmann, A. *et al.* Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing data sets. *J. Clin. Microbiol.* **57**, 466–485 (2019).

180. Junier *et al.* Viral Metagenomics in the Clinical Realm: Lessons Learned from a Swiss-Wide Ring Trial. *Genes (Basel).* **10**, 655 (2019).

181. 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. Diagnostics* **22**, 196–207 (2020).

182. Huson, D. H. *et al.* MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Comput. Biol.* **12**, e1004957 (2016).

183. Breitwieser, F. P. & Salzberg, S. L. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *bioRxiv* 084715 (2016). doi:10.1101/084715

184. Pedersen, T. L., Nookaew, I., Wayne Ussery, D. & Månsson, M. PanViz: interactive visualization of the structure of functionally annotated pangenomes. *Bioinformatics* btw761 (2017). doi:10.1093/bioinformatics/btw761

185. Wagner, J. *et al.* Metaviz: interactive statistical and visual analysis of metagenomic data. *Nucleic Acids Res.* **46**, 2777–2787 (2018).

186. Eren, A. M. *et al.* Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ* **2015**, (2015).

187. RStudio, I. shiny: Web Application Framework for R. (2020). Available at: https://shiny.rstudio.com/.

188. Hafen, R. & Continuum Analytics, Inc. rbokeh: R Interface for Bokeh. (2020).

189. Morgan, M., Pages, H., Obenchain, V. & Hayden, N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. *R Packag. version* **1**, 677–689 (2016).

190. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. (2019).

191. Wickham, H. *R Packages*. (O'Reilly Media, Inc, Usa, 2015).

192. Vilsker, M. *et al.* Genome Detective: An automated system for virus identification from high-throughput sequencing data. *Bioinformatics* (2019). doi:10.1093/bioinformatics/bty695

193. Amid, C. *et al.* The COMPARE Data Hubs. *Database (Oxford).* **2019**, 136 (2019).

194. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406–410 (2017).

195. Cadar, D. *et al.* Widespread activity of multiple lineages of Usutu virus, Western Europe, 2016. *Eurosurveillance* **22**, (2017).

196. Paz, S. & Semenza, J. C. Environmental drivers of West Nile fever epidemiology in Europe and Western Asia--a review. *International journal of environmental research and public health* **10**, 3543–3562 (2013).

197. Sikkema, R. S. *et al.* Detection of West Nile virus in a common whitethroat (Curruca communis) and Culex mosquitoes in the Netherlands, 2020. *Euro Surveill.* **25**, 1–6 (2020).

198. Folly, A. J. *et al.* Detection of Usutu virus infection in wild birds in the United Kingdom, 2020. *Euro Surveill.* **25**, (2020).

199. Javelle, E., Gautret, P. & Raoult, D. Towards the risk of yellow fever transmission in Europe. *Clinical Microbiology and Infection* **25**, 10–12 (2019).

200. Faria, N. R. *et al.* Genomic and epidemiological monitoring of yellow fever virus transmission potential. *bioRxiv* 299842 (2018). doi:10.1101/299842

201. Oude Munnink, B. B. *et al.* Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat. Med.* **26**, 1405–1410 (2020).

202. Grubaugh, N. D., Faria, N. R., Andersen, K. G. & Pybus, O. G. Genomic Insights into Zika Virus Emergence and Spread. *Cell* **172**, 1160–1162 (2018).

203. Lessler, J. *et al.* Assessing the global threat from Zika virus. *Science* **353**, (2016).

204. Lustig, Y., Sofer, D., Bucris, E. D. & Mendelson, E. Surveillance and Diagnosis of West Nile Virus in the Face of Flavivirus Cross-Reactivity. *Front. Microbiol.* **9**, 2421 (2018).

205. Thomas, S. J., Endy, T. P., Rothman, A. L. & Barrett, A. D. Flaviviruses (Dengue, Yellow Fever, Japanese Encephalitis, West Nile Encephalitis, St. Louis Encephalitis, Tick-Borne Encephalitis, Kyasanur Forest Disease, Alkhurma Hemorrhagic Fever, Zika). *Mand. Douglas, Bennett's Princ. Pract. Infect. Dis.* **2**, 1881-1903.e6 (2015).

206. Goldani, L. Z. Yellow fever outbreak in Brazil, 2017. *Brazilian J. Infect. Dis.* **21**, 123–124 (2017).

207. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).

208. Nieuwenhuijse, D. F. *et al.* Setting a baseline for global urban virome surveillance in sewage. *Sci. Rep.* **10**, 1–13 (2020).

209. Magi, A., Giusti, B. & Tattini, L. Characterization of MinION nanopore data for resequencing analyses. *Brief. Bioinform.* **18**, bbw077 (2016).

210. Oude Munnink, B. B., Nieuwenhuijse, D. F., Sikkema, R. S. & Koopmans, M. Validating Whole Genome Nanopore Sequencing, using Usutu Virus as an Example. *J. Vis. Exp.* **2020**, e60906 (2020).

211. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).

212. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).

213. Oude Munnink, B. B. *et al.* The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat. Med. 2021 279* **27**, 1518–1524 (2021).

214. Harvey, E. & Holmes, E. C. Diversity and evolution of the animal virome. *Nat. Rev. Microbiol. 2022* 1–14 (2022). doi:10.1038/s41579-021-00665-x

215. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 1–19 (2019).

216. Neill, J. D., Bayles, D. O. & Ridpath, J. F. Simultaneous rapid sequencing of multiple RNA virus genomes. *J VIROL METHODS* **201**, 68–72 (2014).

217. Faria, N. R. *et al.* Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science (80-.).* **361**, 894–899 (2018).

218. Lim, S. M., Koraka, P., Osterhaus, A. D. M. E. & Martina, B. E. E. Development of a strand-specific real-time qRT-PCR for the accurate detection and quantitation of West Nile virus RNA. *J. Virol. Methods* **194**, 146–153 (2013).

219. Lanciotti, R. S. *et al.* Genetic and Serologic Properties of Zika Virus Associated with an Epidemic, Yap State, Micronesia, 2007. *Emerg. Infect. Dis.* **14**, 1232 (2008).

220. Domingo, C. *et al.* Advanced yellow fever virus genome detection in point-of-care facilities and reference laboratories. *J. Clin. Microbiol.* **50**, 4054–4060 (2012).

221. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

222. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

223. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

224. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

225. Bonaldo, M. C. *et al.* Genome analysis of yellow fever virus of the ongoing outbreak in Brazil reveals polymorphisms. *Mem Inst Oswaldo Cruz* **112**, 447–451 (2017).

226. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).

227. Nanopore Store, R10 flow cells. Available at: https://store.nanoporetech.com/flowcells/spoton-flow-cell-mk-i-r10.html.

228. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).

229. GitHub – nanoporetech/flappie: Flip-flop basecaller for Oxford Nanopore reads. Available at: https://github.com/nanoporetech/flappie.

230. Lühken, R. *et al.* Distribution of Usutu Virus in Germany and Its Effect on Breeding Bird Populations. *Emerg. Infect. Dis.* **23**, 1994–2001 (2017).

231. Becker, N. *et al.* Epizootic emergence of Usutu virus in wild and captive birds in Germany. *PLoS One* **7**, (2012).

232. Diagne, M. *et al.* Usutu Virus Isolated from Rodents in Senegal. *Viruses* **11**, 181 (2019).

233. Bakonyi, T. *et al.* Usutu virus infections among blood donors, Austria, July and August 2017 – Raising awareness for diagnostic challenges. *Eurosurveillance* **22**, (2017).

234. Zaaijer, H. L., Slot, E., Molier, M., Reusken, C. B. E. M. & Koppelman, M. H. G. M. Usutu virus infection in Dutch blood donors. *Transfusion* trf.15444 (2019). doi:10.1111/trf.15444

235. Cadar, D. *et al.* Blood donor screening for West Nile virus (WNV) revealed acute Usutu virus (USUV) infection, Germany, September 2016. *Eurosurveillance* **22**, 30501 (2017).

236. Pierro, A. *et al.* Detection of specific antibodies against West Nile and Usutu viruses in healthy blood donors in northern Italy, 2010–2011. *Clin. Microbiol. Infect.* **19**, E451–E453 (2013).

237. Pecorari, M. *et al.* First human case of Usutu virus neuroinvasive infection, Italy, August-September 2009. *Euro Surveill.* **14**, (2009).

238. Simonin, Y. *et al.* Human Usutu Virus Infection with Atypical Neurologic Presentation, Montpellier, France, 2016. *Emerg. Infect. Dis.* **24**, 875–878 (2018).

239. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).

240. R.R. Wick. GitHub - rrwick/Porechop: adapter trimmer for Oxford Nanopore reads. Available at: https://github.com/rrwick/porechop.

241. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10 (2011).

242. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

243. BBMap download | SourceForge.net. Available at: https://sourceforge.net/projects/bbmap/.

244. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455–477 (2012).

245. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).

246. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* **5**, 536–544 (2020).

247. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **0**, 533–534 (2020).

248. Oude Munnink, B. B. *et al.* Genomic monitoring to understand the emergence and spread of Usutu virus in the Netherlands, 2016-2018. *Sci. Rep.* **10**, 2798 (2020).

249. Khoury, M. J. *et al.* From public health genomics to precision public health: A 20-year journey. *Genetics in Medicine* **20**, 574–582 (2018).

250. Armstrong, G. L. *et al.* Pathogen Genomics in Public Health. *N. Engl. J. Med.* **381**, 2569–2580 (2019).

251. Polonsky, J. A. *et al.* Outbreak analytics: A developing data science for informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**, 20180276 (2019).

252. Modjarrad K Millett P, Gsell PS, Roth C, Kieny MS, M. V. S. Developing Global Norms for Sharing Data and Results during Public Health Emergencies. *PLOS Pathog.* **13**, (2016).

253. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challenges* **1**, 33–46 (2017).

254. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, (2017).

255. Covid-19 < European Bioinformatics Institute. Available at: https://www.ebi.ac.uk/covid-19.

256. Kraaij – Dirkzwager, M. *et al.* Middle East respiratory syndrome coronavirus (MERS-CoV) infections in two returning travellers in the Netherlands, May 2014. *Eurosurveillance* **19**, 20817 (2014).

257. Timen, A. Response to Imported Case of Marburg Hemorrhagic Fever, the Netherlands. *Emerg. Infect. Dis.* **15**, 1171–1175 (2009).

258. Corman, V. M. *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* **25**, 2000045 (2020).

259. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).

260. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* (2020). doi:10.1093/molbev/msaa015

261. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, (2016).

262. Ayres, D. L. *et al.* BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. *Syst. Biol* **61**, 170–173 (2012).

263. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016 (2018).

264. Hasegawa, M., Kishino, H. & Yano, T. aki. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).

265. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

266. Kluytmans-van den Bergh, M. F. Q. Q. *et al.* Prevalence and Clinical Presentation of Health Care Workers With Symptoms of Coronavirus Disease 2019 in 2 Dutch Hospitals During an Early Phase of the Pandemic. *JAMA Netw Open* **3**, e209673 (2020).

267. WHO. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). Available at: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.

268. WHO. WHO Regional Office for the Eastern Mediterranean. MERS situation update. (2020). Available at: http://www.emro.who.int/%0Apandemic-epidemic-diseases/mers-cov/mers-situation-updatejanuary-%0A2020.html.

269. Christian E.A. Alderweireld *et al.* COVID-19: patiënt nul in Nederland | Nederlands Tijdschrift voor Geneeskunde. *Ned Tijdschr Geneeskd* (2020).

270. Kluytmans, M. *et al.* SARS-CoV-2 infection in 86 healthcare workers in two Dutch hospitals in March 2020. *medRxiv* 2020.03.23.20041913 (2020). doi:10.1101/2020.03.23.20041913

271. Reusken, C. B. *et al.* Rapid assessment of regional SARS-CoV-2 community transmission through a convenience sample of healthcare workers, the Netherlands, March 2020. *Eurosurveillance* **25**, 2000334 (2020).

272. NOS. Care threatens to be squeezed by a shortage of mouth masks. (2020). Available at: https://nos.nl/artikel/2324830-zorg-dreigt-in-de-knel-te-komen-door-tekort-aan-mondkapjes.

273. Rijksinstituut voor Volksgezondheid en Milieu. Guidance on infection prevention for hospitals: isolation guidlines. (2011). Available at: https://www.rivm.nl/wip-richtlijn-strikte-isolatie-zkh.

274. Rijksinstituut voor Volksgezondheid en Milieu. Personal protective equipment directive. (2015). Available at: https://www.rivm.nl/%0Adocumenten/wip-richtlijn-persoonlijke-beschermingsmiddelen-zkh.

275. Hoek, R. A. S. *et al.* Incidence of viral respiratory pathogens causing exacerbations in adult cystic fibrosis patients. *Scand. J. Infect. Dis.* **45**, 65–69 (2013).

276. van der Vries, E. *et al.* Molecular assays for quantitative and qualitative detection of influenza virus and oseltamivir resistance mutations. *J Mol Diagn* **15**, 347–354 (2013).

277. ISO. Medical laboratories: requirements for quality and competence. (2012). Available at: https://www.iso.org/standard/56115.%0Ahtml.

278. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).

279. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018).

280. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).

281. Hill, V. & Baele, G. Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. *Mol Biol Evol* **36**, 2620–2628 (2019).

282. McMichael, T. M. *et al.* Epidemiology of Covid-19 in a Long-Term Care Facility in King County, Washington. *N. Engl. J. Med.* (2020). doi:10.1056/NEJMoa2005412

283. Zhan, M., Qin, Y., Xue, X. & Zhu, S. Death from Covid-19 of 23 Health Care Workers in China. *N. Engl. J. Med.* (2020). doi:10.1056/NEJMc2005696

284. Durante-Mangoni, E. *et al.* Low rate of severe acute respiratory syndrome coronavirus 2 spread among health-care personnel using ordinary personal protection equipment in a medium-incidence setting. *Clin Microbiol Infect* (2020).

285. Houlihan, C. F. *et al.* Use of Whole-Genome Sequencing in the Investigation of a Nosocomial Influenza Virus Outbreak. *J Infect Dis* **218**, 1485–1489 (2018).

286. Taiaroa, G. *et al.* Direct RNA sequencing and early evolution of SARS-CoV-2. *bioRxiv* 2020.03.05. 976167 (2020). doi:10.1101/2020.03.05.976167

287. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.* (2020). doi:10.1056/nejmoa2001316

288. Gorbalenya, A. E. *et al.* Severe acute respiratory syndrome-related coronavirus: The species and its viruses-a statement of the Coronavirus Study Group. doi:10.1101/2020.02.07.937862

289. Pongpirul, W. A., Pongpirul, K., Ratnarathon, A. C. & Prasithsirikul, W. Journey of a Thai Taxi Driver and Novel Coronavirus. *N. Engl. J. Med.* NEJMc2001621 (2020). doi:10.1056/NEJMc2001621

290. Phan, L. T. *et al.* Importation and Human-to-Human Transmission of a Novel Coronavirus in Vietnam. *N. Engl. J. Med.* (2020). doi:10.1056/nejmc2001272

291. Rothe, C. *et al.* Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. *N. Engl. J. Med.* (2020). doi:10.1056/nejmc2001468

292. Coronavirus disease 2019. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. (Accessed: 18th August 2020)

293. Laboratory guidance. Available at: https://www.who.int/publications/m/item/molecular-assays-to-diagnose-covid-19-summary-table-of-available-protocols. (Accessed: 18th August 2020)

294. Ziegler, K. *et al.* SARS-CoV-2 samples may escape detection because of a single point mutation in the N gene. *Eurosurveillance* **25**, 2001650 (2020).

295. Artesi, M. *et al.* A recurrent mutation at position 26340 of SARS-CoV-2 is associated with failure of the E Gene quantitative reverse transcription-PCR utilized in a commercial dual-target diagnostic assay. *J. Clin. Microbiol.* **58**, (2020).

296. Hall, R. J., Draper, J. L., Nielsen, F. G. G. G. & Dutilh, B. E. Beyond research: a primer for considerations on using viral metagenomics in the field and clinic. *Front. Microbiol.* **6**, 224 (2015).

297. Towner, J. S. *et al.* Newly Discovered Ebola Virus Associated with Hemorrhagic Fever Outbreak in Uganda. *PLOS Pathog.* **4**, e1000212 (2008).

298. Marra, M. A. *et al.* The genome sequence of the SARS-associated coronavirus. *Science (80-.).* **300**, 1399–1404 (2003).

299. Stalin Raj, V. *et al.* Isolation of MERS Coronavirus from a Dromedary Camel, Qatar, 2014. *Emerg. Infect. Dis.* **20**, 1339 (2014).

300. Munnink, B. B. O. *et al.* A novel astrovirus-like RNA virus detected in human stool. *Virus Evol.* **2**, vew005 (2016).

301. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biol.* **9**, e1001177 (2011).

302. Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLOS Biol.* **19**, e3001135 (2021).

303. Izquierdo-Lara, R. *et al.* Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium. *Emerg. Infect. Dis.* **27**, 1405 (2021).

304. Wylie, T. N., Wylie, K. M., Herter, B. N. & Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **25**, 1910–20 (2015).

305. Crits-Christoph, A. *et al.* Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. *MBio* **12**, 1–9 (2021).

306. Rothman, J. A. *et al.* RNA viromics of Southern California wastewater and detection of SARS-CoV-2 single-nucleotide variants. *Appl. Environ. Microbiol.* **87**, (2021).

307. Brouwer, A. F. *et al.* Epidemiology of the silent polio outbreak in Rahat, Israel, based on modeling of environmental surveillance data. *Proc. Natl. Acad. Sci.* **115**, E10625–E10633 (2018).

308. van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L. & Guschanski, K. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Mol. Ecol. Resour.* **20**, 1171–1181 (2020).

309. Porter, A. F. *et al.* Metagenomic Identification of Viral Sequences in Laboratory Reagents. (2021). doi:10.3390/v13112122

310. Naccache, S. N. *et al.* The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol* **87**, 11966–11977 (2013).

311. Kufner, V. *et al.* Two Years of Viral Metagenomics in a Tertiary Diagnostics Unit: Evaluation of the First 105 Cases. *Genes 2019, Vol. 10, Page 661* **10**, 661 (2019).

312. Wang, H. *et al.* Clinical diagnostic application of metagenomic next-generation sequencing in children with severe nonresponding pneumonia. *PLoS One* **15**, e0232610 (2020).

313. Wilson, M. R. *et al.* Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N. Engl. J. Med.* **380**, 2327–2340 (2019).

314. López-Labrador, F. X. *et al.* Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure. *J. Clin. Virol.* **134**, 104691 (2021).

315. de Vries, J. J. C. *et al.* Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J. Clin. Virol.* **141**, 104908 (2021).

316. Hess, J. F. *et al.* Library preparation for next generation sequencing: A review of automation strategies. *Biotechnol. Adv.* **41**, 107537 (2020).

317. Murphy, T. W., Hsieh, Y. P., Zhu, B., Naler, L. B. & Lu, C. Microfluidic platform for next-generation sequencing library preparation with low-input samples. *Anal. Chem.* **92**, 2519–2526 (2020).

318. Lamprecht, A.-L. *et al.* Towards FAIR principles for research software. *Data Sci.* **3**, 37–59 (2020).

319. Jang, M. *Linux Annoyances for Geeks: Getting the Most Flexible System in the World*. (O'Reilly Media, Inc, 2006).

320. Köster, J. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, (2021).

321. DI Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol. 2017 354* **35**, 316–319 (2017).

322. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

323. Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2021).

324. Piro, V. C., Dadi, T. H., Seiler, E., Reinert, K. & Renard, B. Y. ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics* **36**, i12–i20 (2020).

325. Allesøe, R. L. *et al.* Automated download and clean-up of family-specific databases for kmer-based virus identification. *Bioinformatics* **37**, 705–710 (2021).

326. Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**, D593–D598 (2012).

327. Edgar, R. C. *et al.* Petabase-scale sequence alignment catalyses viral discovery. *Nat. 2022 6027895* **602**, 142–147 (2022).

328. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol. 2018 371* **37**, 29–37 (2018).

329. Ladner, J. T. *et al.* Standards for sequencing viral genomes in the era of high-throughput sequencing. *MBio* **5**, 1–5 (2014).

330. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, (2021).

331. Poen Id, M. J. *et al.* Comparison of sequencing methods and data processing pipelines for whole genome sequencing and minority single nucleotide variant (mSNV) analysis during an influenza A/H5N8 outbreak. (2020). doi:10.1371/journal.pone.0229326

332. Nicholls, S. M. *et al.* Probabilistic recovery of cryptic haplotypes from metagenomic data.

333. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol. 2019 410* **4**, 1727–1736 (2019).

334. Baaijens, J. A. *et al.* Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. *medRxiv* 2021.08.31.21262938 (2021). doi:10.1101/2021.08.31.21262938

335. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

336. Wilkinson, E. *et al.* A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science (80-.).* **374**, 423–431 (2021).

337. Holmes, E. C. *et al.* The origins of SARS-CoV-2: A critical review. *Cell* **184**, 4848–4856 (2021).

338. Chen, Z. *et al.* Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat. Genet. 2022 544* **54**, 499–507 (2022).

339. Inzaule, S. C., Tessema, S. K., Kebede, Y., Ogwell Ouma, A. E. & Nkengasong, J. N. Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *Lancet Infect. Dis.* **21**, e281–e289 (2021).

340. Ribeiro, C. dos S. *et al.* How ownership rights over microorganisms affect infectious disease control and innovation: A root-cause analysis of barriers to data sharing as experienced by key stakeholders. *PLoS One* **13**, e0195885 (2018).

341. Woolhouse, M. E. J., Rambaut, A. & Kellam, P. Lessons from Ebola: Improving infectious disease surveillance to inform outbreak management. *Sci. Transl. Med.* **7**, (2015).

342. Callaway, E. Microbiomes raise privacy concerns. *Nature* **521**, 136 (2015).

343. Yakowitz Bambauer, J. Tragedy of the Data Commons. *SSRN Electron. J.* (2011). doi:10.2139/SSRN.1789749

344. Escribano, N., Galicia, D. & Ariño, A. H. The tragedy of the biodiversity data commons: a data impediment creeping nigher? *Database* **2018**, 33 (2018).

345. Neumann, J. & Brase, J. DataCite and DOI names for research data. *J. Comput. Aided. Mol. Des.* **28**, 1035–1041 (2014).

Nederlands samenvatting

12

12.1 Metagenoom sequencen voor surveillance van voedsel- en water-overdraagbare virussen

De huidige surveillance van virale pathogenen is vooral gebaseerd op informatie van patiënten waarbij diagnostiek wordt aangevraagd door huisartsen of van patienten die zijn opgenomen het ziekenhuis. Daarmee geeft de surveillance een sterk verbogen beeld, aangezien de meeste infecties niet tot een bezoek aan de huisarts of tot ziekenhuisopname leiden. Bovendien is de surveillance beperkt tot bekende pathogenen waarvoor bestrijdingsprogramma's zijn ontwikkeld. Om beter zicht te krijgen op de circulatie van alle virale pathogenen in de populatie is daarom een andere aanpak nodig. Metagenoom sequencen, ookwel "shotgun" sequencen of "agnostisch" sequencen genoemd, is een methode waarbij al het DNA en RNA van een organisme wordt gesequenced zonder dat het genoom van het organisme vooraf bekend hoeft te zijn. Daarnaast kunnen meerdere pathogenen tegelijkertijd in kaart worden gebracht. Om die redenen wordt deze methode gebruikt voor generieke virus detectie. Een specifiek geval waar monitoring op basis van systematisch metagenoom sequencen van toepassing zou kunnen zijn is bij het testen van voedsel- en omgevingsmonsters op de aanwezigheid van pathogenen. In **hoofdstuk 2** verkennen we de potentie van metagenoom sequencen voor het monitoren van virussen die via voedsel en/of water verspreiden, en bespreken technische mogelijkheden en beperkingen van de huidige methoden, inclusief bioinformatica analyse en data visualisatie.

12.2 Een startpunt voor het wereldwijd monitoren van virussen in rioolwater

Naast gerichtte methodes is metagenoom sequencen van rioolwater met wisselend succes ingezet als toevoeging en uitbereiding bij het monitoren van virussen. Een van de basisvragen bij metagenoom sequencing van zulke complexe monsters is wat een "normale" samenstelling is van het zogenaamde viroom, en wat kan worden gezien als viraal "signaal" voor surveillance. Met dat doel is in **hoofdstuk 3** een wereldwijde cross-sectionele "foto" gemaakt door rond hetzelfde moment het viroom van rioolwater monsters van 62 grote steden verspreid over alle continenten van de wereld te sequencen. Dit liet seizoenspatronen zien van humane pathogene virussen en ook verschillen tussen de hoeveelheden insect en planten gerelateerde virussen in het noordelijk en het zuidelijk halfrond.

12.3 Visualisatie van complexe metagenoom sequencing datasets

De resultaten van metagenoom sequencing experimenten van monsters met een grote diversiteit aan virussen zijn lastig om te analyseren voor wetenschappers zonder programmeer kennis. Om dit te vergemakkelijken is in **hoofdstuk 4** een viroom browser applicatie ontwikkeld, waarmee gebruikers complexe metagenoom analyse resultaten kunnen doorzoeken met de hulp van interactieve kwaliteitsfilters. Daarnaast maakt de viroom browser het makkelijk om specifieke interessante sequenties uit de dataset te halen voor verder onderzoek. Ook kan een stratificatie worden gemaakt van de viroom annotatie data, om zo verschillende monsters met elkaar te vergelijken.

12.4 Vergelijken van verschillende sequencing methodes voor arbo-virussen

Uitbraakonderzoek maakt in toenemende mate gebruik van “whole genome sequencing” (WGS). Daarvoor kunnen meerdere sequencing methodes worden gebruikt, die variëren in sensitiviteit, doorlooptijd en kosten. In **hoofdstuk 5** beschrijven we een grondige vergelijking van amplicon-, agnostisch- of capture-gebaseerde sequencing methodes voor vier arbo-virussen. Uit deze vergelijking concluderen we dat alle methodes kunnen worden gebruik voor het detecteren van de aanwezigheid van deze virussen maar dat voor het verkrijgen van een volledig genoom een capture gebaseerde of een amplicon gebaseerde methode de voorkeur heeft. Deze vergelijking kan laboratoria helpen om een keuze te maken voor een sequencing methode passend bij de specifieke vraagstelling en de beschikbare middelen.

12.5 Methodes voor het valideren van nanopore gebaseerde genoom sequencing

In **hoofdstuk 6** wordt stap voor stap een methode voor de evaluatie van Nanopore sequencing resultaten uitgelegd in een video artikel. We gebruiken het sequencen van het Usutu virus als een voorbeeld en gebruiken Illumina sequencen als de gouden standaard validatie methode. Gebaseerd op de resultaten van het onderzoek is het mogelijk om minimale kwaliteits standaarden te zetten voor het verkrijgen van een betrouwbaar en volledig genoom en kunnen software settings zo worden gezet dat een goede kwaliteit consensus sequentie kan worden gegenereerd na het verwerken van de ruwe data. Het is echter wel zo dat door de snelle technologische verbeteringen in sequence kwaliteit en doorlooptijd regelmatig moet worden heroverwogen om de parameters aan te passen. Naar aanleiding van die verbeteringen kunnen er namelijk wellicht meer monsters tegelijk worden verwerkt tegen een lagere kostprijs, wat nodig is om routinematig sequencen mogelijk te maken binnen de moleculaire diagnostiek.

12.6 Bijna real-time SARS-CoV-2 sequencen in het begin van de pandemie

In het begin van 2020 dook een nieuw SARS-achtig coronavirus (SARS-CoV-2) op in de Wuhan provincie in China en verspreidde zich snel over de wereld. Doordat het volledige genoom snel werd gesequenced en gedeeld konden we vroegtijdig een amplicon gebaseerd sequencing protocol opzetten om het virus te sequencen. In maart 2020 werden de eerste gevallen van SARS-CoV-2 in Nederland gevonden door middel van een specifieke RT-PCR. In **hoofdstuk 7** beschrijven we hoe we door het sequencen van het meerendeel van deze gevallen een phylogenetische analyse konden uitvoeren die hielp om een data gedreven besluit te maken over de aanpak van de uitbraak. Deze snelle karakterisatie en bijna real-time sequentie analyse met directe invloed op beslissingen voor de publieke gezondheidszorg laat zien wat de kracht is van genoom sequencen voor pathogeen detectie en surveillance.

12.7 Onderzoek naar SARS-CoV-2 clusters in ziekenhuizen

Na de initiele introductie van SARS-CoV-2 in Nederland verspreidde het virus zich, met meerdere kleinere en grotere uitbraken tot gevolg. Om de aanpak van de bestrijding te ondersteunen is het belangrijk om deze uitbraken te onderzoeken en de omvang en origine ervan te bepalen. In **hoofdstuk 8** beschrijven we hoe het combineren van epidemiologische data en genoom sequencen gebruikt werd om inzicht te krijgen in de verspreiding van SARS-CoV-2 bij zorgmedewerkers van drie grote ziekenhuizen in het zuiden van Nederland in maart 2020. Op basis van de analyses kon worden geconcludeerd dat er geen wijdverspreide ziekenhuistransmissie plaatvond tussen medewerkers en patienten en vice versa. De genoom sequenties bleken daarbij van toegevoegde waarde voor het lokale infectiepreventie beleid.

12.8 Afwijkingen vinden in diagnostische en amplicon primers

Tijdens een uitbraak neemt het aantal virussen met mutaties in het genoom toe, door willekeurige fouten die worden gemaakt tijdens virus replicatie. Deze mutaties kunnen plaatsvinden in de gebieden waar de primers van een diagnostische RT-PCR binden wat ervoor kan zorgen dat deze minder goed werken. Om te bepalen of een nieuwe virus variant nog kan worden gedetecteerd is de eerste check vaak of het mutaties bevat in de regio die wordt getarget door de RT-PCR. Vooral als er veel varianten circuleren en er veel verschillende primers gebruikt worden is dit een bewerkelijk proces. In **hoofdstuk 9** beschijven we een gebruiksvriendelijke bioinformatische tool die snel een overzicht geeft welke primers wel of niet overeenkomen met een set virus varianten. Op basis van de positie en de soort mismatch van de primer kan actie worden ondernomen om de primers te testen en aan te passen. Dit geeft wetenschappers de mogelijkheid om snel en makkelijk een eerste check te doen op de diagnostische en/of amplicon sequencing primers die ze gebruiken.

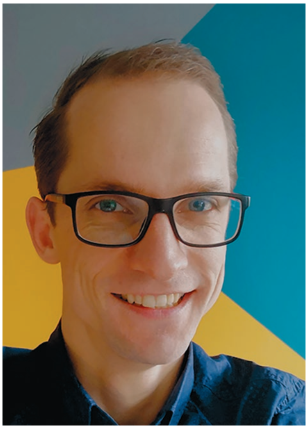
Appendices

About the author

PhD portfolio

List of publications

Acknowledgements



About the author

David Frederik Nieuwenhuijse was born on October 24, 1989, in Woerden, The Netherlands. After completing high school in 2008, he pursued a Bachelor's degree in Biotechnology at Wageningen UR, where he developed a fascination for the inner workings of the cell. In 2012, he completed his Bachelor's degree with a thesis on the role of PIN auxin transporters in root nodule formation in the plant model Medicago truncatula.

David continued his studies at Wageningen, earning a double Master's degree in Biotechnology and Bioinformatics to focus on the computational side of molecular biology. His Biotechnology Master's degree focused on microbial molecular and cell biology, and his thesis at the Cell Biology and Immunology department, completed under the supervision of Prof. Huub Savelkoul, examined the effect of probiotics in the diet of a rapidly aging mouse model on the fitness of their bone marrow-derived immune cells. His Bioinformatics Master's thesis, completed at the Systems and Synthetic Biology department under the supervision of Dr. Peter Schaap, examined the comparison of pathogenic and non-pathogenic Streptococcus species from a protein domain point of view, resulting in a peer-reviewed publication in Plos One.

As part of his final internship, David worked in the Experimental Urology department of Prof. Guido Jenster at the Erasmus Medical Center in Rotterdam. He analyzed sequencing data of long non-coding RNA molecules from the blood of prostate cancer patients and evaluated their prognostic value.

David began his PhD studies in 2016 under supervision of Prof. Marion Koopmans at the viroscience department of the Erasmus Medical Center in Rotterdam. During his PhD he investigated and applied metagenomic and amplicon-based sequencing techniques for the detection and the public health surveillance of pathogenic viruses. As part of his work he developed several analysis tools with a focus on making complex datasets interpretable and at the start of the 2019 SARS-CoV-2 pandemic he developed bioinformatics tools to help scale-up the Dutch sequencing effort.

PhD portfolio

Name	David Frederik Nieuwenhuijse
Research department	Department of Viroscience, Erasmus MC
Research school	Post-graduate Molecular Medicine (MolMed)
PhD period	2016-2022
Promotor	Prof. Dr. Marion P. G. Koopmans
Co-promotor	Dr. Bas B. Oude Munnink

Education

2016 – 2022	**PhD program** Department of Viroscience, ErasmusMC, Rotterdam, The Netherlands
2012-2015	**Master Bioinformatics** Wageningern University and Research Center, Wageningen, The Netherlads
2012-2015	**Master Biotechnology** Wageningern University and Research Center, Wageningen, The Netherlads
2008-2012	**Bachelor Biotechnology** Wageningern University and Research Center, Wageningen, The Netherlads

PhD training

Courses

2016	Course in Virology	1.4 ECTs
2017	Programming with Python	1.0 ECTs
2017	Virus Pathogen Resource Workshop	0.3 ECTs
2018	Bayesians statistics and JASP	0.6 ETCs
2021	Scientific Integrity	0.3 ETCs
2022	Personal Leadership and Communication	1.0 ETCs

Seminars and Workshops

2016	Metagenomics workshop	1.0 ECTs
2018	Computational Pangenomics workshop	1.2 ECTs
2018	Virus Evolution and Molecular Epidemiology workshop	1.5 ECTs
2018	Workshop Snakemake	0.3 ECTs

Oral & poster presentations (10 ETCs)

2016	6th International Calicivirus Conference, United States Presentation: **Fecal-oral route pathogen surveillance using shotgun metagenomics**
2017	Applied Bioinformatics and Public Health Microbiology conference, United Kingdom Poster: **Facilitating virome sequencing data analysis**
2017	COMPARE General Meeting, The Netherlands Poster: **Facilitating high-throughput virome sequencing data analysis**
2018	COMPARE General Meeting, Denmark Poster: **Viral sequence classification using deep learning algorithms**
2018	Utrecht Bioinformatics Center Symposium, The Netherlands Poster/Presentation: **ViRNN: Viral short-read sequence classification using deep learning**
2018	European Conference on Computational Biology, The Netherlands Poster: **ViRNN: Viral short-read sequence classification using deep learning**
2019	COMPARE General Meeting, The Netherlands Poster/Presentation: **Viral diversity in global urban sewage**
2019	Molecular Medicine day, The Netherlands Poster: **Viral diversity in global urban sewage**
2019	European Congress of Virology, The Netherlands Presentation: **Viral diversity in global urban sewage**
2022	VEO General Meeting, The Netherlands Poster: **NAW: A flexible analysis workflow for real-time amplicon sequencing**

Teaching (13.2 ECTs)

2016-2022	**Biannual Basic programming in R course**, full week teaching
2016/2017	**I&I Summer Csourse**, lecture on data clustering
08-2018 – 10-2018	**MSc student**, daily supervision
09-2020 – 06-2021	**HBO student**, daily supervision
09-2021 – 01-2022	**HBO student**, daily supervision
06-2022 – 10-2022	**HBO student**, daily supervision
2021/2022	**Biomedical Research in Practice**, lecture on bioinformatics
2022	**SHARP course on Nanopore Sequencing**, lecture on nanopore sequencing
2022	**SeqNeth Bioinformatics course**, lecture on nanopore sequencing

List of publications

DF Nieuwenhuijse, A van der Linden, RHG Kohl, RS Sikkema, MPG Koopmans, BB Oude Munnink. (2022)
Towards reliable whole genome sequencing for outbreak preparedness and response. BMC genomics 23 (1), 1-9.

S Popping, R Molenkamp, JD Weigel, PGNJ Mutsaers, JC Voermans, S van Boheemen, MC Shamier, RS Sikkema**, DF Nieuwenhuijse**, BB Oude Munnink, MPG Koopmans, JJA van Kampen. (2022)
Diminished amplification of SARS-CoV-2 ORF1ab in a commercial dual-target qRT-PCR diagnostic assay. Journal of virological methods 300, 114397, 1.

L Lu, RS Sikkema, FC Velkers, **DF Nieuwenhuijse**, EAJ Fischer, PA Meijer, NBouwmeester-Vincken, A Rietveld, MCA Wegdam-Blans, P Tolsma, M Koppelman, LAM Smit, RW Hakze-van der Honing, WHM van der Poel, AN van der Spek, MAH Spierenburg, RJ Molenaar, J de Rond, M Augustijn, M Woolhouse, A Stegeman, S Lycett, BB Oude Munnink, MPG Koopmans. (2021)
Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands. Nature communications 12 (1), 1-12, 32.

BB Oude Munnink, N Worp, **DF Nieuwenhuijse**, RS Sikkema, B Haagmans, RAM Fouchier, MPG Koopmans. (2021)
The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. Nature medicine 27 (9), 1518-1524, 106.

LH Boogaard, RS Sikkema, JHGM van Beek, HJ Brockhoff, E Dalebout, B de Heus, SL Niemansburg, **DF Nieuwenhuijse**, D Stougje, E Verspui, BB Oude Munnink, MPG Koopmans, EB Fanoy. (2021)
A mixed-methods approach to elucidate SARS-CoV-2 transmission routes and clustering in outbreaks in native workers and labour migrants in the fruit and vegetable packaging industry in South Holland, the Netherlands, May to July 2020. International Journal of Infectious Diseases 109, 24-32, 4.

R Izquierdo-Lara, G Elsinga, L Heijnen, BB Oude Munnink, CME Schapendonk, **DF Nieuwenhuijse**, M Kon, L Lu, FM Aarestrup, S Lycett, G Medema, MPG Koopmans, M De Graaf. (2021) **Monitoring SARS-CoV-2 circulation and diversity through community wastewater sequencing, the Netherlands and Belgium.** Emerging infectious diseases 27 (5), 1405, 116.

DF Nieuwenhuijse, BB Oude Munnink, MPG Koopmans. (2021) **viromeBrowser: A Shiny App for Browsing Virome Sequencing Analysis Results.** Viruses 13 (3), 437.

BB Oude Munnink, RS Sikkema, **DF Nieuwenhuijse**, RJ Molenaar, E Munger, R Molenkamp, A Van Der Spek, P Tolsma, A Rietveld, M Brouwer, N Bouwmeester-Vincken, F Harders, R Hakze-van Der Honing, MCA Wegdam-Blans, RJ Bouwstra, C GeurtsvanKessel, AA Van Der Eijk, FC Velkers, LAM Smit, A Stegeman, WHM Van Der Poel, MPG Koopmans. (2021) **Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans.** Science 371 (6525), 172-177, 755.

JJE Rovers, LS Van de Linde, N Kenters, EM Bisseling, **DF Nieuwenhuijse**, BB Oude Munnink, A Voss, M Nabuurs-Franssen. (2020) **Why psychiatry is different-challenges and difficulties in managing a nosocomial outbreak of coronavirus disease (COVID-19) in hospital care.** Antimicrobial Resistance & Infection Control 9 (1), 1-8, 15.

RS Sikkema, SD Pas, **DF Nieuwenhuijse**, Á O'Toole, J Verweij, A van der Linden, I Chestakova, CM Schapendonk, M Pronk, P Lexmond, T Bestebroer, RJ Overmars, S van Nieuwkoop, W Van den Bijllaardt, RG Bentvelsen, MML van Rijen, AGM Buiting, AJG van Oudheusden, BM Diederen, AMC Bergmans, AA van der Eijk, R Molenkamp, A Rambaut, A Timen, JAJW Kluytmans, BB Oude Munnink, MFQ Kluytmans van den Bergh, MPG Koopmans. (2020) **COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study.** The Lancet Infectious Diseases 20 (11), 1273-1280, 234.

BB Oude Munnink, **DF Nieuwenhuijse**, M Stein, Á O'Toole, M Haverkate, M Mollers, SK Kamga, CM Schapendonk, M Pronk, P Lexmond, A van der Linden, T Bestebroer, I Chestakova, RJ Overmars, S van Nieuwkoop, R Molenkamp, AA van der Eijk, C GeurtsvanKessel, H Vennema, A Meijer, A Rambaut, J van Dissel, RS Sikkema, A Timen, MPG Koopmans. (2020)
Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nature medicine 26 (9), 1405-1410, 263.

DF Nieuwenhuijse, BB Oude Munnink, MVT Phan, the Global Sewage Surveillance project consortium, P Munk, S Venkatakrishnan, FM Aarestrup, M Cotton, MPG Koopmans. (2020)
Setting a baseline for global urban virome surveillance in sewage. Scientific Reports 10 (1), 1-13, 28.

BB Oude Munnink, **DF Nieuwenhuijse**, RS Sikkema, MPG Koopmans. (2020)
Validating whole genome nanopore sequencing, using Usutu virus as an example. JoVE, e60906, 12.

BB Oude Munnink, E Münger, **DF Nieuwenhuijse**, R Kohl, A van der Linden, CME Schapendonk, H Van Der Jeugd, M Kik, JM Rijks, CBEM Reusken, MPG Koopmans. (2020)
Genomic monitoring to understand the emergence and spread of Usutu virus in the Netherlands, 2016–2018. Scientific reports 10 (1), 1-10, 27.

KTT Kwok, **DF Nieuwenhuijse**, MVT Phan, MPG Koopmans. (2020)
Virus metagenomics in farm animals: a systematic review. Viruses 12 (1), 107, 41.

S Strubbia, J Schaeffer, BB Oude Munnink, A Besnard, MVT Phan, **DF Nieuwenhuijse**, M De Graaf, CME Schapendonk, C Wacrenier, M Cotten, MPG Koopmans, FS Le Guyader. (2019)
Metavirome sequencing to evaluate norovirus diversity in sewage and related bioaccumulated oysters
Frontiers in microbiology 10, 2394, 20.

RS Hendriksen, P Munk, P Njage, B Van Bunnik, L McNally, O Lukjancenko, T Röder, **DF Nieuwenhuijse**, S Karlsmose Pedersen, J Kjeldgaard, RS Kaas, PTLC Clausen, JK Vogt, P Leekitcharoenphon, MGM Van De Schans, T Zuidema, AM de Roda Husman, S Rasmussen, B Petersen, C Amid, G Cochrane, T Sicheritz-Ponten, H Schmitt, JRM Alvarez, A Aidara-Kane, S J Pamp, O Lund, T Hald, M Woolhouse, MPG Koopmans, H Vigre, TN Petersen, FM Aarestrup. (2019)
Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nature communications 10 (1), 1-12, 417.

C Amid, N Pakseresht, N Silvester, S Jayathilaka, O Lund, L D Dynovski, B Á Pataki, D Visontai, BB Xavier, BTF Alako, A Belka, JLB Cisneros, M Cotten, GB Haringhuizen, PW Harrison, D Höper, S Holt, C Hundahl, A Hussein, RS Kaas, X Liu, R Leinonen, S Malhotra-Kumar, **DF Nieuwenhuijse**, N Rahman, C dos S Ribeiro, JE Skiby, D Schmitz, J Stéger, JM Szalai-Gindl, MCF Thomsen, SM Cacciò, I Csabai, A Kroneman, MPG Koopmans, FM Aarestrup, G Cochrane. (2019)
The COMPARE data hubs. Database, 2019, 23.

BM Laksono, RD de Vries, RJ Verburgh, EG Visser, A de Jong, PLA Fraaij, WLM Ruijs, **DF Nieuwenhuijse**, HJ van den Ham, MPG Koopmans, MC van Zelm, ADME Osterhaus, RL de Swart. (2018)
Studies into the mechanism of measles-associated immune suppression during a measles outbreak in the Netherlands. Nature communications 9 (1), 1-10, 78.

RD de Vries, AF Altenburg, NJ Nieuwkoop, E De Bruin, SE van Trierum, MR Pronk, MM Lamers, M Richard, **DF Nieuwenhuijse**, MPG Koopmans, JHCM Kreijtz, RAM Fouchier, ADME Osterhaus, G Sutter, GF Rimmelzwaan. (2018)
Induction of cross-clade antibody and T-cell responses by a modified vaccinia virus Ankara–Based influenza A (H5N1) vaccine in a randomized phase 1/2a clinical trial
The Journal of infectious diseases 218 (4), 614-623, 20.

DF Nieuwenhuijse, MPG Koopmans. (2017)
Metagenomic sequencing for surveillance of food-and waterborne viral diseases. Frontiers in Microbiology 8, 230.

E Saccenti, **DF Nieuwenhuijse**, JJ Koehorst, VAP Martins dos Santos, PJ Schaap (2015)
Assessing the Metabolic Diversity of Streptococcus from a Protein Domain Point of View.
PloS one 10 (9), e0137908.

Acknowledgements

First and foremost, I would like to thank **Marion Koopmans**, my promotor, for the opportunity to explore my interests in what has now resulted in a wrapped-up thesis. Thank you for your patience with me since it is not one of my strengths to quickly finish my research projects. Also, thank you for letting me “swim” and letting me follow my interests while supporting me in my work. I haven’t felt pressured in my work, about which I have heard many other PhD candidates complain, for which I’m thankful, even if that may have also contributed to my delays. Thank you for allowing me to stay at the department to further pursue my interests and further develop the bioinformatics “glue” of the Viroscience department. It is still surprising to me how often you find the time in your extremely busy schedule to talk about my research projects and how quickly you can zoom in and comment on the details of a project. I’m looking forward to continuing to work in the field of virus bioinformatics as there is still so much to develop and to learn.

Second, I would like to thank **Bas Oude Munnink**, my co-promotor. You have been there almost from the beginning and I’m happy that you were there when things in the group or in science “politics” became heated. You always look (and are) undisturbed by anything and that quality was nice to be able to rely on sometimes. It was also great to be part of the COVID sequencing team with you during the heat of the pandemic and work day and night to get the phylogenetic trees back to the collaborators in as real-time as possible.

On that topic I would also like to thank **Reina**, our other team member during the pandemic. Before then we did not work together that much, but it was nice to work in a team with you and to get the right metadata in the right excel sheets.

Miranda, **Claudia**, **Henk-Jan**, **Dennis**, **Annelies** thank you for letting me be part of the early bioinformatics crew in the start of my PhD, you helped me a lot to keep my ideas biologically interpretable and bioinformatically sound.

Also, a great thank you to everyone in the (ever growing) **public health virology group**, you adopted me during the last part of my PhD for which I’m grateful and thank you for all the sequencing work you do in the lab, because without your efforts there is no data to analyze.

Thank you, **Nele** and **Jasmin**, my PhD lunch buddies, for the interesting lunch discussions on the Dutch language, and other weird stuff. Although I spent most of my time behind my computer and you guys spent most of your time in the lab, I really valued the time we spent together outside of work. I hope that I inspired you both to keep on bouldering and although you are living in Germany and are travelling the world, I hope to still meet you at least once a year during Oktoberfest in München!

Emma, **Louella**, **Nathalie**, my post-Corona office desk buddies, you guys made it difficult for me to work from home. **Emma** thank you for our discussions about science and life in general. **Louella**, thank you for letting me beat you at limbo dancing at (almost) every Viroscience borrel. **Nathalie**, thank you for joining the bioinformatics crew, you are the new sewage metagenomics expert now, let's publish some great manuscripts in the future!

Alwin, **Nele**, **Jasmin**, **Do**, **Suzan**, **Lauren**, **Kirsty**, **Janko**, **Emma**, **Louella**, **Nnomzie**, **Babette**, **Div**, **Felicity**, **Shreo**, **Stef** it was a pleasure to (attempt to) infect you with the bouldering virus. Also because of you, bouldering has become a great hobby of mine, so thank you for joining! Let me know if you can go on Tuesday or Wednesday or Thursday, Monday?

Pieter, **Terrens**, **Dennis**, mijn "old boys" netwerk, ik ben als laatste aan de beurt om de PhD binnen te slepen (heel misschien is er nog hoop voor Dennis?). Bedankt boys voor de gesprekken over of naast het PhD leven, ook tijdens onze jaarlijkse last-minute "old boys" weekenden. Laten we die traditie in stand proberen te houden.

Aafke, m'n chickie, verloofde, en ooit binnenkort vrouw, zonder onze studies hadden we elkaar nooit ontmoet en zonder jou was het verder studeren niet gelukt. Dank je voor het luisteren en sparren tijdens de gebruikelijke ups en downs van het PhD traject. Het helpt enorm om net zo'n nerd als ik als partner te hebben die nota bene ook een PhD in de virologie nastreeft. Ik hoop dat ook jij binnenkort je PhD weet af te ronden zodat we als Doctors D 'n A door het leven kunnen.

Last but not least, **pap** en **mam**, bedankt voor het maken van een goed stel hersenen, en voor het altijd voeden van mijn interesse voor de wetenschap al vanaf kleins af aan. Mijn onderzoekende, zelfstandige en eigenwijze karakter heb ik zeker van jullie geërfd en daarvoor ben ik jullie erg dankbaar.

Sequence Appendix

>My_SARS-CoV-2_Genome

nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnncgatctcttgtagatctgttctctaaacga
actttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcagtataattaataactaattactgtcgttgac
aggacacgagtaactcgtctatcttctgcaggctgcttacggtttcgtccgtgttgcagccgatcatcagcacatctaggtttT
gtccgggtgtgaccgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaactcagtttgcct
gttttacaggttcgcgacgtgctcgtacgtggctttggagactccgtggaggaggtcttatcagaggcacgtcaacatcttaa
agatggcacttgtggcttagtagaagttgaaaaaggcgttttgcctcaacttgaacagccctatgtgttcatcaaacgttcgg
atgctcgaactgcacctcatggtcatgttatggttgagctggtagcagaactcgaaggcattcagtacggtcgtagtggtga
gacacttggtgtccttgtccctcatgtgggcgaaataccagtggcttaccgcaaggttcttcttcgtaagaacggtaataaag
gagctggtggccatagGtacggcgccgatctaaagtcatttgacttaggcgacgagcttggcactgatccttatgaagattt
tcaagaaaactggaacactaaacatagcagtggtgttacccgtgaactcatgcgtgagcttaacggaggggcatacactc
gctatgtcgataacaacttctgtggccctgatggctaccctcttgagtgcattaaagaccttctagcacgtgctggtaaagctt
catgcactttgtccgaacaactggactttattgacactaagaggggtgtatactgctgccgtgaacatgagcatgaaattgc
ttggtacacggaacgttctgaaaagagctatgaattgcagacaccttttgaaattaaattggcaaagaaatttgacaccttc
aatggggaatgtccaaattttgtatttcccttaaattccataatcaagactattcaaccaagggttgaaaagaaaaagcttg
atggctttatgggtagaattcgatctgtctatccagttgcgtcaccaaatgaatgcaaccaaatgtgcctttcaactctcatga
agtgtgatcattgtggtgaaacttcatggcagacgggcgattttgttaaagccacttgcgaattttgtggcactgagaatttg
actaaagaaggtgccactacttgtggttacttaccccaaaatgctgttgttaaaatttattgtccagcatgtcacaattcaga
agtaggacctgagcatagtcttgccgaataccataatgaatctggcttgaaaaccattcttcgtaagggtggtcgcactatt
gcctttggaggctgtgtgttctcttatgttggttgccataacaagtgtgcctattgggttccacgtgctagcgctaacataggtt
gtaaccatacaggtgttgttggagaaggttccgaaggtcttaatgacaaccttcttgaaatactTcaaaaagagaaagtca
acatcaatattgttggtgactttaaacttaatgaagagatcgccattattttggcatctttttctgcttccacaagtgcttttgtg
gaaactgtgaaaggtttggattataaagcattcaaacaaattgttgaatcctgtggtaattttaaagttacaaaaggaaaag
ctaaaaaaggtgcctggaatattggtgaacagaaatcaatactgagtcctctttatgcatttgcatcagaggctgctcgtgtt
gtacgatcaattttctcccgcactcttgaaactgctcaaaattctgtgcgtgttttacagaaggccgctataacaatactagat
ggaatttcacagtattcactgagactcattgatgctatgatgttcacatctgatttggctactaacaatctagttgtaatggcc
tacattacaggtggtgttgttcagttgacttcgcagtggctaactaacatctttggcactgtttatgaaaaactcaaacccgtc
cttgattggcttgaagagaagtttaaggaaggtgtagagtttcttagagacggttgggaaattgttaaatttatctcaacctg
tgcttgtgaaattgtcggtggacaaattgtcacctgtgcaaaggaaattaaggagagtgttcagacattctttaagcttgtaa
ataaatttttggctttgtgtgctgactctatcattattggtggagctaaacttaaagccttgaatttaggtgaaacatttgtcac
gcactcaaagggattgtacagaaagtgtgttaaatccagagaagaaactggcctactcatgcctctaaaagccccaaaag
aaattatcttcttagagggagaaacacttcccacagaagtgttaacagaggaagttgtcttgaaaactggtgatttacaacc
attagaacaacctactagtgaagctgttgaagctccattggttggtacaccagtttgtattaacgggcttatgttgctcgaaa
tcaaagacacagaaaagtactgtgcccttgcacctaatatgatggtaacaaacaataccttcacactcaaaggcggtgcac
caacaaaggttacttttggtgatgacactgtgatagaagtgcaaggttacaagagtgtgaatatcaTttttgaacttgatga
aaggattgataaagtacttaatgagaagtgctctgcctatacagttgaactcggtacagaagtaaatgagttcgcctgtgtt
gtggcagatgctgtcataaaaactttgcaaccagtatctgaattacttacaccactgggcattgatttagatgagtggagtat
ggctacatactacttatttgatgagtctggtgagtttaaattggcttcacatatgtattgttctttTtaccctccagatgaggat

gaagaagaaggtgattgtgaagaagaagagtttgagccatcaactcaatatgagtatggtactgaagatgattaccaagg
taaacctttggaatttggtgccacttctgctgctcttcaacctgaagaagagcaagaagaagattggttagatgatgatagt
caacaaactgttggtcaacaagacggcagtgaggacaatcagacaactactattcaaacaattgttgaggttcaacctcaa
ttagagatggaacttacaccagttgttcagactattgaagtgaatagttttagtggttatttaaaacttactgacaatgtatac
attaaaaatgcagacattgtggaagaagctaaaaaggtaaaaccaacagtggttgttaatgcagccaatgtttaccttaaa
catggaggaggtgttgcaggagccttaaataaggctactaacaatgccatgcaagttgaatctgatgattacatagctacta
atggaccacttaaagtgggtggtagttgtgttttaagcggacacaatcttgctaaacactgtcttcatgttgtcggcccaaat
gttaacaaaggtgaagacattcaacttcttaagagtgcttatgaaaattttaatcagcacgaagttctacttgcaccattatt
atcagctggtatttttggtgctgaccctatacattctttaagagtttgtgtagatactgttcgcacaaatgtctacttagctgtct
ttgataaaaatctctatgacaaacttgtttcaagctttttggaaatgaagagtgaaaagcaagttgaacaaaagatcgctga
gattcctaaagaggaagttaagccatttataactgaaagtaaaccttcagttgaacagagaaaacaagatgataagaaaa
tcaaagcttgtgttgaagaagttacaacaactctggaagaaactaagttcctcacagaaaacttgttactttatattgacatt
aatggcaatcttcatccagattctgccactcttgttagtgacattgacatcactttcttaaagaaagatgctccatatatagtg
ggtgatgttgttcaagagggtgttttaactgctgtggttatacctactaaaaaggctAgtggcactactgaaatgctagcga
aagctttgagaaaagtgccaacagacaattatataaccacttacccgggtcagggtttaaatggttacactgtagaggagg
caaagacagtgcttaaaaagtgtaaaagtgcTttttacattctaccatctattatctctaatgagaagcaagaaattcttgg
aactgtttcttggaatttgcgagaaatgcttgcacatgcagaagaaacacgcaaattaatgcctgtctgtgtggaaactaaa
gccatagtttcaactatacagcgtaaatataagggtattaaaatacaagagggtgtggttgattatggtgctagattttactt
ttacaccagtaaaacaactgtagcgtcacttatcaacacacttaacgatctaaatgaaactcttgttacaatgccacttggct
atgtaacacatggcttaaatttggaagaagctgctcggtatatgagatctctcaaagtgccagctacagtttctgtttcttcac
ctgatgctgttacagcgtataatggttatcttacttcttcttctaaaacacctgaagaacattttattgaaaccatctcacttgc
tggttcctataaagattggtcctattctggacaatctacacaactaggtatagaatttcttaagagaggtgataaaagtgtat
attacactagtaatcctaccacattccacctagatggtgaagttatcacctttgacaatcttaagacacttctttctttgagag
aagtgaggactattaaggtgtttacaacagtagacaacattaacctccacacgcaagttgtggacatgtcaatgacatatg
gacaacagtttggtccaacttatttggatggagctgatgttactaaaataaaacctcataattcacatgaaggtaaaacatt
ttatgttttacctaatgatgacactctacgtgttgaggcttttgagtactaccacacaactgatcctagttttctgggtaggtac
atgtcagcattaaatcacactaaaaagtggaaatacccacaagttaatggtttaacttctattaaatgggcagataacaact
gttatcttgccactgcattgttaacactccaacaaatagagttgaagtttaatccacctgctctacaagatgcttattacaga
gcaagggctggtgaagctgctaacttttgtgcacttatcttagcctactgtaataagacagtaggtgagttaggtgatgttag
agaaacaatgagttacttgtttcaacatgccaatttagattcttgcaaaagagtcttgaacgtggtgtgtaaaacttgtggac
aacagcagacaacccttaagggtgtagaagctgttatgtacatgggcacactttcttatgaacaatttaagaaaggtgttca
gataccttgtacgtgtggtaaacaagctacaaaatatctagtacaacaggagtcaccttttgttatgatgtcagcaccacctg
ctcagtatgaacttaagcatggtacatttacttgtgctagtgagtacactggtaattaccagtgtggtcactataaacatata
acttctaaagaaactttgtattgcatagacggtgctttacttacaaagtcctcagaatacaaaggtcctattacggatgttttc
tacaaagaaaacagttacacaacaaccataaaaccagttacttataaattggatggtgttgtttgtacagaaattgacccta
agttggacaattattataagaaagacaattcttatttcacagagcaaccaattgatcttgtaccaaaccaaccatatccaaa
cgcaagcttcgataattttaagtttgtatgtgataatatcaaatttgctgatgatttaaaccagttaactggttataagaaacc
tgcttcaagagagcttaaagttacattttccctgacttaaatggtgatgtggtggctattgattataaacactacacaccctc
ttttaagaaaggagctaaattgttacataaacctattgtttggcatgttaacaatgcaactaataaagccacgtataaacca
aatacctggtgtatacgttgtctttggagcacaaaaccagttgaaacatcaaattcgtttgatgtactgaagtcagaggacg

cgcagggaatggataatcttgcctgcgaagatctaaaaccagtctctgaagaagtagtggaaaatcctaccatacagaaa
gacgttcttgagtgtaatgtgaaaactaccgaagttgtaggagacattatacttaaaccagcaaataatagtttaaaaatta
cagaagaggttggccacacagatctaatggctgcttatgtagacaattctagtcttactattaagaaacctaatgaattatct
agagtattaggtttgaaaacccttgctactcatggtttagctgctgttaatagtgtcccttgggatactatagctaattatgct
aagccttttcttaacaaagttgttagtacaactactaacatagttacacggtgtttaaaccgtgtttgtactaattatatgcctt
atttctttactttattgctacaattgtgtacttttactagaagtacaaattctagaattaaagcatctatgccgactactatagc
aaagaatactgttaagagtgtcggtaaattttgtctagaggcttcatttaattatttgaagtcacctaatttttctaaactgata
aatattataatttggtttttactattaagtgtGtgcctaggttctttaatctactcaaccgctgctttaggtgttttaatgtctaat
ttaggcatgccttcttactgtactggttacagagaaggctatttgaactctactaatgtcactattgcaacctactgtactggtt
ctataccttgtagtgtttgtcttagtggtttagattctttagacacctatccttctttagaaactatacaaattaccatttcatctt
ttaaatgggatttaactgcttttggcttagttgcagagtggtttttggcatatattcttttcactaggtttttctatgtacttggat
tggctgcaatcatgcaattgtttttcagctattttgcagtacattttattagtaattcttggcttatgtggttaataattaatcttg
tacaaatggccccgatttcagctatggttagaatgtacatcttctttgcatcattttattatgtatggaaaagttatgtgcatgt
tgtagacggttgtaattcatcaacttgtatgatgtgttacaaacgtaatagagcaacaagagtcgaatgtacaactattgtta
atggtgttagaaggtccttttatgtctatgctaatggaggtaaaggcttttgcaaactacacaattggaattgtgttaattgtg
atacattctgtgctggtagtacatttattagtgatgaagttgcgagagacttgtcactacagtttaaaagaccaataaatcct
actgaccagtcttcttacatcgttgatagtgttacagtgaagaatggttccatccatctttactttgataaagctggtcaaaag
acttatgaaagacattctctctctcatttgttaacttagacaacctgagagctaataacactaaaggttcattgcctattaat
gttatagtttttgatggtaaatcaaaatgtgaagaatcatctgcaaaatcagcgtctgtttactacagtcagcttatgtgtcaa
cctatactgttactagatcaggcattagtgtctgatgttggtgatagtgcggaagttgcagttaaaatgtttgatgcttacgtt
aatacgttttcatcaacttttaacgtaccaatggaaaaactcaaaacactagttgcaactgcagaagctgaacttgcaaag
aatgtgtccttagacaatgtcttatctacttttatttcagcagctcggcaagggtttgttgattcagatgtagaaactaaagat
gttgttgaatgtcttaaattgtcacatcaatctgacatagaagttactggcgatagttgtaataactatatgctcacctataac
aaagttgaaaacatgacaccccgtgaccttggtgcttgtattgactgtagtgcgcgtcatattaatgcgcaggtagcaaaaa
gtcacaacattgctttgatatggaacgttaaagatttcatgtcattgtctgaacaactacgaaaacaaatacgtagtgctgct
aaaaagaataacttaccttttaagttgacatgtgcaactactagacaagttgttaatgttgtaacaacaaagatagcactta
agggtggtaaaattgttaataattggttgaagcagttaattaaagttacacttgtgttcctttttgttgctgctattttctattta
ataacacctgttcatgtcatgtctaaacatactgacttttcaagtgaaatcataggatacaaggctattgatggtggtgtcac
tcgtgacatagcatctacagatacttgttttgctaacaaacatgctgattttgacacatggtttagccagcgtggtggtagtta
tactaatgacaaagcttgcccattgattgctgcagtcataacaagagaagtgggttttgtcgtgcctggtttgcctggcacga
tattacgcacaactaatggtgactttttgcatttcttacctagagtttttagtgcagttggtaacatctgttacacaccatcaaa
acttatagagtacactgactttgcaacatcagcttgtgttttggctgctgaatgtacaatttttaaagatgcttctggtaagcc
agtaccatattgttatgataccaatgtactagaaggttctgttgcttatgaaagtttacgccctgacacacgttatgtgctcat
ggatggctctattattcaatttcctaacacctaccttgaaggttctgttagagtggtaacaacttttgattctgagtactgtagg
cacggcacttgtgaaagatcagaagctggtgtttgtgtatctactagtggtagatgggtacttaacaatgattattacagatc
tttaccaggagttttctgtggtgtagatgctgtaaatttaTttactaatatgtttacaccactaattcaacctattggtgctttgg
acatatcagcatctatagtagctggtggtattgtGgctatcgtagtaacatgccttgcctactattttatgaggtttagaaga
gcttttggtgaatacagtcatgtagttgcctttaatactttactattccttatgtcattcaTtgtactctgtttaacaccagtttac
tcattcttacctggtgtttattctgttatttacttgtacttgacattttatcttactaatgatgtttctttttagcacatattcagtg
gatggttatgttcacacctttagtacctttctggataacaattgcttatatcatttgtatttccacaaagcatttctattggttctt

tagtaattacctaaagagacgtgtagtctttaatggtgtttcctttagtacttttgaagaagctgcgctgtgcaccttttttgtta
aataaagaaatgtatctaaagttgcgtagtgatgtgctattacctcttacgcaatataatagatacttagctctttataataa
gtacaagtattttagtggagcaatggatacaactagctacagagaagctgcttgttgtcatctTgcaaaggctctcaatgac
ttcagtaactcaggttctgatgttctttaccaaccaccacaaaTctctatcacctcagctgttttgcagagtggttttagaaaa
atggcattcccatctggtaaagttgagggttgtatggtacaagtaacttgtggtacaactacacttaacggtctttggcttgat
gacgtagtttactgtccaagacatgtgatctgcacctctgaagaTatgcttaaccctaattatgaagatttactcattcgtaa
gtctaatcataatttcttggtacaggctggtaatgttcaactcagggttattggacattctatgcaaaattgtgtacttaagctt
aaggttgatacagccaatcctaagacacctaagtataagtttgttcgcattcaaccaggacagactttttcagtgttagcttg
ttacaatggttcaccatctggtgtttaccaatgtgctatgagAcAcaatttcactattaagggttcattccttaatggttcatgt
ggtagtgttggttttaacatagattatgactgtgtctctttttgttacatgcaccatatggaattaccaactggagttcatgctg
gcacagacttagaaggtaacttttatggaccttttgttgacaggcaaacagcacaagcagctggtacggacacaactatta
cagttaatgttttagcttggttgtacgctgctgttataaatggagacaggtggtttctcaatcgatttaccacaactcttaatg
actttaaccttgtggctatgaagtacaattatgaacctctaacacaagaccatgttgacatactaggacctctttctgctcaa
actggaattgccgttttagatatgtgtgcttcattaaaagaattactgcaaaatggtatgaatggacgtaccatattgggtag
tgctttattagaagatgaatttacaccttttgatgttgttagacaatgctcaggtgttactttccaaagtgcagtgaaaagaac
aatcaagggtacacaccactggttgttactcacaattttgacttcacttttagttttagtccagagtactcaatggtctttgttc
ttttttttgtatgaaaatgcctttttaccttttgctatgggtattattgctatgtctgcttttgcaatgatgtttgtcaaacataagc
atgcatttctctgtttgtttttgttaccttctcttgccactgtagcttattttaatatggtctatatgcctgctagttgggtgatgcg
tattatgacatggttggatatggttgatactagtttg---------aagctaaaagactgtgttatgtatgcatcagctgtagtgtt
actaatccttatgacagcaagaactgtgtatgatgatggtgctaggagagtgtggacacttatgaatgtcttgacactcgttt
ataaagtttattatggtaatgctttagatcaagccatttccatgtgggctcttataatctctgttacttctaactactcaggtgt
agttacaactgtcatgtttttggccagaggtattgtttttatgtgtgttgagtattgccctattttcttcataactggtaatacac
ttcagtgtataatgctagtttattgtttcttaggctatttttgtacttgttactttggcctcttttgtttactcaaccgctactttag
actgactcttggtgtttatgattacttagtttctacacaggagtttagatatatgaattcacagggactactcccacccaagaa
tagcatagatgccttcaaactcaacattaaattgttgggtgttggtggcaaaccttgtatcaaagtagccactgtacagtcta
aaatgtcagatgtaaagtgcacatcagtagtcttactctcagttttgcaacaactcagagtagaatcatcatctaaattgtgg
gctcaatgtgtccagttacacaatgacattctcttagctaaagatactactgaagcctttgaaaaaatggtttcactactttct
gttttgctttccatgcagggtgctgtagacataaacaagctttgtgaagaaatgctggacaacagggcaaccttacaagcta
tagcctcagagtttagttcccttccatcatatgcagcttttgctactgctcaagaagcttatgaAcaggctgttgctaatggtg
attctgaagttgttcttaaaaagttgaagaagtctttgaatgtggctaaatctgaatttgaccgtgatgcagccatgcaacgt
aagttggaaaagatggctgatcaagctatgacccaaatgtataaacaggctagatctgaggacaagagggcaaaagtta
ctagtgctatgcagacaatgcttttcactatgcttagaaagttggataatgatgcactcaacaacattatcaacaatgcaag
agatggttgtgttcccttgaacataatacctcttacaacagcagccaaactaatggttgtcataccagactataacacatata
aaaatacgtgtgatggtacaacatttacttatgcatcagcattgtgggaaatccaacaggttgtagatgcagatagtaaaat
tgttcaacttagtgaaattagtatggacaattcacctaatttagcatggcctcttattgtaacagctttaagggccaattctgc
tgtcaaattacagaataatgagcttagtcctgttgcactacgacagatgtcttgtgctgccggtactacacaaactgcttgca
ctgatgacaatgcgttagcttactacaacacaacaaagggaggtaggtttgtacttgcactgttatccgatttacaggatttg
aaatgggctagattccctaagagtgatggaactggtactatTtatacagaactggaaccaccttgtaggtttgttacagaca
cacctaaaggtcctaaagtgaagtatttatactttattaaaggattaaacaacctaaatagaggtatggtacttggtagttta
gctgccacagtacgtctacaagctggtaatgcaacagaagtgcctgccaattcaactgtattatctttctgtgcttttgctgta

gatgctgctaaagcttacaaagattatctagctagtgggggacaaccaatcactaattgtgttaagatgttgtgtacacaca
ctggtactggtcaggcaataacagttacaccggaagccaatatggatcaagaatcctttggtggtgcatcgtgttgtctgtac
tgccgttgccacatagatcatccaaatcctaaaggattttgtgacttaaaaggtaagtatgtacaaatacctacaacttgtgc
taatgaccctgtgggttttacacttaaaaacacagtctgtaccgtctgcggtatgtggaaaggttatggctgtagttgtgatc
aactccgcgaacccatgcttcagtcagctgatgcacaatcgtttttaaacgggtttgcggtgtaagtgcagcccgtcttacac
cgtgcggcacaggcactagtactgatgtcgtatacagggcttttgacatctacaatgataaagtagctggttttgctaaattc
ctaaaaactaattgttgtcgcttccaagaaaaggacgaagatgacaatttaattgattcttactttgtagttaagagacaca
ctttctctaactaccaacatgaagaaacaatttataatttacttaaggattgtccagctgttgctaaacatgacttctttaagtt
tagaatagacggtgacatggtaccacatatatcacgtcaacgtcttactaaatacacaatggcagacctcgtctatgcttta
aggcattttgatgaaggtaattgtgacacattaaaagaaatacttgtcacatacaattgttgtgatgatgattatttcaataa
aaaggactggtatgattCtgtagaaaacccagatatattacgcgtatacgccaacttaggtgaacgtgtacgccaagctttg
ttaaaaacagtacaattctgtgatgccatgcgaaatgctggtattgttggtgtactgacattagataatcaagatctcaatgg
taactggtatgatttcggtgatttcatacaaaccacgccaggtagtggagttcctgttgtagattcttattattcattgttaatg
cctatattaaccttgaccagggctttaactgcagagtcacatgttgacactgacttaacaaagccttacattaagtgggattt
gttaaaatatgacttcacggaagagaggttaaaactctttgaccgttattttaaatattgggatcagacataccacccaaatt
gtgttaactgtttggatgacagatgcattctgcattgtgcaaactttaatgttttattctctacagtgttcccacTtacaagtttt
ggaccactagtgagaaaaatatttgttgatggtgttccatttgtagtttcaactggataccacttcagagagctaggtgttgt
acataatcaggatgtaaacttacatagctctagacttagttttaaggaattacttgtgtatgctgctgaccctgctatgcacgc
tgcttctggtaatctattactagataaacgcactacgtgcttttcagtagctgcacttactaacaatgttgcttttcaaactgtc
aaacccggtaattttaacaaagacttctatgactttgctgtgtctaagggtttctttaaggaaggaagttctgttgaattaaaa
cacttcttctttgctcaggatggtaatgctgctatcagcgattatgactactatcgttataatctaccaacaatgtgtgatatca
gacaactactatttgtagttgaagttgttgataagtactttgattgttacgatggtggctgtattaatgctaaccaagtcatcg
tcaacaacctagacaaatcagctggttttccatttaataaatggggtaaggctagactttattatgattcaatgagttatgag
gatcaagatgcacttttcgcatatacaaaacgtaatgtcatccctactataactcaaatgaatcttaagtatgccattagtgc
aaagaatagagctcgcaccgtagctggtgtctctatctgtagtactatgaccaatagacagtttcatcaaaaattattgaaa
tcaatagccgccactagaggagctactgtagtaattggaacaagcaaattctatggtggttggcacaacatgttaaaaact
gtttatagtgatgtagaaaaccctcaccttatgggttgggattatcctaaatgtgatagagccatgcctaacatgcttagaat
tatggcctcacttgttcttgctcgcaaacatacaacgtgttgtagcttgtcacaccgtttctatagattagctaatgagtgtgct
caagtattgagtgaaatggtcatgtgtggcggttcactatatgttaaaccaggtggaacctcatcaggagatgccacaactg
cttatgctaatagtgtttttaacatttgtcaagctgtcacggccaatgttaatgcacttttatctactgatggtaacaaaattgc
cgataagtatgtccgcaatttacaacacagactttatgagtgtctctatagaaatagagatgttgacacagactttgtgaatg
agttttacgcatatttgcgtaaacatttctcaatgatgatactTtctgacgatgctgttgtgtgtttcaatagcacttatgcatc
tcaaggtctagtggctagcataaagaactttaagtcagttctttattatcaaaacaatgtttttatgtctgaagcaaaatgttg
gactgagactgaccttactaaaggacctcatgaattttgctctcaacatacaatgctagttaaacagggtgatgattatgtgt
accttccttacccagatccatcaagaatcctaggggccggctgttttgtagatgatatcgtaaaaacagatggtacacttatg
attgaacggttcgtgtctttagctatagatgcttacccacttactaaacatcctaatcaggagtatgctgatgtctttcatttgt
acttacaatacataagaaagctacatgatgagttaacaggacacatgttagacatgtattctgttatgcttactaatgataa
cacttcaaggtattgggaacctgagttttatgaggctatgtacacaccgcatacagtcttacaggctgttggggcttgtgttct
ttgcaattcacagacttcattaagatgtggtgcttgcatacgtagaccattcttatgttgtaaatgctgttacgaccatgtcat
atcaacatcacataaattagtcttgtctgttaatccgtatgtttgcaatgctccaggttgtgatgtcacagatgtgactcaact

```
ttacttaggaggtatgagctattattgtaaatcacataaaccacccattagttttccattgtgtgctaatggacaagtttttggt
ttatataaaaatacatgtgttggtagcgataatgttactgactttaatgcaattgcaacatgtgactggacaaatgctggtga
ttacattttagctaacacctgtactgaaagactcaagctttttgcagcagaaacgctcaaagctactgaggagacatttaaa
ctgtcttatggtattgctactgtacgtgaagtgctgtctgacagagaattacatctttcatgggaagttggtaaacctagacc
accacttaaccgaaattatgtctttactggttatcgtgtaactaaaaacagtaaagtacaaataggagagtacacctttgaa
aaaggtgactatggtgatgctgttgtttaccgaggtacaacaacttacaaattaaatgttggtgattattttgtgctgacatca
catacagtaatgccattaagtgcacctacactagtgccacaagagcactatgttagaattactggcttatacccaacactca
atatctcagatgagttttctagcaatgttgcaaattatcaaaaggttggtatgcaaaagtattctacactccagggaccacct
ggtactggtaagagtcattttgctattggcctagctctctactacccttctgctcgcatagtgtatacagcttgctctcatgccg
ctgttgatgcactatgtgagaaggcattaaaatatttgcctatagataaatgtagtagaattatacctgcacgtgctcgtgta
gagtgttttgataaattcaaagtgaattcaacattagaacagtatgtcttttgtactgtaaatgcattgcctgagacgacagT
agatatagttgtctttgatgaaatttcaatggccacaaattatgatttgagtgttgtcaatgccagattaTgtgctaagcacta
tgtgtacattggcgaccctgctcaattacctgcaccacgcacattgctaactaagggcacactagaaccagaatatttcaat
tcagtgtgtagacttatgaaaactataggtccagacatgttcctcggaacttgtcggcgttgtcctgctgaaattgttgacact
gtgagtgctttggtttatgataataagcttaaagcacataaagacaaatcagctcaatgctttaaaatgttttataagggtgt
tatcacgcatgatgtttcatctgcaattaacaggccacaaataggcgtggtaagagaattccttacacgtaaccctgcttgg
agaaaagctgtctttatttcaccttataattcacagaatgctgtagcctcaaagattttgggactaccaactcaaactgttgat
tcatcacagggctcagaatatgactatgtcatattcactcaaaccactgaaacagctcactcttgtaatgtaaacagatttaa
tgttgctattaccagagcaaaagtaggcatactttgcataatgtctgatagagacctttatgacaagttgcaatttacaagtc
ttgaaattccacgtaggaatgtggcaactttacaagctgaaaatgtaacaggactctttaaagattgtagtaaggtaatcac
tgggttacatcctacacaggcacctacacacctcagtgttgacactaaattcaaaactgaaggtttatgtgttgacGtacct
ggcatacctaaggacatgacctatagaagactcatctctatgatgggttttaaaatgaattatcaagttaatggttaccctaa
catgtttatcacccgcgaagaagctataagacatgtacgtgcatggattggcttcgatgtcgaggggtgtcatgctactaga
gaagctgttggtaccaatttacctttacagctaggtttttctacaggtgttaacctagttgctgtacctacaggttatgttgata
cacctaataatacagattttccagagttagtgctaaaccaccgcctggagatcaatttaaacacctcataccacttatgtac
aaaggacttccttggaatgtagtgcgtataaagattgtacaaatgttaagtgacacacttaaaaatctctctgacagagtcg
tatttgtcttatgggcacatggctttgagttgacatctatgaagtattttgtgaaaataggacctgagcgcacctgttgtctat
gtgatagacgtgccacatgcttttccactgcttcagacacttatgcctgttggcatcattctattggatttgattacgtctataa
tccgtttatgattgatgttcaacaatggggttttacaggtaacctacaaagcaaccatgatctgtattgtcaagtccatggta
atgcacatgtagctagttgtgatgcaatcatgactaggtgtctagctgtccacgagtgctttgttaagcgtgttgactggact
attgaatatcctataattggtgatgaactgaagattaatgcggcttgtagaaaggttcaacacatggttgttaaagctgcatt
attagcagacaaattcccagttcttcacgacattggtaaccctaaagctattaagtgtgtacctcaagctgatgtagaatgg
aagttctatgatgcacagccttgtagtgacaaagcttataaaatagaagaattattctattcttatgccacacattctgacaa
attcacagatggtgtatgcctattttggaattgcaatgtcgatagatatcctgctaattccattgtttgtagatttgacactaga
gtgctatctaaccttaacttgcctggttgtgatggtggcagtttgtatgtaaataaacatgcattccacacaccagcttttgat
aaaagtgcttttgttaatttaaaacaattaccatttttctattactctgacagtccatgtgagtctcatggaaaacaagtagtg
tcagatatagattatgtaccactaaagtctgctacgtgtataacacgttgcaatttaggtggtgctgtctgtagacatcatgct
aatgagtacagattgtatctcgatgcttataacatgatgatctcagctggctttagcttgtgggtttacaaacaatttgatact
tataacctctggaacacttttacaagacttcagagtttagaaaatgtggcttttaatgttgtaaataagggacactttgatgg
acaacagggtgaagtaccagtttctatcattaataacactgtttacacaaaagttgatggtgttgatgtagaattgtttgaa
```

```
ataaaacaacattacctgttaatgtagcatttgagctttgggctaagcgcaacattaaaccagtaccagaggtgaaaatact
caataatttgggtgtggacattgctgctaatactgtgatctgggactacaaaagagatgctccagcacatatatctactattg
gtgtttgttctatgactgacatagccaagaaaccaaTtgaaacgatttgtgcaccactcactgtctttttgatggtagagttg
atggtcaagtagacttatttagaaatgcccgtaatggtgttcttattacagaGggtagtgttaaaggtttacaaccatctgta
ggtcccaaacaagctagtcttaatggagtcacattaattggagaagccgtaaaaacacagttcaattattataagaaagtt
gatggtgttgtccaacaattacctgaaacttactttactcagagtagaaatttacaagaatttaaacccaggagtcaaatgg
aaattgatttcttagaattagctatggatgaattcattgaacggtataaattagaaggctatgccttcgaacatatcgtttatg
gagattttagtcatagtcagttaggtggtttacatctactgattggactagctaaacgttttaaggaatcaccttttgaattag
aagattttattcctatggacagtacagttaaaaactatttcataacagatgcgcaaacaggttcatctaagtgtgtgtgttct
gttattgatttattacttgatgattttgttgaaataataaaatcccaagatttatctgtagtttctaaggttgtcaaagtgactat
tgactatacagaaatttcatttatgctttggtgtaaagatggccatgtagaaacattttacccaaaattacaatctagtcaag
cgtggcaaccgggtgttgctatgcctaatctttacaaaatgcaaagaatgctattagaaaagtgtgaccttcaaaattatgg
tgatagtgcaacattacctaaaggcataatgatgaatgtcgcaaaatatactcaactgtgtcaatatttaaacacattaaca
ttagctgtaccctataatatgagagttatacattttggtgctggttctgataaaggagttgcaccaggtacagctgttttaaga
cagtggttgcctacgggtacgctgcttgtcgattcagatcttaatgactttgtctctgatgcagattcaactttgattggtgatt
gtgcaactgtacatacagctaataaatgggatctcattattagtgatatgtacgaccctaagactaaaaatgttacaaaaga
aaatgactctaaagagggtttttcacttacatttgtgggtttatacaacaaaagctagctcttggaggttccgtggctataa
agataacagaacattcttggaatgctgatctttataagctcatgggacacttcgcatggtggacagcctttgttactaatgtg
aatgcgtcatcatctgaagcattttaattggatgtaattatcttggcaaaccacgcgaacaaatagatggttatgtcatgca
tgcaaattacatattttggaggaatacaaatccaattcagttgtcttcctattctttatttgacatgagtaaatttccccttaaa
ttaaggggtactgctgttatgtctttaaaagaaggtcaaatcaatgatatgattttatctcttcttagtaaaggtagacttata
attagagaaaacaacagagttgttatttctagtgatgttcttgttaacaactaaacgaacaatgtttgtttttcttgttttattg
ccactagtctctagtcagtgtgttaatcttaTaaccagaactcaat---------catacactaattctttcacacgtggtgtttat
taccctgacaaagttttcagatcctcagttttacattcaactcaggacttgttcttacctttcttttccaatgttacttggttccat
gcta------tctctgggaccaatggtactaagaggtttgataaccctgtcctaccatttaatgatggtgtttattttgcttccact
gagaagtctaacataataagaggctggatttttggtactactttagattcgaagacccagtccctacttattgttaataacgc
tactaatgttgttattaaagtctgtgaatttcaattttgtaatgatccatttttggAtgtttattaccacaaaaacaacaaaagt
tggatggaaagtgagttcagagtttattctagtgcgaataattgcacttttgaatatgtctctcagccttttcttatggaccttg
aaggaaaacagggtaatttcaaaaatcttagggaatttgtgtttaagaatattgatggttattttaaaatatattctaagcac
acgcctattaatttagGgcgtgatctccctcagggtttttcggctttagaaccattggtagatttgccaataggtattaacatc
actaggtttcaaactttacttgctttacatagaagttatttgactcctggtgattcttcttcaggttggacagctggtgctgcag
cttattatgtgggttatcttcaacctaggacttttctattaaaatataatgaaaatggaaccattacagatgctgtagactgtg
cacttgaccctctctcagaaacaaagtgtacgttgaaatccttcactgtagaaaaaggaatctatcaaacttctaactttaga
gtccaaccaacagaatctattgttagatttcctaatattacaaacttgtgcccttttgAtgaagtttttaacgccaccagattt
gcatctgtttatgcttggaacaggaagagaatcagcaactgtgttgctgattattctgtcctatataattccgcaCcattttTc
Gcttttaagtgttatggagtgtctcctactaaattaaatgatctctgctttactaatgtctatgcagattcatttgtaattagag
gtAatgaagtcagCcaaatcgctccagggcaaactggaaaTattgctgattataattataaattaccagatgattttacag
gctgcgttatagcttggaattctaacaaGcttgattctaaggttggtggtaattataattaccGgtatagattgtttaggaag
tctaatctcaaaccttttgagagagatatttcaactgaaatctatcaggccggtaAcaAaccttgtaatggtgttgCaggtG
ttaattgttactttcctttacaatcatatggtttccGacccactTatggtgttggtCaccaaccatacagagtagtagtacttt
```

```
cttttgaacttctacatgcaccagcaactgtttgtggacctaaaaagtctactaatttggttaaaaacaaatgtgtcaatttca
acttcaatggtttaacaggcacaggtgttcttactgagtctaacaaaaagtttctgcctttccaacaatttggcagagacatt
gctgacactactgatgctgtccgtgatccacagacacttgagattcttgacattacaccatgttcttttggtggtgtcagtgtt
ataacaccaggaacaaatacttctaaccaggttgctgttctttatcaggGtgttaactgcacagaagtccctgttgctattca
tgcagatcaacttactcctacttggcgtgtttattctacaggttctaatgtttttcaaacacgtgcaggctgtttaataggggct
gaaTatgtcaacaactcatatgagtgtgacatacccattggtgcaggtatatgcgctagttatcagactcagactaaGtctc
Atcggcgggcacgtagtgtagctagtcaatccatcattgcctacactatgtcacttggtgcagaaaattcagttgcttactct
aataactctattgccatacccacaaattttactattagtgttaccacagaaattctaccagtgtctatgaccaagacatcagt
agattgtacaatgtacatttgtggtgattcaactgaatgcagcaatcttttgttgcaatatggcagtttttgtacacaattaaa
AcgtgctttaactggaatagctgttgaacaagacaaaaacacTcaagaagtttttgcacaagtcaaacaaatttacaaaa
caccaccaattaaaTattttggtggttttaatttttcacaaatattaccagatccatcaaaaccaagcaagaggtcatttatt
gaagatctacttttcaacaaagtgacacttgcagatgctggcttcatcaaacaatatggtgattgccttggtgatattgctgct
agagacctcatttgtgcacaaaagtttaacggccttactgttttgccacctttgctcacagatgaaatgattgctcaatacact
tctgcactgttagcgggtacaatcacttctggttggacctttggtgcaggtgctgcattacaaataccatttgctatgcaaatg
gcttataggtttaatggtattggagttacacagaatgttctctatgagaaccaaaaattgattgccaaccaatttaatagtgc
tattggcaaaattcaagactcactttcttccacagcaagtgcacttggaaaacttcaagatgtggtcaaccaTaatgcacaa
gctttaaacacgcttgttaaacaacttagctccaaAtttggtgcaatttcaagtgttttaaatgatatcctttcacgtcttgaca
aagttgaggctgaagtgcaaattgataggttgatcacaggcagacttcaaagtttgcagacatatgtgactcaacaattaat
tagagctgcagaaatcagagcttctgctaatcttgctgctactaaaatgtcagagtgtgtacttggacaatcaaaaagagtt
gatttttgtggaaagggctatcatcttatgtccttccctcagtcagcacctcatggtgtagtcttcttgcatgtgacttatgtccc
tgcacaagaaaagaacttcacaactgctcctgccatttgtcatgatggaaaagcacactttcctcgtgaaggtgtctttgttt
caaatggcacacactggtttgtaacacaaaggaatttttatgaaccacaaatcattactacagacaacacatttgtgtctgg
taactgtgatgttgtaataggaattgtcaacaacacagtttatgatcctttgcaacctgaattagaTtcattcaaggaggagt
tagataaatattttaagaatcatacatcaccagatgttgatttaggtgacatctctggcattaatgcttcagttgtaaacattc
aaaaagaaattgaccgcctcaatgaggttgccaagaatttaaatgaatctctcatcgatctccaagaacttggaaagtatg
agcagtatataaaatggccatggtacatttggctaggttttatagctggcttgattgccatagtaatggtgacaattatgcttt
gctgtatgaccagttgctgtagttgtctcaagggctgttgttcttgtggatcctgctgcaaatttgatgaagacgactctgagc
cagtgctcaaaggagtcaaattacattacacataaacgaacttatggatttgtttatgagaatcttcacaattggaactgta
actttgaagcaaggtgaaatcaaggatgctactccttcagattttgttcgcgctactgcaacgataccgatacaagcctcact
ccctttcggatggcttattgttggcgttgcacttcttgctgtttttcagagcgcttccaaaatcataacTctcaaaaagagatg
gcaactagcactctccaagggtgttcactttgtttgcaacttgctgttgttgtttgtaacagtttactcacaccttttgctcgttg
ctgctggccttgaagccccttttctctatctttatgctttagtctacttcttgcagagtataaactttgtaagaataataatgag
gctttggctttgctggaaatgccgttccaaaaacccattactttatgatgccaactattttctttgctggcatactaattgttac
gactattgtataccttacaatagtgtaacttcttcaattgtcattacttcaggtgatggcacaacaagtcctatttctgaacat
gactaccagattggtggttatactgaaaaatgggaatctggagtaaaagactgtgttgtattacacagttacttcacttcag
actattaccagctgtactcaactcaattgagtacagacaTtggtgttgaacatgttaccttcttcatctacaataaaattgttg
atgagcctgaagaacatgtccaaattcacacaatcgacggttcatccggagttgttaatccagtaatggaaccaatttatga
tgaaccgacgacgactactagcgtgcctttgtaagcacaagctgatgagtacgaacttatgtactcattcgtttcggaagag
aTaggtacgttaatagttaatagcgtacttctttttcttgctttcgtggtattcttgctagttacactagccatccttactgcgct
tcgattgtgtgcgtactgctgcaatattgttaacgtgagtcttgtaaaaccttctttttacgtttactctcgtgttaaaaatctga
```

attcttctagagttcctgatcttctggtctaaacgaactaaatattatattagtttttctgtttggaactttaattttagccatggc
aAattccaacggtactattaccgttgaagagcttaaaaagctccttgaaGaatggaacctagtaataggtttcctattcctt
acatggatttgtcttctacaatttgcctatgccaacaggaataggtttttgtatataattaagttaattttcctctggctgttatg
gccagtaactttaActtgttttgtgcttgctgctgtttacagaataaattggatcaccggtggaattgctatcgcaatggcttg
tcttgtaggcttgatgtggctcagctacttcattgcttctttcagactgtttgcgcgtacgcgttccatgtggtcattcaatccag
aaactaacattcttctcaacgtgccactccatggcactattctgaccagaccgcttctagaaagtgaactcgtaatcggagct
gtgatccttcgtggacatcttcgtattgctggacaccatctaggacgctgtgacatcaaggacctgcctaaagaaatcactgt
tgctacGtcacgaacgctttcttattacaaattgggagcttcgcagcgtgtagcaggtgactcaggttttgctgcatacagtc
gctacaggattggcaactataaattaaacacagaccattccagtagcagtgacaatattgctttgcttgtacagtaagtgac
aacagatgtttcatctcgttgactttcaggttactatagcagagatattactaattattatgaggacttttaaagtttccatttg
gaatcttgattacatcataaacctcataattaaaaatttatctaagtcactaactgagaataaatattctcaattagatgaag
agcaaccaatggagattgattaaacgaacatgaaaattattcttttcttggcactgataacactcgctacttgtgagctttat
cactaccaagagtgtgttagaggtacaacagtacttttaaaagaaccttgctcttctggaacatacgagggcaattcaccat
ttcatcctctagctgataacaaatttgcactgacttgctttagcactcaatttgcttttgcttgtcctgacggcgtaaaacacgt
ctatcagttacgtgccagatcagtttcacctaaactgttcatcagacaagaggaagttcaagaactttactctccaatttttct
tattgttgcggcaatagtgtttataacactttgcttcacactcaaaagaaagacagaatgattgaactttcattaattgacttc
tatttgtgctttttagcctttctgTtattccttgttttaattatgcttattatcttttggttctcacttgaactgcaagatcataatg
aaacttgtcacgcctaaaTgaacatgaaatttcttgttttcttaggaatcatcacaactgtagctgcatttcaccaagaatgt
agtttacagtcatgtactcaacatcaaccatatgtagttgatgacccgtgtcctattcacttctattctaaatggtatattaga
gtaggagctagaaaatcagcacctttaattgaattgtgcgtggatgaggctggttctaaatcacccattcagtacatcgatat
cggtaattatacagtttcctgtttaccttttacaattaattgccaggaacctaaattgggtagtcttgtagtgcgttgttcgttct
atgaagactttttagagtatcatgacgttcgtgttgttttagatttcatctaaacgaacaaactTaaatgtctgataatggacc
ccaaaatcagcgaaatgcaTTccgcattacgtttggtggGccctcagattcaactggcagtaaccagaatg---------gtg
gggcgcgatcaaaacaacgtcggccccaaggtttacccaataatactgcgtcttggttcaccgctctcactcaacatggcaa
ggaagaccttaaattccctcgaggacaaggcgttccaattaacaccaatagcagtccagatgaccaaattggctactaccg
aagagctaccagacgaattcgtggtggtgacggtaaaatgaaagatctcagtccaagatggtatttctactacctaggaac
tgggccagaagctggacttccctatggtgctaacaaagacggcatcatatgggttgcaactgagggagccttgaatacacc
aaaagatcacattggcacccgcaatcctgctaacaatgctgcaatcgtgctacaacttcctcaaggaacaacattgccaaa
aggcttctacgcagaagggagcagaggcggcagtcaagcctcttctcgttcctcatcacgtagtcgcaacagttcaagaaa
ttcaactccaggcagcagtaAACgaacttctcctgctagaatggctggcaatggcggtgatgctgctcttgctttgctgctg
cttgacagattgaaccagcttgagagcaaaatgtctggtaaaggccaacaacaacaaggccaaactgtcactaagaaatc
tgctgctgaggcttctaagaagcctcggcaaaaacgtactgccactaaagcatacaatgtaacacaagctttcggcagacg
tggtccagaacaaacccaaggaaattttggggaccaggaactaatcagacaaggaactgattacaaacattggccgcaa
attgcacaatttgcccccagcgcttcagcgttcttcggaatgtcgcgcattggcatggaagtcacaccttcgggaacgtggtt
gacctacacaggtgccatcaaattggatgacaaagatccaaatttcaaagatcaagtcattttgctgaataagcatattgac
gcatacaaaacattcccaccaacagagcctaaaaaggacaaaaagaagaaggctgatgaaactcaagccttaccgcaga
gacagaagaaacagcaaactgtgactcttcttcctgctgcagatttggatgatttctccaaacaattgcaacaatccatgag
cCgtgctgactcaactcaggcctaaactcatgcagaccacacaaggcagatgggctatataaacgttttcgcttttccgttta
cgatatatagtctactcttgtgcagaatgaattctcgtaactacatagcacaagtagatgtagttaactttaatctcacatag
caatctttaatcagtgtgtaacattagggaggacttgaaagagccaccacattttcaccgaggccacgcggagtacgatcg

agtgtacagtgaacaatgctagggagagctgcctatatggaagagccctaatgtgtaaaattaattttagtagtgnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
n