frontiers in Microbiology

FOCUSED REVIEW
published: 01 October 2015
doi: 10.3389/fmicb.2015.01069

Recovering full-length viral genomes from metagenomes

Saskia L. Smits[1‡], Rogier Bodewes[1†‡], Aritz Ruiz-González[2,3,4], Wolfgang Baumgärtner[5], Marion P. Koopmans[1,6], Albert D. M. E. Osterhaus[1,7] and Anita C. Schürch[1†]*

[1] Department of Viroscience, Erasmus Medical Center, Rotterdam, Netherlands, [2] Department of Zoology and Animal Cell Biology, University of the Basque Country (UPV/EHU), Vitoria-Gasteiz, Spain, [3] Systematics, Biogeography and Population Dynamics Research Group, Lascaray Research Center, University of the Basque Country (UPV/EHU), Vitoria-Gasteiz, Spain, [4] Conservation Genetics Laboratory, National Institute for Environmental Protection and Research, Bologna, Italy, [5] Department of Pathology, University of Veterinary Medicine Hannover, Hannover, Germany, [6] Centre for Infectious Diseases Research, Diagnostics and Screening, National Institute for Public Health and the Environment, Bilthoven, Netherlands, [7] Center for Infection Medicine and Zoonoses Research, Hannover, Germany

Infectious disease metagenomics is driven by the question: "what is causing the disease?" in contrast to classical metagenome studies which are guided by "what is out there?" In case of a novel virus, a first step to eventually establishing etiology can be to recover a full-length viral genome from a metagenomic sample. However, retrieval of a full-length genome of a divergent virus is technically challenging and can be time-consuming and costly. Here we discuss different assembly and fragment linkage strategies such as iterative assembly, motif searches, k-mer frequency profiling, coverage profile binning, and other strategies used to recover genomes of potential viral pathogens in a timely and cost-effective manner.

Keywords: metagenomics, viruses, virus discovery, assembly, k-mer analysis, coverage analysis, motif discovery, zoonotic pathogens

OPEN ACCESS

Edited by:
Akihide Ryo, Yokohama City University, Japan

Reviewed by:
Efthymios Ladoukakis, National Technical University of Athens, Greece
Makoto Kuroda, National Institute of Infectious Diseases, Japan

Correspondence:

Anita C. Schürch *is a bioinformatician and postdoc at the Department of Medical Microbiology, University Medical Center Utrecht, Netherlands. Her main interest lies in the characterization of etiological agents of infectious diseases.*
a.c.schurch@umcutrecht.nl

†Present Address:
Rogier Bodewes, Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands; Anita C. Schürch, Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, Netherlands

‡These authors have contributed equally to this work.

Received: *19 August 2015*
Accepted: *17 September 2015*
Published: *01 October 2015*

Citation:
Smits SL, Bodewes R, Ruiz-González A, Baumgärtner W, Koopmans MP, Osterhaus ADME and Schürch AC (2015) Recovering full-length viral genomes from metagenomes. Front. Microbiol. 6:1069. doi: 10.3389/fmicb.2015.01069

Introduction

Infectious viral diseases, both emerging, and re-emerging, pose a continuous health threat and disease burden to humans. In recent years, we have seen an increasing number of emerging virus disease outbreaks in human and animals alike with substantial health and economic impact. They are most likely related to accelerating environmental and anthropogenic changes, such as increased mobility and demographic changes, which alter the rate and nature of contact between animal and human populations. The influenza A viruses are probably the most notorious viruses which have shown their potential for repeated cross species transmission and pandemic potential (Claas et al., 1998; Koopmans, 2013; de Graaf and Fouchier, 2014; Bodewes et al., 2015). Schmallenberg virus caused an outbreak in ruminants with major impact on international trade of susceptible animals and animal products such as semen and embryos with more than 15 countries imposing restrictions on imports of live cattle from the European Union (EU) (Beer et al., 2013). Lately, Middle East Respiratory Syndrome (MERS) coronavirus was causing renewed concern as it spread from the Middle East to the Republic of Korea with 186 confirmed human cases, including 36 deaths in July 2015 (http://www.who.int/csr/don/21-july-2015-mers-korea/en/). Ebola virus has fruit bats (*Pteripodidae*) as natural hosts and the current epidemic is thought to be introduced into the human population by zoonotic transmission (Marí Saíz et al., 2014). Many of the most important human pathogens are either zoonotic or originated as zoonoses before adapting to humans (Taylor et al., 2001; Kuiken et al., 2005; Woolhouse and Gowtage-Sequeria, 2005; Cutler, 2010; Morse et al., 2012) and humanity is continuously being exposed to novel animal pathogens.

Breakthroughs in the field of metagenomics have had far-reaching effects on the identification and characterization of newly emerging viral pathogens (Fauci and Morens, 2012). Virus discovery metagenomics assays rely on sequence-independent amplification of nucleic acids from clinical samples, in combination with next-generation sequencing platforms and bioinformatics tools for sequence analysis. They are relatively simple and fast, and allow detection of hundreds of viruses simultaneously and unknown viruses even if they are highly divergent from those that are already described (Rosario and Breitbart, 2011; Miller et al., 2013; Smits and Osterhaus, 2013). If the new viral genome shows considerable similarity to previously characterized virus genomes present in public databases, the identification of a new virus and its genomic characterization can be finalized in a matter of days and a fraction of the costs compared to a few years ago. This is of utmost importance for timely disease outbreak management. Here, the guiding questions are: Is the group of diseased persons normal for the time of year and/or geographic area? If so, which pathogen(s) is causing the disease? Who gets infected? How do people get infected? What is the source of infection? What are transmission routes? How can infection be prevented, treated and/or contained? The fast discovery of a partial or full-length viral genome can also serve as basis for development of specific molecular diagnostic assays to confirm suspect cases and for development of vaccines and antivirals. This was exemplified after the discoveries of Schmallenberg virus and MERS Coronavirus, where molecular diagnostic protocols were made available within a matter of days after the discovery of the pathogen (Beer et al., 2013; Pollack et al., 2013).

Despite this promise, however, most new discoveries made through metagenomics in fact are viruses that belong to already known virus families as current data analysis strategies rely mostly on **similarity searches** against annotated sequences in public databases (Woyke et al., 2006; Chew and Holmes, 2009; Schmieder and Edwards, 2012; Garcia-Garcerà et al., 2013; Prachayangprecha et al., 2014; Schürch et al., 2014). Significant problems in characterization of full-length viral genomes from metagenomic datasets are encountered when dealing with a highly divergent new virus with no closely related genomes in public databases. Additional time-consuming experimental approaches are required to obtain full-length genome sequences (Van Leeuwen et al., 2010; Siegers et al., 2014). By optimally mining the metagenomic content, for example through effective assembly, k-mer frequency profiling, motif search, coverage profile binning, or other fragment linkage strategies, the likelihood, and speed of finding viral reads and the level of viral genome completeness can be increased. This also increases the number of reads for which a source can be assigned in metagenomes. An example was described by us recently (Smits et al., 2014) and showed that, using BLAST searches, 27.67, 5.82, and 0.11% of all reads were identified as being from the viral target genome (**Figure 1**), whereas after viral **genome finishing**, 69.52, 13.58, and 26.14% reads were tagged as belonging to the viral genome (**Figure 1**). At the same time, genome completeness increased from 7291, 7682, and 24,734 bases in the initial fragments, to full-length or nearly full-length genomes of 11, 15.5, and 33 kb (**Figure 2**). With the *in silico* methods described here, the need for laboratory follow-up can be minimized, thereby providing the necessary information in a timely, efficient, and more cost-effective manner.

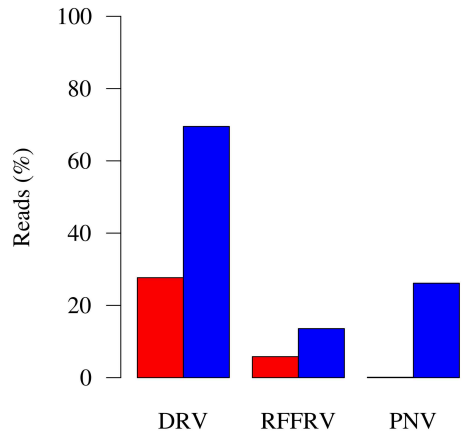

FIGURE 1 | Metagenome composition before and after genome finishing. Red bars: Percentage of reads of the metagenome assigned to target genome after similarity searches and before genome finishing. Blue bars: Percentage of reads of the metagenome assigned to target genome after genome finishing strategies were applied to the target genome. DRV, dolphin rhabdovirus; RFFRV, red fox fecal rhabdovirus; PNV, python nidovirus.

KEY CONCEPT 1 | Similarity searches
Similarity searches in metagenomics, also referred to as homology searches, refers to searching of sequences databases for matching sequences, often with BLAST (Altschul, 1997), in order to assign a taxonomy to a query sequence.

KEY CONCEPT 2 | Genome finishing
After metagenome assembly, genomes of individual organisms are often incomplete and shattered into different contigs. While in case of known viral genomes, the genomes can be resolved by mapping to a reference genome, in novel, and divergent viral genomes it is often necessary to link contigs first. Further steps to retrieve a full-length genome involve confirmation of linkage, augmenting assemblies, determination of the order of the contigs and closing gaps between them.

Recovery of full-length genomes of novel viruses from metagenome data consists of four different steps: First, the assembly of the reads into long fragments, second, assignment of at least one contig (seed) as originating from the target virus, third, the linkage of other fragments to the seed contig to receive a draft genome and fourth, gap closing and finalizing of the draft genome to receive a full-length genome (**Figure 3**).

Assembly

A metagenome dataset consists not only of reads of the (novel viral) target genome but also of all other genetic material present in the sequenced sample, at varying concentrations and sequencing depth. If the viral load in the sample is sufficient to produce overlapping sequencing reads, it is in theory possible to

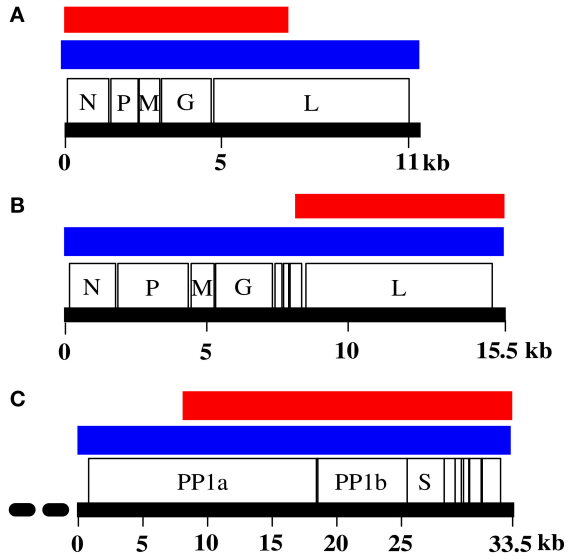

FIGURE 2 | Genome completeness before and after genome finishing. **(A)** Genome of dolphin rhabdovirus (DRV). **(B)** Genome of red fox fecal rhabdovirus (RFFRV). **(C)** Genome of python nidovirus (PNV). Indicated is contig size after initial genome assembly (red) and after genome finishing (blue). Striped line at the 5′ end of PNV indicates putative unresolved 5′ end.

create longer, contiguous sequences from the viral genome. A first step in retrieval of a viral genome from a metagenome is therefore to assemble all reads, followed by identification of (part of) the target genome. Unfortunately, only very few assemblers were developed with virus discovery metagenomics in mind. A specific challenge here is the much greater sequence diversity in viruses compared with host or bacterial genomes. In contrast to assembly in general metagenomics where assembly of all sequencing reads is necessary to shed light into population and functions within the metagenomic community (Howe and Chain, 2015), only assembly of a single or few targets (namely the virus in question) but to a high level of completeness is pursued. These targets might have highly uneven coverage and can be very dissimilar to sequences of known pathogens which make reference-guided assembly impossible.

A comparison of the performance of standard, whole-genome shotgun DNA sequence assemblers in viral metagenomes has been published recently (Vázquez-Castellanos et al., 2014). None of the tested assemblers achieved convincing results. Whole-genome assemblers are known to perform poorly on metagenomes because they assume even coverage (Pop, 2009; Laserson et al., 2011; Peng et al., 2011; Lai et al., 2012; Namiki et al., 2012; Scholz et al., 2012). An assembler that was designed to work with uneven depth encountered in metagenomes is IDBA-UD (Peng et al., 2011) and it has been tested for viral genome assembly in simulated viral metagenomes with good results (Aguirre de Cárcer et al., 2014).

If the genome sequence of the host is known and the target virus not endogenous, the reads generated from the host can be subtracted prior to assembly by mapping which can increase virus contig lengths (Daly et al., 2015).

Other successful approaches in assembly of viral genomes from mixed samples are strategies that involve **iterative assembly**. The genome of Bas-Congo virus that was associated with an outbreak of human cases of acute hemorrhagic fever in the Democratic Republic of Congo in 2009 (Grard et al., 2012), was assembled with software (PRICE) that applies iterative rounds of contig extension on paired-end reads (Ruby et al., 2013). This targeted way of assembly is depending on a seed sequence, such as a read or contig that was initially identified by sequence similarity to a known virus. The seed is then extended by establishing local assemblies at both ends with the remaining reads which are merged to form a new contig. The target genome grows with each extension step. PRICE has been applied to a number of metagenome datasets, and aided, among others, in the discovery of a frequent contaminant in spin columns (Naccache et al., 2013), the identification of novel rhabdoviruses and bunyaviruses in mosquitos (Coffey et al., 2014), and the assembly of a novel nidovirus from a ball python (Stenglein et al., 2014). A similar, iterative assembly approach is applied by IVA (Iterative Virus Assembler, Hunt et al., 2015), developed for RNA virus genomes with uneven sequencing depth. Extension steps are carried out more conservatively than with PRICE, leading to more accurate assemblies of HIV and Influenza sequencing samples. The usefulness of IVA for metagenome and virus discovery data has yet to be shown.

KEY CONCEPT 3 | Iterative assembly

Sequence assembly consists of searches of overlaps, alignment, and merging of sequences. Computational limitations however prohibit most assemblers to perform exhaustive overlap searches. In iterative assembly, the resulting contiguous sequences (contigs) and singletons of the initial assembly are subjected to assembly by the same or a different assembly algorithm. This process is repeated until no new contigs can be found. Iterative assembly disregards coverage information and is therefore well-suited for metagenomics samples where coverage biases might exist.

Both assemblers were designed for paired-end short read data. To achieve the assembly of long 454 metagenome reads, an algorithm was developed that sequentially uses an overlap-layout-consensus (OLC) assembler (Newbler, Roche 454) followed by several rounds of greedy assembly (CAP3, Huang and Madan, 1999) until convergence (Schürch et al., 2014). Application of this algorithm to a data set derived from a cell culture supernatant containing a virus isolated from tissue of a dead white-beaked dolphin (*Lagenorhynchus albirostris*) lead to long contigs of a novel dolphin rhabdovirus that was closely related to fish rhabdoviruses of the genera *Perhabdovirus* and *Vesiculovirus* (Siegers et al., 2014). These contigs were combined with fragments derived by other assembly methods to retrieve a full viral genome (Smits et al., 2014). Such a sequential assembly strategy for short-read data using an OLC and a de Bruijn graph assembler is also proposed by Deng et al. (Deng et al., 2015) and shows great promise for metagenome assembly in general.

Identification of a Seed Fragment

After assembly it is essential to identify at least one fragment as originating from the target virus. This is most often performed

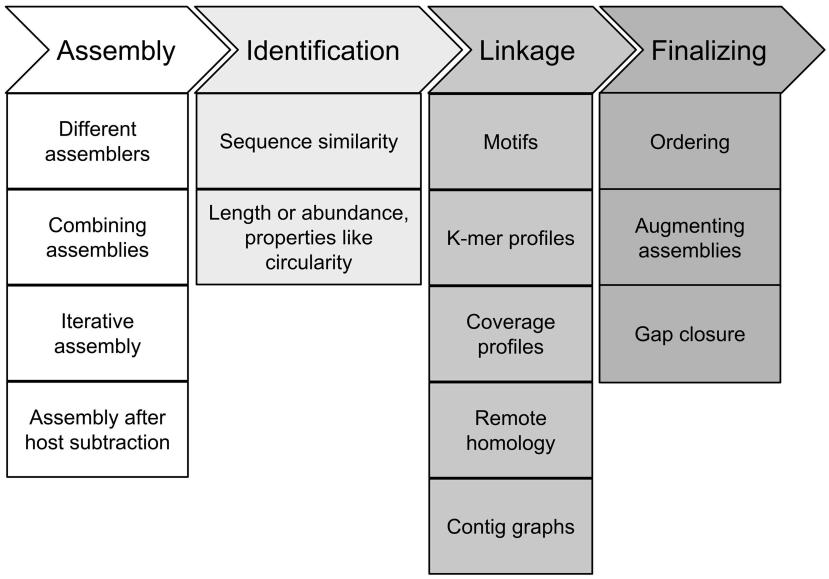

FIGURE 3 | Steps in recovery of full-length viral genomes from metagenomes. For description see article text.

by sequence similarity to known viruses, for example by BLAST analysis (Mokili et al., 2012). In the absence of sequence similarity, contig length, and abundance of the fragment can be an indication for the presence of an abundant organism with a comparably short genome within the metagenome (Bellas et al., 2015). Additional information gained from the sequence alone, like circularity, can also give an indication for the possible viral origin of a fragment (Mokili et al., 2013).

Linkage of Fragments

Even after optimal or nearly-optimal assembly of a metagenome, no full viral target genome might have been recovered, because of drops in coverage or miss-assemblies that prohibit the extension of contigs to full length genomes. Or, in case of viruses with segmented genomes, segments can only be separately assembled. In these cases, additional methods are necessary to link the fragments together. This can be achieved by motif searches, coverage profile binning, kmer profiling, and other methods.

Motif Discovery and Search

DNA, RNA, or protein motifs with or without assigned functions are present in the genetic material of viruses (and other life forms). Sequence motifs that are potentially useful for linkage of fragments are motifs that occur several times in a single genome. These motifs can be useful for linkage of missing content, or the linkage of distinct segments. Well-known examples are the conserved motifs that act as mRNA start and poly(A)/stop site in viruses of the order *Mononegavirales (*non-segmented negative-strand RNA viruses) (Kolakofsky et al., 2005; Penno et al., 2015) or motifs that act as leader sequence or transcription-regulation sequences in viruses of the order *Nidovirales* (families *Coronaviridae, Arteriviridae*, and *Roniviridae*) (Pasternak et al., 2006). These elements are conserved in each genome, but the motifs are distinct for the different families. Other positive-strand RNA viruses synthesize subgenomic RNAs with similar or different mechanisms (Miller and Koev, 2000; Lozano and Martínez-Salas, 2015) which can often be linked to the presence of conserved motifs in the genome. The existence of these sequence motifs in *Mononegavirales* and *Nidovirales* could provide an ideal target for motif search and detection strategies for linkage of fragments. Also viruses with segmented genomes, like Influenza A, have been shown to harbor conserved sequence motifs (ElHefnawi et al., 2011).

If the presence of a sequence motif is not obvious, *de novo* motif discovery on a putative seed contig can be performed with motif discovery methods, e.g., with MEME (Bailey et al., 2015), given that this contig is long enough to harbor several occurrences of the motif. Matches of these motifs to other sequences then can be detected with MAST (Bailey et al., 2015). A wide range of applications for motif discovery and searches are available (Das and Dai, 2007) which can accommodate differing data properties and individual preference.

An example of this is the retrieval of a full genome of a red fox fecal rhabdovirus from feces of red foxes from Spain (Bodewes et al., 2014b) that was enabled by detection of a highly conserved junctional motif (Smits et al., 2014). The occurrence of this motif in other, unassigned fragments, was used to link all fragments of the viral genome. The motif was highly specific for the red fox fecal rhabdovirus genome which suggested, given sufficiently long contigs, that motif search could be an effective method to link fragments.

Coverage Profile Binning

Binning of coverage profiles refers to the clustering of contigs based on their read coverage, utilizing the differences in

abundance of the organism within the metagenome sample. To this end, the read coverage for each contig has to be known, either by extraction of this information from the assembly or by mapping of the reads to the generated contigs. The coverage per base is calculated and contigs with a similar coverage are binned.

Binning of metagenomic contigs by coverage profiles is widely used in general metagenomics (Alneberg et al., 2014) and can be used to recover rare bacterial species from metagenomes (Albertsen et al., 2013). In viral metagenomes, coverage profile binning was used to assemble viral genomes across different human gut metagenomes avoiding the use of reference strains (Nielsen et al., 2014) and to verify cross-assemblies of a novel bacteriophage present in 73% of all publicly available gut metagenomes (Dutilh et al., 2014). However, if amplification is applied during metagenome processing, a coverage bias can be introduced to the sequencing data (Karlsson et al., 2013; Rosseel et al., 2013). In such cases, coverage profile binning does not always have the desired effect of linking fragments from the same source (Smits et al., 2014). Therefore, coverage profile binning can be applied in cases in which amplification is unnecessary or the introduction of a coverage bias was excluded.

K-mer Profiling

Differing sequence composition of the target genome compared to other genomes present in the metagenome can be used to cluster sequences by k-mer frequency profiling. **K-mer profiles** are used in a wide range of sequence similarity searches in bioinformatics. In metagenomics, k-mer frequency profiling is applied for alignment-free similarity analyses between sequences (Sims et al., 2009; Trifonov and Rabadan, 2010; Comin et al., 2015), especially for sequence assembly, quality control (Plaza Onate et al., 2015) and for taxonomic profiling and binning methods (Edwards et al., 2012; Silva et al., 2014; Dröge et al., 2015). There, k-mer analysis allows to overcome data analysis challenges associated with growing data volumes and short read lengths. The sensitivity of taxonomic classification for viral metagenome datasets however does not reach the *de facto* gold standard level of tBLASTX analyses (Vázquez-Castellanos et al., 2014), especially for longer reads(Edwards et al., 2012) and is, like BLAST searches, limited by the quality and comprehensiveness of the reference data set. This is especially true for taxonomic profiling methods, which analyse k-mer content in relation to a set of taxon-specific marker genes or genomes. Binning methods use k-mer profiles to cluster sequences based on similarity of their profiles and allow draft genome recovery (Dröge et al., 2015). Ranking of k-mer profiles, based on their similarity to the seed contig, which is a method similar to binning, was successfully used to link fragments of two rhabdovirus genomes (Smits et al., 2014). Other applications of k-mer analysis in viral metagenomes include nucleotide composition analysis, which are special cases of k-mers (1- and 2-mers) which was successfully used to infer a host for novel picorna-like viral sequences found in gut metagenomes (Kapoor et al., 2010).

KEY CONCEPT 4 | K-mer profiles
Frequency of all possible DNA sequences of length k within one DNA sequence. For example, if $k = 4$, one possible k-mer would be CGTA. In total, a k-mer profile contains frequency information on 4^k (256, in case of 4-mers) k-mers.

Other Methods for Linkage of Fragments

If the sequence similarity of a part of the viral target genomes to sequences in the databases is low, it is sometimes still possible to apply remote homology detection methods (Kuchibhatla et al., 2014), such as PSI-BLAST and HMMER3 (Altschul, 1997; Finn et al., 2011) and profile-profile comparison (HHpred Kuchibhatla et al., 2014), HHblits (Remmert et al., 2012), FFAS (Jaroszewski et al., 2011), WebPRC (Brandt and Heringa, 2009). To apply remote homology detection it is necessary that the viral family of the target genome was identified, for example by the identification of a highly conserved stretch, and that sequence profiles of this families are present in the respective databases [e.g., in pFAM (Finn et al., 2014) or Uniprot (Magrane and Consortium, 2011)] or can be produced from multiple sequence alignments.

Another method to link fragments is to extract all information obtained from an assembly by looking at the original contig graphs (Mulyukov and Pevzner, 2002). Information on adjacency of contigs can be found in such graphs and extracted for genome finishing. This method has successfully been applied when sequencing the gram-negative bacteria *Rickettsia prowazekii* (Nagarajan et al., 2010). Contig adjacency information is especially useful if repeats exist that are longer than the read length, and contigs were split at positions where the reconstruction was ambiguous.

Finalizing

After linkage of fragments it can be necessary to determine the order of the fragments by generating overlaps between contigs. Order might be inferred from paired end data, if available, but can also be deduced from augmented assemblies, produced by applying different assembly and alignment parameters. Wrong assemblies at contig ends can be the reason that no further overlap has been found by an assembler. Editing of contig ends improves wrong assemblies and facilitates subsequent gap closing with a different assembler.

Combination of different assemblies of the data can help to close potential gaps and confirm the sequences.

Concluding Remarks

In this review, we discussed some methods for assembly and linkage of viral genomes from metagenomes. They overcome often-encountered challenges associated with the extraction of full-length viral genomes from metagenomes: no or little similarity to viral sequences in databases, uneven coverage of the target genomes, and intrapopulation diversity or sequencing errors leading to incomplete contigs.

In silico genome finishing methods for recovery of a full-length viral genome from a metagenome are cost-effective and fast because they avoid re-sequencing. Selection of the most effective methods for assembly and finishing of viral genomes depends highly on the sample: host, originating tissue, library

preparation, sequencing technique, and depth will lead to different methods for finishing. In general, recovery consists of four different steps (**Figure 3**): assembly, identification, linkage, and finalizing. All methods are depending on the presence of sufficient sequencing data of the target genome, which in turn is related to a sufficiently high load of the viral target genome in the metagenome sample, and the quality of sample preparation and sequencing protocols in the first place. If target genome loads are not high enough, re-sequencing of the metagenome sample can be a solution, with either a complementary sequencing technique, mate-pair, or paired-end reads, or longer sequences or at a higher depth (Nagarajan et al., 2010; Grard et al., 2012).

In some cases, even though the coverage of the organism is sufficient, viral genomes cannot completely be resolved *in silico* despite exploitation of all possibilities described here. This was the case for the genome of a python nidovirus (PNV, Bodewes et al., 2014a; Smits et al., 2014) which had an unresolvable 5′ end and the PNV genome (accession nr KJ935003) putatively lacks some genome information there. Since also results of 5′RACE PCRs were inconclusive, most likely due to fragmentation of the RNA in the original lung tissue, other options are needed to obtain the complete genome. For instance, another next-generation sequencing platform could be used to confirm the findings of the 454 data. In addition, it could be of interest to set up an *in vitro* culture system to obtain a high titer virus stock with less background and use that as input material, like it was done for the DRV (Osterhaus et al., 1993).

To effectively use k-mer profiles, coverage profile binning and motif search to link fragments, it is essential that long contigs (>1 kb) are produced by the assembly. Coverage and k-mer profile analysis perform better on long fragments because of the noise associated with the high numbers of features in case of k-mers or introduced by uneven coverage (Smits et al., 2014). Motif search is limited by occurrence of the motif, which is more likely in long contigs. Therefore, it is advisable to apply the linkage methods only to well-assembled, long fragments. Successful assembly on the other hand starts with the selection of an assembler suited for sequencing technique and read length, and, if applicable, makes use of paired-end information. By combining assemblies (thus the contigs) or assemblers (thus iterative assembly), the length of contigs can be increased. However, in order to directly recover full-length viral genomes from metagenomes, the development of efficient and dedicated metagenome assemblers that consider the characteristics of viromes and viral genomes is needed.

Author Contributions

RB and AS conceived the study. AS and SS designed the experiments. AS and RB carried out the research. AS and SS prepared the first draft of the manuscript. RB, SS, AR, and WB contributed materials. SS, MK, and AO participated in the discussion and writing of the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Funding

This work was partially funded by the Virgo Consortium, funded by the Dutch government project number FES0908, by Netherlands Genomics Initiative (NGI) project number 050-060-452 and ZonMW TOP project 91213058. It was also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 643476 (COMPARE) and grant agreement No 634650 (VIROGENESIS). It was also supported in part by a grant from the Niedersachsen-Research Network on Neuroinfectiology of the Ministry of Science and Culture of Lower Saxony, Germany. AR holds a Post doc fellowship awarded by the Department of Education, Universities and Research of the Basque Government (Ref. DKR-2012-64) and was partially supported by the Research group on "Systematics, Biogeography, and Population Dynamics" (Basque Government; Ref. IT317-10; GIC10/76).

Acknowledgments

The authors wish to thank Jurre Y. Siegers for characterization of the DRV genome.

References

Aguirre de Cárcer, D., Angly, F. E., and Alcamí, A. (2014). Evaluation of viral genome assembly and diversity estimation in deep metagenomes. *BMC Genomics* 15:989. doi: 10.1186/1471-2164-15-989

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538. doi: 10.1038/nbt.2579

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res.* 43, W39–W49. doi: 10.1093/nar/gkv416

Beer, M., Conraths, F. J., and van der Poel, W. H. M. (2013). "Schmallenberg virus"–a novel orthobunyavirus emerging in Europe. *Epidemiol. Infect.* 141, 1–8. doi: 10.1017/S0950268812002245

Bellas, C. M., Anesio, A. M., and Barker, G. (2015). Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Front. Microbiol.* 6:656. doi: 10.3389/fmicb.2015.00656

Bodewes, R., Bestebroer, T. M., van der Vries, E., Verhagen, J. H., Herfst, S., Koopmans, M. P., et al. (2015). Avian Influenza A(H10N7) virus-associated mass deaths among harbor seals. *Emerg. Infect. Dis.* 21, 720–722. doi: 10.3201/eid2104.141675

Bodewes, R., Lempp, C., Schürch, A. C., Habierski, A., Hahn, K., Lamers, M., et al. (2014a). Novel divergent nidovirus in a python with pneumonia. *J. Gen. Virol.* 95, 2480–2485. doi: 10.1099/vir.0.068700-0

Bodewes, R., Ruiz-Gonzalez, A., Schürch, A. C., Osterhaus, A. D. M. E., and Smits, S. L. (2014b). Novel divergent rhabdovirus in feces of red fox, Spain. *Emerg. Infect. Dis.* 20, 2172–2174. doi: 10.3201/eid2012.140236

Brandt, B. W., and Heringa, J. (2009). webPRC: the Profile Comparer for alignment-based searching of public domain databases. *Nucleic Acids Res.* 37, W48–W52. doi: 10.1093/nar/gkp279

Chew, Y. V., and Holmes, A. J. (2009). Suppression subtractive hybridisation allows selective sampling of metagenomic subsets of interest. *J. Microbiol. Methods* 78, 136–143. doi: 10.1016/j.mimet.2009.05.003

Claas, E. C., Osterhaus, A. D., van Beek, R., De Jong, J. C., Rimmelzwaan, G. F., Senne, D. A., et al. (1998). Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet* 351, 472–477. doi: 10.1016/S0140-6736(97)11212-0

Coffey, L. L., Page, B. L., Greninger, A. L., Herring, B. L., Russell, R. C., Doggett, S. L., et al. (2014). Enhanced arbovirus surveillance with deep sequencing: identification of novel rhabdoviruses and bunyaviruses in Australian mosquitoes. *Virology* 448, 146–158. doi: 10.1016/j.virol.2013.09.026

Comin, M., Leoni, A., and Schimd, M. (2015). Clustering of reads with alignment-free measures and quality values. *Algorithms Mol. Biol.* 10, 4. doi: 10.1186/s13015-014-0029-x

Cutler, S. J. (2010). Relapsing fever–a forgotten disease revealed. *J. Appl. Microbiol.* 108, 1115–1122. doi: 10.1111/j.1365-2672.2009.04598.x

Daly, G. M., Leggett, R. M., Rowe, W., Stubbs, S., Wilkinson, M., Ramirez-Gonzalez, R. H., et al. (2015). Host subtraction, filtering and assembly validations for novel viral discovery using next generation sequencing data. *PLoS ONE* 10:e0129059. doi: 10.1371/journal.pone.0129059

Das, M. K., and Dai, H.-K. (2007). A survey of DNA motif finding algorithms. *BMC Bioinformatics* 8(Suppl. 7):S21. doi: 10.1186/1471-2105-8-S7-S21

de Graaf, M., and Fouchier, R. A. M. (2014). Role of receptor binding specificity in influenza A virus transmission and pathogenesis. *EMBO J.* 33, 823–841. doi: 10.1002/embj.201387442

Deng, X., Naccache, S. N., Ng, T., Federman, S., Li, L., Chiu, C. Y., et al. (2015). An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. *Nucleic Acids Res.* 43, e46. doi: 10.1093/nar/gkv002

Dröge, J., Gregor, I., and McHardy, A. C. (2015). Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics* 31, 817–824. doi: 10.1093/bioinformatics/btu745

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5, 4498. doi: 10.1038/ncomms5498

Edwards, R. A., Olson, R., Disz, T., Pusch, G. D., Vonstein, V., Stevens, R., et al. (2012). Real time metagenomics: using k-mers to annotate metagenomes. *Bioinformatics* 28, 3316–3317. doi: 10.1093/bioinformatics/bts599

ElHefnawi, M., Alaidi, O., Mohamed, N., Kamar, M., El-Azab, I., Zada, S., et al. (2011). Identification of novel conserved functional motifs across most Influenza A viral strains. *Virol. J.* 8:44. doi: 10.1186/1743-422X-8-44

Fauci, A. S., and Morens, D. M. (2012). The perpetual challenge of infectious diseases. *N. Engl. J. Med.* 366, 454–461. doi: 10.1056/NEJMra1108296

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Garcia-Garcerà, M., Garcia-Etxebarria, K., Coscollà, M., Latorre, A., and Calafell, F. (2013). A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin. *PLoS ONE* 8:e74914. doi: 10.1371/journal.pone.0074914

Grard, G., Fair, J. N., Lee, D., Slikas, E., Steffen, I., Muyembe, J.-J., et al. (2012). A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathog.* 8:e1002924. doi: 10.1371/journal.ppat.1002924

Howe, A., and Chain, P. S. G. (2015). Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Front. Microbiol.* 6:678. doi: 10.3389/fmicb.2015.00678

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877.

Hunt, M., Gall, A., Ong, S. H., Brener, J., Ferns, B., Goulder, P., et al. (2015). IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* 31, 2374–2376. doi: 10.1093/bioinformatics/btv120

Jaroszewski, L., Li, Z., Cai, X., Weber, C., and Godzik, A. (2011). FFAS server: novel features and applications. *Nucleic Acids Res.* 39, W38–W44. doi: 10.1093/nar/gkr441

Kapoor, A., Simmonds, P., Lipkin, W. I., Zaidi, S., and Delwart, E. (2010). Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses. *J. Virol.* 84, 10322–10328. doi: 10.1128/JVI.00601-10

Karlsson, O. E., Belák, S., and Granberg, F. (2013). The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. *Biosecur. Bioterror.* 11(Suppl. 1), S227–S234. doi: 10.1089/bsp.2013.0008

Kolakofsky, D., Roux, L., Garcin, D., and Ruigrok, R. W. H. (2005). Paramyxovirus mRNA editing, the "rule of six" and error catastrophe: a hypothesis. *J. Gen. Virol.* 86, 1869–1877. doi: 10.1099/vir.0.80986-0

Koopmans, M. (2013). The expanding list of zoonotic influenza viruses. *Lancet. Respir. Med.* 1, 756–757. doi: 10.1016/S2213-2600(13)70233-9

Kuchibhatla, D. B., Sherman, W. A., Chung, B. Y. W., Cook, S., Schneider, G., Eisenhaber, B., et al. (2014). Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently "orphan" viral proteins. *J. Virol.* 88, 10–20. doi: 10.1128/JVI.02595-13

Kuiken, T., Leighton, F. A., Fouchier, R. A. M., LeDuc, J. W., Peiris, J. S. M., Schudel, A., et al. (2005). Public health. Pathogen surveillance in animals. *Science* 309, 1680–1681. doi: 10.1126/science.1113310

Lai, B., Ding, R., Li, Y., Duan, L., and Zhu, H. (2012). A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics* 28, 1455–1462. doi: 10.1093/bioinformatics/bts162

Laserson, J., Jojic, V., and Koller, D. (2011). Genovo: de novo assembly for metagenomes. *J. Comput. Biol.* 18, 429–443. doi: 10.1089/cmb.2010.0244

Lozano, G., and Martínez-Salas, E. (2015). Structural insights into viral IRES-dependent translation mechanisms. *Curr. Opin. Virol.* 12, 113–120. doi: 10.1016/j.coviro.2015.04.008

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford).* 2011:bar009. doi: 10.1093/database/bar009

Marí Saíz, A., Weiss, S., Nowak, K., Lapeyre, V., Zimmermann, F., Dux, A., et al. (2014). Investigating the zoonotic origin of the West African Ebola epidemic. *EMBO Mol. Med.* 7, 17–23. doi: 10.15252/emmm.201404792

Miller, R. R., Montoya, V., Gardy, J. L., Patrick, D. M., and Tang, P. (2013). Metagenomics for pathogen detection in public health. *Genome Med.* 5, 81. doi: 10.1186/gm485

Miller, W. A., and Koev, G. (2000). Synthesis of subgenomic RNAs by positive-strand RNA viruses. *Virology* 273, 1–8. doi: 10.1006/viro.2000.0421

Mokili, J. L., Dutilh, B. E., Lim, Y. W., Schneider, B. S., Taylor, T., Haynes, M. R., et al. (2013). Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PLoS ONE* 8:e58404. doi: 10.1371/journal.pone.0058404

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004

Morse, S. S., Mazet, J. A. K., Woolhouse, M., Parrish, C. R., Carroll, D., Karesh, W. B., et al. (2012). Prediction and prevention of the next pandemic zoonosis. *Lancet* 380, 1956–1965. doi: 10.1016/S0140-6736(12)61684-5

Mulyukov, Z., and Pevzner, P. A. (2002). EULER-PCR: finishing experiments for repeat resolution. *Pac. Symp. Biocomput.* 7, 199–210. doi: 10.1142/9789812799623_0019

Naccache, S. N., Greninger, A. L., Lee, D., Coffey, L. L., Phan, T., Rein-Weston, A., et al. (2013). 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. doi: 10.1128/JVI.02323-13

Nagarajan, N., Cook, C., Di Bonaventura, M., Ge, H., Richards, A., Bishop-Lilly, K. A., et al. (2010). Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC Genomics* 11:242. doi: 10.1186/1471-2164-11-242

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40, e155. doi: 10.1093/nar/gks678

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828. doi: 10.1038/nbt.2939

Osterhaus, A. D., Broeders, H. W., Teppema, J. S., Kuiken, T., House, J. A., Vos, H. W., et al. (1993). Isolation of a virus with rhabdovirus morphology from a white-beaked dolphin (*Lagenorhynchus albirostris*). *Arch. Virol.* 133, 189–193.

Pasternak, A. O., Spaan, W. J. M., and Snijder, E. J. (2006). Nidovirus transcription: how to make sense...? *J. Gen. Virol.* 87, 1403–1421. doi: 10.1099/vir.0.81611-0

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 27, i94–i101. doi: 10.1093/bioinformatics/btr216

Penno, C., Sharma, V., Coakley, A., O'Connell Motherway, M., van Sinderen, D., Lubkowska, L., et al. (2015). Productive mRNA stem loop-mediated transcriptional slippage: crucial features in common with intrinsic terminators. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1984–E1993. doi: 10.1073/pnas.1418384112

Plaza Onate, F., Batto, J.-M., Juste, C., Fadlallah, J., Fougeroux, C., Gouas, D., et al. (2015). Quality control of microbiota metagenomics by k-mer analysis. *BMC Genomics* 16:183. doi: 10.1186/s12864-015-1406-7

Pollack, M. P., Pringle, C., Madoff, L. C., and Memish, Z. A. (2013). Latest outbreak news from ProMED-mail: novel coronavirus – Middle East. *Int. J. Infect. Dis.* 17, e143–e144. doi: 10.1016/j.ijid.2012.12.001

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief. Bioinform.* 10, 354–366. doi: 10.1093/bib/bbp026

Prachayangprecha, S., Schapendonk, C. M. E., Koopmans, M. P., Osterhaus, A. D. M. E., Schürch, A. C., Pas, S. D., et al. (2014). Exploring the potential of next-generation sequencing in detection of respiratory viruses. *J. Clin. Microbiol.* 52, 3722–3730. doi: 10.1128/JCM.01641-14

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9, 173–175. doi: 10.1038/nmeth.1818

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 1, 289–297. doi: 10.1016/j.coviro.2011.06.004

Rosseel, T., van Borm, S., Vandenbussche, F., Hoffmann, B., van den Berg, T., Beer, M., et al. (2013). The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing. *PLoS ONE* 8:e76144. doi: 10.1371/journal.pone.0076144

Ruby, J. G., Bellare, P., and Derisi, J. L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda).* 3, 865–880. doi: 10.1534/g3.113.005967

Schmieder, R., and Edwards, R. (2012). Insights into antibiotic resistance through metagenomic approaches. *Future Microbiol.* 7, 73–89. doi: 10.2217/fmb.11.135

Scholz, M. B., Lo, C.-C., and Chain, P. S. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* 23, 9–15. doi: 10.1016/j.copbio.2011.11.013

Schürch, A. C., Schipper, D., Bijl, M. A., Dau, J., Beckmen, K. B., Schapendonk, C. M. E., et al. (2014). Metagenomic survey for viruses in Western Arctic caribou, Alaska, through iterative assembly of taxonomic units. *PLoS ONE* 9:e105227. doi: 10.1371/journal.pone.0105227

Siegers, J. Y., van de Bildt, M. W. G., van Elk, C. E., Schürch, A. C., Tordo, N., Kuiken, T., et al. (2014). Genetic relatedness of dolphin rhabdovirus with fish rhabdoviruses. *Emerg. Infect. Dis.* 20, 1081–1082. doi: 10.3201/eid2006.131880

Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E., and Edwards, R. A. (2014). FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* 2:e425. doi: 10.7717/peerj.425

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2677–2682. doi: 10.1073/pnas.0813249106

Smits, S. L., Bodewes, R., Ruiz-Gonzalez, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D. M. E., et al. (2014). Assembly of viral genomes from metagenomes. *Front. Microbiol.* 5:714. doi: 10.3389/fmicb.2014.00714

Smits, S. L., and Osterhaus, A. D. (2013). Virus discovery: one step beyond. *Curr. Opin. Virol.* 3, e1–e6. doi: 10.1016/j.coviro.2013.03.007

Stenglein, M. D., Jacobson, E. R., Wozniak, E. J., Wellehan, J. F. X., Kincaid, A., Gordon, M., et al. (2014). Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius. *MBio* 5, e01484–e01414. doi: 10.1128/mBio.01484-14

Taylor, L. H., Latham, S. M., and Woolhouse, M. E. (2001). Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356, 983–989. doi: 10.1098/rstb.2001.0888

Trifonov, V., and Rabadan, R. (2010). Frequency analysis techniques for identification of viral genetic data. *MBio* 1:e00156-10. doi: 10.1128/mBio.00156-10

Van Leeuwen, M., Williams, M. M. W., Koraka, P., Simon, J. H., Smits, S. L., and Osterhaus, A. D. M. E. (2010). Human picobirnaviruses identified by molecular screening of diarrhea samples. *J. Clin. Microbiol.* 48, 1787–1794. doi: 10.1128/JCM.02452-09

Vázquez-Castellanos, J. F., García-López, R., Pérez-Brocal, V., Pignatelli, M., and Moya, A. (2014). Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* 15:37. doi: 10.1186/1471-2164-15-37

Woolhouse, M. E. J., and Gowtage-Sequeria, S. (2005). Host range and emerging and reemerging pathogens. *Emerg. Infect. Dis.* 11, 1842–1847. doi: 10.3201/eid1112.050997

Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., Gloeckner, F. O., et al. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443, 950–955. doi: 10.1038/nature05192

Conflict of Interest Statement: Dr. Albert D. M. E. Osterhaus is partly employed by Viroclinics Biosciences B.V., Rotterdam, Netherlands. The other 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.