frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE

published: 18 December 2014

doi: 10.3389/fmicb.2014.00714

Assembly of viral genomes from metagenomes

Saskia L. Smits[1,2†], Rogier Bodewes[1†], Aritz Ruiz-Gonzalez[3,4,5], Wolfgang Baumgärtner[6], Marion P. Koopmans[1,7], Albert D. M. E. Osterhaus[1,2,8] and Anita C. Schürch[1]*

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

[2] Viroclinics Biosciences, Rotterdam, Netherlands

[3] Department of Zoology and Animal Cell Biology, University of the Basque Country (UPV/EHU), Vitoria-Gasteiz, Spain

[4] Systematics, Biogeography and Population Dynamics Research Group, Lascaray Research Center, University of the Basque Country (UPV/EHU), Vitoria-Gasteiz, Spain

[5] Conservation Genetics Laboratory, National Institute for Environmental Protection and Research (ISPRA), Bologna, Italy

[6] Department of Pathology, University of Veterinary Medicine Hannover, Hannover, Germany

[7] Centre for Infectious Diseases Research, Diagnostics and Screening, National Institute for Public Health and the Environment, Bilthoven, Netherlands

[8] Center for Infection Medicine and Zoonoses Research, Hannover, Germany

Edited by:
Richard J. Hall, Institute of Environmental Science and Research, New Zealand

Reviewed by:
Hirokazu Kimura, National Institute of Infectious Diseases, Japan
Karen Dawn Weynberg, Australian Institute of Marine Science, Australia
Patrick Jon Biggs, Massey University, New Zealand

***Correspondence:**
Anita C. Schürch, Department of Viroscience, Erasmus Medical Center, PO Box 2040, Rotterdam 3000 CA, Netherlands
e-mail: a.schurch@erasmusmc.nl

† *These authors have contributed equally to this work.*

Viral infections remain a serious global health issue. Metagenomic approaches are increasingly used in the detection of novel viral pathogens but also to generate complete genomes of uncultivated viruses. *In silico* identification of complete viral genomes from sequence data would allow rapid phylogenetic characterization of these new viruses. Often, however, complete viral genomes are not recovered, but rather several distinct contigs derived from a single entity are, some of which have no sequence homology to any known proteins. *De novo* assembly of single viruses from a metagenome is challenging, not only because of the lack of a reference genome, but also because of intrapopulation variation and uneven or insufficient coverage. Here we explored different assembly algorithms, remote homology searches, genome-specific sequence motifs, k-mer frequency ranking, and coverage profile binning to detect and obtain viral target genomes from metagenomes. All methods were tested on 454-generated sequencing datasets containing three recently described RNA viruses with a relatively large genome which were divergent to previously known viruses from the viral families *Rhabdoviridae* and *Coronaviridae*. Depending on specific characteristics of the target virus and the metagenomic community, different assembly and *in silico* gap closure strategies were successful in obtaining near complete viral genomes.

Keywords: virus, pathogen, metagenome, virome, virus discovery, assembly, viral metagenomics

INTRODUCTION

Human and animal populations are continuously confronted with emerging viral infections (Delwart, 2007; Lipkin, 2010; Smits and Osterhaus, 2013). In a proportion of patients and animals suffering from disease, no pathogens can be detected using a range of sensitive diagnostic assays, suggesting the presence of unidentified viruses in human and animal populations (Bloch and Glaser, 2007; Denno et al., 2012). Classically, new viruses were identified by standard molecular detection methods, virus replication in tissue culture or animal experiments. Nowadays, in order to discover and characterize new or (re-) emerging viruses, metagenome sequencing is increasingly being used to identify viral pathogens. In addition, these techniques are more and more often being used to generate complete genomes of uncultivated viruses, but also other organisms (Delwart, 2007; Lipkin, 2010; Iverson et al., 2012; Albertsen et al., 2013; Smits and Osterhaus, 2013; Handley et al., 2014).

Metagenomic strategies to virus discovery rely on sequence-independent amplification of nucleic acids combined with next generation sequencing platforms instead of targeting specific genomic loci, thereby generating DNA sequences (i.e., reads) that align to various genomic locations for the numerous genomes present in the sample, including non-microbes (Sharpton, 2014). Common random amplification methods are multiple displacement amplification (MDA) or sequence-independent single-primer amplification (SISPA) (Hutchison et al., 2005; Spits et al., 2006; Delwart, 2007; Djikeng et al., 2008; Lipkin, 2010; Smits and Osterhaus, 2013). The advantages of sequence-independent amplification are simplicity and relative speed and the ability to identify and sequence hundreds of viruses simultaneously thereby allowing detection of new or previously unrecognized viruses that are highly divergent from already described ones (Bodewes et al., 2014a,c). Inherent to the approach is that a large fraction of the metagenome consists of sequences of other organisms than the viral targets, including host sequences, archaea, bacteria, and bacteriophages, despite physical enrichment strategies for virus particles that are often applied (Van Leeuwen et al., 2010; Kostic et al., 2012; Van Den Brand et al., 2012; Wylie et al., 2012; Bodewes et al., 2013; Schurch et al., 2014).

Metagenomic sequence data analysis with the aim to identify viral sequences presents several challenges. Datasets are relatively complex and large. In addition, the obtained viral

reads in metagenomes can either originate from taxonomically informative genomic regions and even provide insight in the biological function of the encoded protein or originate from less conserved genomic regions for which biological functions are difficult to assign. Current strategies rely mostly on filtering steps to remove host nucleic acid from metagenomes either before or after sequencing and analysis of the data, including assembly and homology searches against annotated sequences in public databases (Woyke et al., 2006; Chew and Holmes, 2009; Schmieder and Edwards, 2012; Garcia-Garcera et al., 2013; Prachayangprecha et al., 2014; Schurch et al., 2014). Untargeted metagenomic approaches have enabled the identification of numerous newly emerging or previously unidentified viral pathogens in recent years. However, obtaining full-length viral genomes from metagenomic datasets remains challenging.

The number of reads obtained from a specific virus in metagenome samples is correlated to the viral load in the sample under study (De Vries et al., 2012; Prachayangprecha et al., 2014). In some cases, the number of reads in the sample is sufficient to permit enough read overlaps to establish longer contiguous sequences (contigs). However, direct assembly of complete viral and bacterial genomes from metagenomic data can involve a large amount of manual curation (Handley et al., 2014; Sharpton, 2014) as most pathogen genomes are not completely represented by reads and most viral communities are highly diverse (Mavromatis et al., 2007; Mende et al., 2012). Currently, full-length viral genomes are often obtained with additional experimental approaches based on PCR amplification with specific primers designed on obtained reads or contigs and/or 5′ and 3′ RACE PCR in combination with a Sanger sequencing approach (Van Leeuwen et al., 2010; Siegers et al., 2014). However, by optimally mining sequences in metagenomes, the likelihood and speed of identifying viral reads and the level of viral genome completeness can be increased and the need for laboratory follow-up minimized. In the present study, we describe and compare methods to obtain viral target genomes from metagenomes using a retrospective approach on 454-sequencing datasets containing three recently described viruses from the families *Rhabdo-* and *Coronaviridae*.

METHODS

DATASETS

The first metagenome dataset was obtained from a cell culture supernatant (CCS) containing a rhabdovirus-like virus isolated from tissue collected from a stranded white-beaked dolphin (*Lagenorhynchus albirostris*) (Osterhaus et al., 1993; Siegers et al., 2014). Genetic and phylogenetic characterization of the dolphin rhabdovirus (DRV) revealed that it was closely related to rhabdoviruses of the genera *Perhabdovirus* and *Vesiculovirus* found in fish (Siegers et al., 2014). In the second case, a highly divergent rhabdovirus, called red fox fecal rhabdovirus (RFFRV) was identified during a metagenomic survey of feces of red foxes from Spain (*Vulpes vulpes*) (Bodewes et al., 2014c). The last metagenome dataset was from lung tissue of a dead Indian python (*Python molurus*) with pneumonia, in which a novel nidovirus belonging to the family *Coronaviridae* within the order *Nidovirales* was identified. It was the first description of a reptile nidovirus (python nidovirus, PNV) and phylogenetic analysis placed this virus in the subfamily *Torovirinae* (Bodewes et al., 2014a). These datasets were acquired using a random sequence amplification and deep sequencing approach on a 454 GS Junior instrument (Roche) as previously described by Van Leeuwen et al. (2010), Bodewes et al. (2013, 2014a,c). At present full-length genomes (DRV) or expected complete coding sequences (PNV, RFFRV) are available.

ASSEMBLY METHODS

Four different assembly methods, exhaustive iterative assembly (Schurch et al., 2014), CLC Genomics Workbench 6.0.4 assembler (CLC bio, Aarhus, Denmark), Genovo version 0.4 (Laserson et al., 2011), and Newbler 2.5 (Roche), were compared in their efficiency of detecting viral reads in the three metagenome datasets. The originally used method was iterative exhaustive assembly. Iterative exhaustive assembly of sequences is part of a virus discovery pipeline written in the python programming language (Python 2.7) that includes trimming of reads and initial assembly with Newbler (454GS Assembler version 2.7, Roche), with standard parameters. Trimmed reads and initial contigs were subjected to assembly by CAP3 (VersionDate: 12/21/07) (Huang and Madan, 1999) with standard parameters. The resulting singletons and contigs were iteratively assembled by CAP3 until no new contigs were formed.

Subsequently, the trimmed reads were mapped back to the identified taxonomic units with Newbler (454 GSMapper version 2.7, Roche) with standard parameters (Schurch et al., 2014). CLC Genomics Workbench 6.0.4 assembler (CLC bio, Aarhus, Denmark) was run with the previously trimmed reads with automatic bubble and word size. Genovo version 0.4 was run with 40 iterations and otherwise default values (Laserson et al., 2011). Newbler 2.5 (Roche) was run with default values.

DETERMINATION OF TAXONOMIC CONTENT

Contigs and singletons of the iterative assembly approach that were longer than 75 bases were filtered with Dustmasker which is part of the NCBI-BLAST+ 2.2.25 suite of tools for sequences that contain more than 60% low complexity sequences (Camacho et al., 2009). After filtering of low complexity sequences, the remaining taxonomic units were subjected to a BLASTN search against a database that contained only nucleotide sequences from birds (Aves, taxonomic identifier 8782), carnivores (Carnivora, taxID 33554), primates (Primates, taxID 9443), rodents (Rodentia, taxID 9989), and ruminants (Ruminantia, taxID 9845) with an *e*-value cut-off of 0.001 for subtraction of potential host sequences. Sequences without hits in the host-BLAST were then subjected to a BLASTN search against the entire nt database with an *e*-value cut-off of 0.001. All sequences without hits were then subjected to a BLASTX search against protein sequences present in the nr database. BLAST hits were categorized by assigning a taxonomic category.

The percentage of viral reads in the sequence datasets and read coverage of the target genome using different assembly methods were determined by mapping trimmed reads to reference

genomes with GSMapper Version 2.7 (Roche) with a minimum overlap identity of 80%.

REMOTE HOMOLOGY SEARCH

All contigs were translated in six frames. Hidden Markov Models (HMMs) of PFAM families associated with *Rhabdoviridae* (pfam14314, pfam00945, pfam02484, pfam03216, pfam03342, pfam03012, pfam03397, pfam04785, pfam05554, pfam00922, pfam00974, pfam06326) were used to search the translated contigs of the metagenome datasets with rhabdoviruses with HMMER3.1 (Punta et al., 2012). Accordingly, HMMs of 45 PFAM families associated with *Coronaviridae* (pfam05213, pfam06460, pfam04694, pfam09408, pfam08717, pfam08716, pfam08715, pfam06478, pfam06471, pfam05409, pfam03262, pfam03053, pfam02723, pfam01601, pfam01600, pfam00937, pfam08779, pfam12383, pfam12379, pfam12133, pfam12124, pfam12093, pfam11963, pfam11633, pfam11501, pfam11395, pfam11289, pfam11030, pfam10943, pfam09401, pfam08710, pfam06336, pfam06145, pfam05528, pfam04753, pfam03905, pfam03622, pfam03620, pfam03617, pfam03187, pfam02398, pfam01635, pfam09399, pfam01831) were used to search the translated contigs of the PNV metagenome.

MOTIF DISCOVERY AND MOTIF SEARCH

Motif sequence patterns were discovered with MEME Version 4.9.1 (Bailey et al., 2009) by allowing any number of repetitions on the sequence. The best scoring detected motif distributed over the seed contig was then used to search the motif in the collection of all contigs longer than 500 bases in all three datasets with MAST (Bailey et al., 2009) with an *e*-value lower than 0.1.

COVERAGE PROFILE BINNING

The average coverage of all contigs identified using exhaustive iterative assembly was calculated by dividing the number of reads covering the contigs by its length, as determined by the mapping procedure of the virus discovery pipeline. Frequency of binned coverage profiles was visualized in R statistical package version 3.1.

K-mer FREQUENCY RANKING

K-mer frequency was determined with the Bioconductor package biostrings (Pages et al., in press) for 3mers to 8mers for contigs larger than 1 kb in R statistical package version 3.1 (Team, 2012). Absolute differences between the k-mer frequencies of the seed contig and all other contigs were summed among different k-mer lengths and ranked, and visualized in relation to contig size.

ACCESSION NUMBERS

Viral genome sequences used in this study were taken from Genbank, accession numbers KF958252 (DRV), KF823814 (RFFRV), and KJ935003 (PNV).

RESULTS

EVALUATION OF DIFFERENT ASSEMBLY ALGORITHMS

The objective of this study was to test and evaluate methods to increase the likelihood and speed of identifying viral reads and the level of viral genome completeness from metagenomic datasets generated on the 454-sequencing platform. The three 454-sequencing datasets obtained from a CCS, a red fox fecal (RFF) metagenome, and a tiger python lung tissue (TPLT) metagenome contained 69,358, 56,174, and 135,812 sequence reads, respectively (**Table 1**). These reads were analyzed with an automatic analysis pipeline that included stringent quality and length trimming, exhaustive iterative assembly, and low complexity filtering (Schurch et al., 2014). A total of 28,207, 32,455 and 50,024 reads from the CCS and the RFF and TPLT metagenome, respectively were subjected to homology searches (**Table 2**). The analysis showed a high variety among almost all taxonomic categories in the three different datasets (**Table 2**). The overall viral content determined by homology search was relatively low (0.72%) in dataset 3 (TPLT metagenome), and high (30.21 and 68.05%) in datasets 1 and 2 (CCS and RFF metagenome, **Table 2**).

Iterative exhaustive assembly resulted in assembled metagenomes containing between 40–60% of the original total amount of obtained reads (**Table 1**) of which the virome has a large dynamic range of reads depending on the sample under analysis. Unsurprisingly, the CCS dataset from an assumingly relatively pure virus culture supernatant had a high viral content, consisting predominantly of DRV. The viromes of the RFF metagenome showed a much smaller percentage of the RFFRV indicating the presence of multiple different viruses (**Tables 1**, **2**).

Genomes of DRV, RFFRV, and PNV were not completely assembled by the exhaustive iterative assembly approach implemented in the automated analysis pipeline. The largest contigs

Table 1 | Description of deep sequencing datasets.

	CCS	RFF	TPLT
Total number of reads	69358	56174	135812
Assembled metagenome (%)	40.67	57.78	36.83
Reads identified by homology search as obtained from target virus (%)	27.67	5.82	0.11
Reads retrospectively obtained from target virus (%)	69.52	13.58	26.14

Cell culture supernatant (CCS) containing Dolphin rhabdovirus (DRV), red fox feces (RFF) metagenome containing red fox fecal rhabdovirus (RFFRV) and python lung tissue (TPLT) metagenome containing python nidovirus (PNV).

Table 2 | Taxonomic composition of deep sequencing datasets.

	CCS	RFF	TPLT
Unassigned	0.04	0.19	0.97
Virus	68.05	30.21	0.72
Unknown	3.90	10.35	49.39
Eukaryota	27.40	37.90	35.22
Bacteria	0.61	21.34	13.64
Archea	0	0	0.06

Taxonomic composition per read in percentage of cell culture supernatant (CCS) containing Dolphin rhabdovirus (DRV), red fox feces (RFF) metagenome containing red fox fecal rhabdovirus (RFFRV) and python lung tissue (TPLT) metagenome containing python nidovirus (PNV). Unassigned, best BLAST hit without taxonomic assignment. Unknown, no homology to any database entry.

were 7291 bases (64.32% of DRV, DRV seed contig) and 7682 bases (47.9% of RFFRV), respectively, of an expected size of 11 to 15 kb for *Rhabdoviridae* and 24,734 bases (73.68% of PNV) of an expected 30 kb for *Coronaviridae* (**Figures 1A–C**). Interestingly, retrospective mapping of reads to the viral target genomes showed that a large percentage of the sequences identified as "unknowns" by homology searches in the TPLT and RFF metagenome were actually derived from the target genome, most likely from parts of the target genomes without detectable similarity to any other viral protein in the BLAST database (**Tables 1**, **2**).

To evaluate if other assembly algorithms would be able to directly assemble the complete target viral genomes from the deep sequencing data, we compared the contigs assembled from trimmed reads by iterative assembly, CLC Genomics Workbench assembler, Genovo and Newbler. While Genovo and Newbler both produced many small contigs covering part of the target genomes (**Figures 1A–C**), CLC Genomics Workbench assembler and the iterative assembly approach produced a similar number of contigs (two to five). However, large contigs (>1 kb) produced by iterative assembly covered the target genomes more completely than any other set of contigs obtained with other assembly algorithms. None of the assemblers tested here was able to completely assemble the viral genomes from the reads into a single contig (**Figures 1A–C**).

The contigs produced by iterative assembly and CLC Genomics Workbench for DRV were clearly overlapping (**Figure 1A**) and could be fused to a single assembly of a complete DRV genome by manual curation. For RFFRV in dataset 2, contigs assembled by iterative assembly and Genovo overlapped. However, a very small overlap of only five nucleotides between position 4356 and 4361, probably due to the combination of a drop in coverage and a stretch of sequence with low complexity (**Figure 1B**) did not allow us to retrieve a complete viral genome. Moreover, with the exception of the largest contig (the RFFRV seed contig), no other RFFRV contigs had a homolog in the NCBI nucleotide or protein database. The minor overlap, in combination with absence of homology, prevented assembly of a complete RFFRV genome. Similarly, the overlaps between contigs of PNV obtained with different assembly algorithms, in combination with the absence of homology, were insufficient to conclusively obtain a full-length PNV genome. In the TPLT metagenome, a 24.7 kb contig had a stretch with an average identity of 29% amino acid identity to the replicase polyprotein of Berne virus, subfamily *Torovirinae* (**Figure 1C**). This contig was used as seed contig. Overall, the data indicate that iterative exhaustive assembly seems to perform best in terms of production of large contigs and coverage of target genomes compared to other assemblers. Thus, for further analysis we used the set of contigs produced by iterative assembly. It is of note, however, that using a combination of different assembly algorithms may result in a higher level of completeness of target genomes if not complete genome assembly.

REMOTE HOMOLOGY SEARCH

In absence of BLAST-detectable sequence homology to previously described viruses for some stretches of the target genomes we attempted to use methods that are able to detect remote homologs, i.e., profile hidden markov models (**Figures 1A–C**). To retrieve and link contigs of the viral target genomes we used profile HMMs of protein domains present in *Rhabdoviridae* and *Coronaviridae*, respectively. For *Rhabdoviridae*, 12 domains were present in PFAM, from nucleocapsid, spike and matrix proteins. Searching the translated contigs of the CCS containing DRV identified three regions (Vesiculovirus matrix protein—PF06326.7, Rhabdovirus nucleocapsid protein—PF00945.13 and the Rhabdovirus spike protein—PF00974.13) covered by several contigs (**Figure 1A**, contigs smaller than 1 kb not shown). The RFF metagenome did not give any hits (**Figure 1B**). For *Coronaviridae* in the TPLT metagenome, 44 HMMs were present in PFAM, again covering all proteins families. Three *Coronaviridae*-specific HMMs identified several contigs, the Coronavirus NSP13 (F06460.7), the RNA synthesis protein NSP10 (PF09401.5) and the Coronavirus RPol N-terminus (PF06478.8). However, in all three datasets, all identified domains were already identified in contigs with BLAST sequence homology to a closely related virus (**Figures 1A–C**). Identification of regions with remote homology to family-specific domains did therefore not result in acquisition of additional genomic regions that were not identified by the original iterative assembly method in combination with homology search by BLAST.

MOTIF SEARCH

A sequence motif is a DNA pattern that occurs repeatedly in a genome or in a group of related sequences. *De novo* motif discovery is independent of previously described motifs and their function. Motif discovery was performed on the seed contigs that showed homology to viruses of either the *Rhabdoviridae* or *Coronaviridae* family of metagenome 1 and 3 and on two adjacent, clearly overlapping RFFRV contigs of metagenome 2, including the RFFRV seed contig (**Figure 2B**). The highest scoring motif (**Figures S1A–C**) of each seed contig was then used to screen all available contigs of the deep sequencing datasets. Contigs were selected if they contained one or several occurrences of the motif at an *e*-value smaller than 0.01. Four additional DRV-matching contigs smaller than 1 kb (not indicated in **Figure 1A**) were identified in the CCS. In the RFF metagenome, four additional RFFRV contigs were identified and in the TLPT metagenome, one additional PNV contig exhibited the detected pattern (**Figures 1B–C**). No false positive contigs were identified with this method. Moreover, when the PNV motif was used to search all three viral genomes including the rhabdovirus genomes, it exhibited highest specificity for the PNV genome (*e*-value 4e-12), with *e*-values above 1 for RFFRV and DRV. The DRV and the RFFRV motif were most specific for their originating genome, while the *e*-value for the respective other rhabdovirus genome was relatively low (0.62 and 0.015), suggesting those motifs might be conserved among both rhabdovirus genomes. The high specificity of the motifs described here was also demonstrated when scanning all contigs of the three metagenomes: the motif discovered contigs of the respective viral genomes with a high specificity. The detection sensitivity however was restricted by the number of contigs that contained the motif. For example, the eight occurrences of the motif in the RFFRV genome were, with one exception, all found in intergenic regions. Contigs not containing intergenic regions could not be identified with this method.

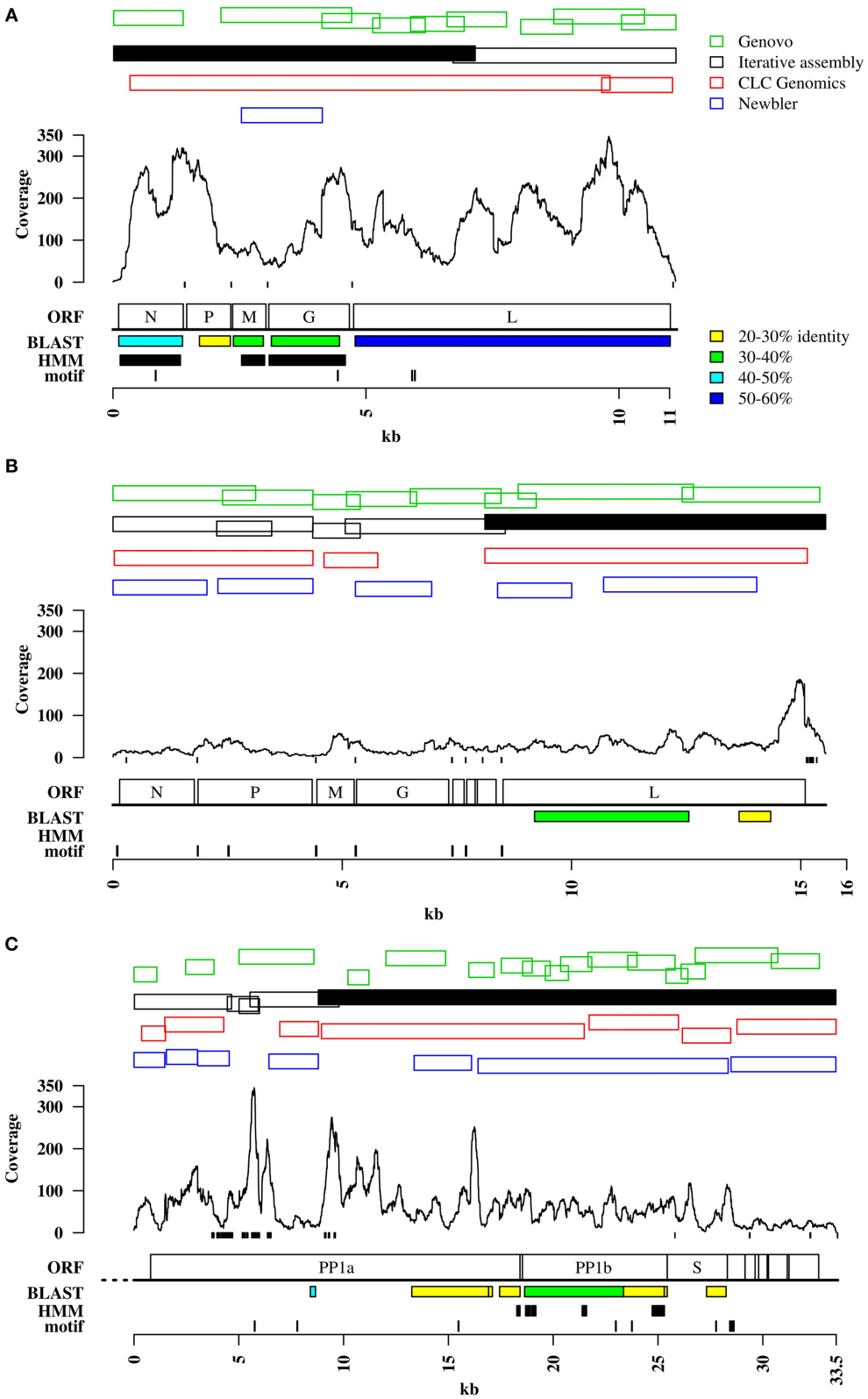

FIGURE 1 | Viral target genomes. Panels **(A–C)** contain information on read coverage and contigs matching the viral genomes of DRV **(A)**, RFFRV **(B)**, and PNV **(C)**, produced by different assembly algorithms. Shown are only contigs larger than 1 kb. Green: Contigs assembled through Genovo as described in the methods. Black outlined: Contigs assembled through iterative assembly. Black solid: Seed contig. Red: Contigs assembled through CLC Genomics workbench assembler. Blue: Contigs assembled through

(Continued)

FIGURE 1 | Continued
Newbler assembler. Small black boxes at the bottom of the read coverage line mark stretches of low sequence complexity. "ORF" indicates the genome organization as described below. "Motif" shows the location of sequence motifs. Motifs are shown in detail in **Figure S1**. "BLAST" shows regions with sequence homology as determined by BLASTX. Colored boxes show sequence identity to the best BLAST hit as indicated on top. "HMM" indicates region with remote homology identified by PFAM profiles, if any. Ruler at the bottom indicates sequence lengths in kilobases. **(A)** DRV, Dolphin rhabdovirus; N, nucleoprotein; P, phosphoprotein; M, matrix protein; G, glycoprotein; L, large protein. **(B)** RFFRV, Red fox fecal rhabdovirus; N, nucleoprotein; P, phosphoprotein; M, matrix protein; G, glycoprotein; L, large protein; no abbrevation, alpha 1,2,3 protein. **(C)** PNV, Python nidovirus; PP1a, polyprotein 1a; PP1b, polyprotein1b; S, spike glycoprotein; no abbreviations, minor membrane protein, membrane protein, nucleocapsid protein, minor membrane protein 2, putative hemagglutinin-neuraminidase protein. Striped line at 5′ end indicates putative unresolved 5′ end.

All occurrences of the detected motifs in the target genomes are indicated in **Figures 1A–C**. For both RFFRV and PNV, motif detection was able to identify contigs from genomic regions lacking BLAST or HMM-detectable homology. This information was sufficient to obtain the complete coding region of the RFFRV genome in combination with the iterative assembly approach.

COVERAGE PROFILE BINNING

In order to find additional contigs by coverage profile binning, the average coverage of every contig of the CCS and the two metagenomes was calculated by dividing the number of reads by the length of the contig. In the CCS dataset, an average coverage of 1.20 reads per base was achieved for the DRV seed contig. Accordingly, another contig that had a coverage of more than 1.1 read/base was obtained from the DRV genome (**Figure 2A**). All other contigs in the CCS dataset showed a lower coverage profile. In the RFF metagenome however, contigs with a similar coverage frequency as the seed contig of RFFRV (coverage of 0.42) were identified as putative plant genes, with a closest homolog to a hypothetical protein of *Medicago truncatula* (BLASTX *e*-value 6e-60, coverage of 0.45) or as part of the *Vulpes vulpes* mitochondrium (*e*-value 0.0, coverage of 0.58) (**Figure 2B**) and as two parts of a novel picobirnavirus, RFF picobirnavirus, isolate 40-2 (Bodewes et al., 2014b) (*e*-value 0.0, coverage of 0.299 and 0.99). In the TPLT metagenome, the seed contig of PNV was covered by 0.5 reads per base. Four of five contigs with a coverage profile of more than 0.2 were matching the PNV genome (**Figure 2C**), with the exception of one contig identified as hypothetical protein of *Clostridium thermocellum* (BLASTX *e*-value 4e-29). In conclusion, coverage profile binning identified additional contigs in two of the datasets.

K-mer FREQUENCIES

Another possible method to detect contigs that lack homology to known viruses but are part of the viral target genome is to determine k-mer frequencies. The frequency of every oligomer at length k was determined for $k = 3, 4, 5, 6, 7$ and 8 and ranked according to their absolute difference with the sum of the k-mer frequencies of the seed contig (**Figure 3**). Low ranking contigs have a similar k-mer frequency profile as the seed contig, whereas high ranking contigs differ in their k-mer frequency profile. For the CCS dataset, one 7.5 kb contig had a closely matching, high ranking k-mer frequency profile and was indeed originating from DRV (**Figure 3A**). Similarly, the two largest contigs (>3.5 kb) of RFFRV had the highest rank when compared to the frequency profile of the seed contig (**Figure 3B**). Two smaller contigs ranked at 14 and 24, suggesting that k-mer frequency profile clustering works better with long sequences. However, for PNV, the highest ranking, largest contig was obtained from the python host genome and contained among others a sequence for cytochrome C oxidase subunit (BLASTX *e*-value 0). Two large contigs matching the PNV genome ranked at 22 and 47, and two small contigs that were obtained from the PNV genome had an even higher rank (**Figure 3C**). While k-mer frequency ranking identified an additional part of DRV, and two large RFFRV contigs, all high-ranking contigs (rank 10 or less) of dataset 3 were unrelated to PNV.

DISCUSSION

We tested and compared different strategies to assist in identifying viral contigs and increasing the level of viral genome completeness from metagenomes. The retrospective nature of this study allowed us to compare the success of the strategies in retrieving three novel viral genomes from 454 metagenomic data. While different metagenome assembly strategies, especially for very large datasets of short read data, apply k-mer clustering or digital normalization and partitioning prior to assembly (Howe et al., 2014; Reddy et al., 2014), we here concentrated on strategies to link contigs after assembly *in silico*. Theoretically, these strategies can also be applied to contigs from metagenomes produced by other sequencing methods. A growing number of metagenome studies apply other next-generation sequencing techniques (e.g., Illumina), but 454 sequencing is still widely applied in viral metagenome studies, sometimes in combination with Illumina or PacBio sequences, because of the large read length (up to 800 bases) (De Vries et al., 2012; Grard et al., 2012; Philippe et al., 2013). In all three datasets presented here, the viral load of the samples was exceptionally high. Despite the high number of reads obtained from the target virus, direct assembly of the full genomes was not possible.

Metagenome assembly is a challenging task, because the number, nature, and abundance of the genomes present in the metagenome is unknown. Whole-genome assemblers are not suited for this task (Laserson et al., 2011; Peng et al., 2011; Lai et al., 2012; Namiki et al., 2012; Scholz et al., 2012). They assume even coverage and recognize high-coverage regions as repeats rather than a highly abundant species, or as an unevenly covered region introduced by amplification bias. Virus discovery relies heavily on low input material methods; therefore an amplification strategy is often necessary. Common random amplification methods, such as MDA or sequence-independent SISPA (Dean et al., 2002; Hutchison et al., 2005; Spits et al., 2006) are known to produce strong amplification biases leading to highly uneven coverage depths (Karlsson et al., 2013; Rosseel et al., 2013).

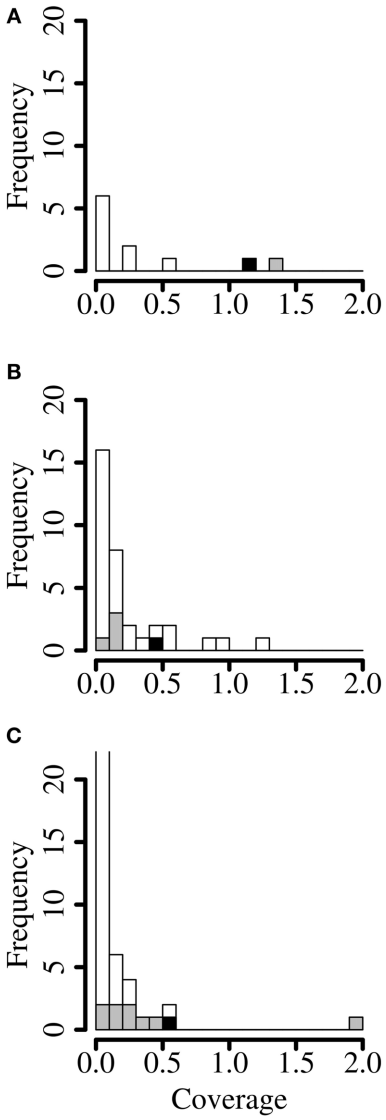

FIGURE 2 | Coverage profile binning. Histograms of coverage (reads per base) of each contig of **(A)** of cell culture supernatant containing Dolphin rhabdovirus, **(B)** red fox feces containing red fox fecal rhabdovirus and **(C)** python lung tissue containing python nidovirus. Gray: contigs mapping to the finished viral genome. Black: seed contig. The first bar in the last panel is truncated for visibility (47%). Shown are only contigs larger than 1 kb.

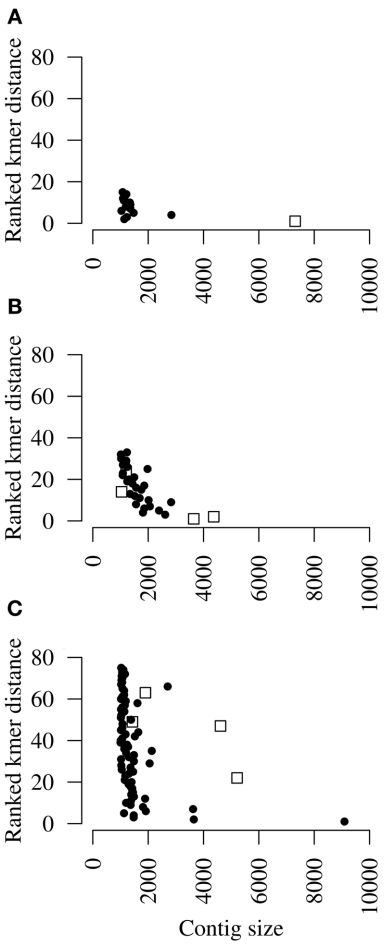

FIGURE 3 | K-mer profiling. Dot plots showing ranked k-mer distance of each contig when compared to the k-mer profile of the seed contig of **(A)** Dolphin rhabdovirus (DRV), **(B)** red fox fecal rhabdovirus (RFFRV), and **(C)** python nidovirus (PNV) in relation to contig lengths. Open boxes indicate contigs that were retrospectively identified as originating from the target genomes. Shown are only contigs larger than 1 kb.

Nevertheless the need for amplification makes these two methods still the most widely used in virus discovery (Allander et al., 2005; Djikeng et al., 2008; Hall et al., 2014). Introduction of amplification bias leads to stretches in the viral genome that are better covered than others. This can not only mislead assembly, it could also hamper detection of additional contigs by coverage profile binning. Accordingly, coverage binning was a successful strategy to link additional contigs for DRV and PNV, but not for RFFRV. Nevertheless, coverage profile binning was successfully applied to assemble viral genomes across a number of human gut metagenomes without the need of a reference (Nielsen et al., 2014) or to verify a cross-assembly of a novel bacteriophage in similar samples (Dutilh et al., 2014). Due to availability of a large amount of (fecal) sample for metagenome studies, amplification can often be avoided, which makes the application of coverage profile binning more straight-forward.

An additional issue in the case of viral metagenomes is the presence of distinct quasispecies sequences which can hamper direct assembly, especially at low sequencing depths. Using stringent assembly parameters that are necessary to avoid chimeras can lead to highly similar singletons or small contigs that are too diverse for assembly into a population sequence. This problem can be overcome by reference-guided assembly by a quasispecies assembler (Prosperi et al., 2013). However, this is currently not possible for divergent viruses which lack a reference genome. While many metagenome assemblers to date were designed to handle short-read data, only very few assemblers are dedicated to assembly of longer (i.e., 454) reads without any further

information such as paired-end or mate-pair information. For this study, we used different assemblers, including an overlap-layout consensus algorithm (Newbler), an assembler that uses a generative probabilistic model (Genovo; Laserson et al., 2011), a de Bruijn graph algorithm (CLC Genomics Workbench), and an assembly strategy applying the combination of an OLC and a greedy algorithm (iterative assembly; Schurch et al., 2014). None of these strategies lead to a full reconstruction of the genomes of the novel viruses, but produced fragmented contigs. Overlaps between the contigs were often not recognized because of misassembled contig ends (not indicated in **Figures 1A–C**). These misassembled ends could represent chimeric contigs, i.e., assembled from reads from different species, but also chimeric reads due to chimera formation during PCR. Chimerism can not only prohibit successful assembly but can also lead to misclassification of the taxonomic content of the metagenome sample (Mavromatis et al., 2007; Pignatelli and Moya, 2011; Mende et al., 2012). Taxonomic "misclassification" of reads in the analysis described here, however, was rather due to the large number of taxonomic units without a homolog in the sequence databases. These reads were then classified as "unknowns." Another challenge for recovery of viral genomes from metagenomes poses the segmented genomes of some viruses, with up to 12 segments for some viruses in the family *Reoviridae*, for example the Colorado tick fever virus (Attoui et al., 2000). Those segments can only be separately assembled, if possible, and need to be linked afterwards. The strategies described in this study can aid in identification of missing segments or contigs.

The strategy with the highest specificity was *de novo* motif discovery in the seed contig, and subsequent motif search in all contigs of the assembled metagenome. The (A/U)CU7 motif detected between open reading frames of RFFRV could serve as a transcription termination/polyadenylation sequence similar to other rhabdoviruses (Whelan et al., 2004).

Adjacent to this termination signal was a stretch of conserved nucleotides which might function as a transcription initiation signal. For the other detected motifs in DRV and PNV no obvious functions can be envisaged. However, their power to detect additional contigs matching the target genomes was only limited by the number of occurrences of the motif in the genome.

K-mer profile ranking detected large viral contigs with a similar profile as the seed contig in the CCS dataset and the RFF metagenome. In both cases, further manual curation or a laboratory follow-up would have been necessary to confirm the predictions made by this technique.

Assembly, in combination with motif discovery enabled retrieval of the complete RFFRV genome, with good results in k-mer frequency clustering. Two additional contigs of the PNV genome were identified by motif search, but linking of the remaining PNV contigs was only possible with frequency methods. Surprisingly, all methods applied here showed good results in retrieval of the full genome of DRV from the CCS dataset. This is most likely due to the high viral load which allowed assembly of the whole genome into two very long contigs in the first place. Therefore, we feel that the development of more efficient and dedicated metagenome assemblers, taking into account the specific characteristics of viral genomes, will lead to improved retrieval of viral pathogen genomes from metagenome sequences.

In conclusion, iterative exhaustive assembly, although highly stringent and thus excluding a large amount of data, is actually performing rather well compared to other assembly algorithms in that it covered the target genomes more completely than any other set of contigs obtained with other assembly algorithms. Nevertheless, the number of identified target virus reads and the level of viral genome completeness can be increased by combining data generated with different assembly algorithms. In addition, various methods can be applied to obtain additional genome fragments although a success rate cannot be predicted beforehand based on our analyses and probably depend largely on the dataset under study. These results indicate that a combination of these methods can be of great value to rapidly obtain additional genome information of a previously unknown virus.

AUTHOR CONTRIBUTIONS

Rogier Bodewes and Anita C. Schürch conceived the study. Anita C. Schürch designed the experiments. Anita C. Schürch, Rogier Bodewes carried out the research. Saskia L. Smits contributed to the design of experiments. Anita C. Schürch prepared the first draft of the manuscript. Rogier Bodewes, Saskia L. Smits, Aritz Ruiz-Gonzalez, Wolfgang Baumgärtner, contributed materials. Saskia L. Smits, Marion P. Koopmans, Albert D. M. E. Osterhaus 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.

GRANT INFORMATION

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. A. Ruiz-Gonzalez 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.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00714/abstract

Figure S1 | (A) Nucleotide sequence motif discovered in the Dolphin rhabdovirus (DRV). **(B)** Nucleotide sequence motif discovered in the red fox fecal rhabdovirus (RFFRV). **(C)** Nucleotide sequence motif discovered in the python nidovirus (PNV).

REFERENCES

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

Allander, T., Tammi, M. T., Eriksson, M., Bjerkner, A., Tiveljung-Lindell, A., and Andersson, B. (2005). Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12891–12896. doi: 10.1073/pnas.0504666102

Attoui, H., Billoir, F., Cantaloube, J. F., Biagini, P., De Micco, P., and De Lamballerie, X. (2000). Strategies for the sequence determination of viral dsRNA genomes. *J. Virol. Methods* 89, 147–158. doi: 10.1016/S0166-0934(00)00212-3

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Bloch, K. C., and Glaser, C. (2007). Diagnostic approaches for patients with suspected encephalitis. *Curr. Infect. Dis. Rep.* 9, 315–322. doi: 10.1007/s11908-007-0049-5

Bodewes, R., Lempp, C., Schurch, 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., Schapendonk, C. M., Van Den Brand, J. M., Osterhaus, A. D., and Smits, S. L. (2014b). Viral metagenomic analysis of feces of wild small carnivores. *Virol. J.* 11:89. doi: 10.1186/1743-422X-11-89

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

Bodewes, R., Van De Bildt, M. W., Schapendonk, C. M., Van Leeuwen, M., Van Boheemen, S., De Jong, A. A., et al. (2013). Identification and characterization of a novel adenovirus in the cloacal bursa of gulls. *Virology* 440, 84–88. doi: 10.1016/j.virol.2013.02.011

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

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

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5261–5266. doi: 10.1073/pnas.082089499

Delwart, E. L. (2007). Viral metagenomics. *Rev. Med. Virol.* 17, 115–131. doi: 10.1002/rmv.532

Denno, D. M., Shaikh, N., Stapp, J. R., Qin, X., Hutter, C. M., Hoffman, V., et al. (2012). Diarrhea etiology in a pediatric emergency department: a case control study. *Clin. Infect. Dis.* 55, 897–904. doi: 10.1093/cid/cis553

De Vries, M., Oude Munnink, B. B., Deijs, M., Canuti, M., Koekkoek, S. M., Molenkamp, R., et al. (2012). Performance of VIDISCA-454 in feces-suspensions and serum. *Viruses* 4, 1328–1334. doi: 10.3390/v4081328

Djikeng, A., Halpin, R., Kuzmickas, R., Depasse, J., Feldblyum, J., Sengamalay, N., et al. (2008). Viral genome sequencing by random priming methods. *BMC Genomics* 9:5. doi: 10.1186/1471-2164-9-5

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G., 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

Garcia-Garcera, M., Garcia-Etxebarria, K., Coscolla, 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

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., et al. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* 195, 194–204. doi: 10.1016/j.jviromet.2013.08.035

Handley, K. M., Bartels, D., O'Loughlin, E. J., Williams, K. H., Trimble, W. L., Skinner, K., et al. (2014). The complete genome sequence for putative H - and S-oxidizer Candidatus Sulfuricurvum sp., assembled *de novo* from an aquifer-derived metagenome. *Environ. Microbiol.* 16, 3443–3462. doi: 10.1111/1462-2920.12453

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4904–4909. doi: 10.1073/pnas.1402564111

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877. doi: 10.1101/gr.9.9.868

Hutchison, C. A. 3rd., Smith, H. O., Pfannkoch, C., and Venter, J. C. (2005). Cell-free cloning using phi29 DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* 102, 17332–17336. doi: 10.1073/pnas.0508809102

Iverson, V., Morris, R. M., Frazar, C. D., Berthiaume, C. T., Morales, R. L., and Armbrust, E. V. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335, 587–590. doi: 10.1126/science.1212665

Karlsson, O. E., Hansen, T., Knutsson, R., Lofstrom, C., Granberg, F., and Berg, M. (2013). Metagenomic detection methods in biopreparedness outbreak scenarios. *Biosecur. Bioterror.* 11(Suppl 1), S146–S157. doi: 10.1089/bsp.2012.0077

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* 22, 292–298. doi: 10.1101/gr.126573.111

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

Lipkin, W. I. (2010). Microbe hunting. *Microbiol. Mol. Biol. Rev.* 74, 363–377. doi: 10.1128/MMBR.00007-10

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., et al. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* 4, 495–500. doi: 10.1038/nmeth1043

Mende, D. R., Waller, A. S., Sunagawa, S., Jarvelin, A. I., Chan, M. M., Arumugam, M., et al. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* 7:e31386. doi: 10.1371/journal.pone.0031386

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. doi: 10.1007/BF01309754

Pages, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (in press). *Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms.* R package version 2.32.30. Available online at: http://www.bioconductor.org/packages/release/bioc/html/Biostrings.html

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

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., et al. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* 341, 281–286. doi: 10.1126/science.1239181

Pignatelli, M., and Moya, A. (2011). Evaluating the fidelity of *de novo* short read metagenomic assembly using simulated data. *PLoS ONE* 6:e19984. doi: 10.1371/journal.pone.0019984

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

Prosperi, M. C., Yin, L., Nolan, D. J., Lowe, A. D., Goodenow, M. M., and Salemi, M. (2013). Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. *Sci. Rep.* 3:2837. doi: 10.1038/srep02837

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucl. Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Reddy, R. M., Mohammed, M. H., and Mande, S. S. (2014). MetaCAA: a clustering-aided methodology for efficient assembly of metagenomic datasets. *Genomics* 103, 161–168. doi: 10.1016/j.ygeno.2014.02.007

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

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

Schurch, A. C., Schipper, D., Bijl, M. A., Dau, J., Beckmen, K. B., Schapendonk, C. M., 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

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* 5:209. doi: 10.3389/fpls.2014.00209

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

Smits, S. L., and Osterhaus, A. D. (2013). Virus discovery: one step beyond. *Curr. Opin. Virol.* doi: 10.1016/j.coviro.2013.03.007. [Epub ahead of print].

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., et al. (2006). Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.* 1, 1965–1970. doi: 10.1038/nprot.2006.326

Team, R. C. (2012). *R: A Language and Environment for Statistical Computing*. Available online at: http://web.mit.edu/r_v3.0.1/fullrefman.pdf

Van Den Brand, J. M., Van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D., et al. (2012). Metagenomic analysis of the viral flora of pine marten and European badger feces. *J. Virol.* 86, 2360–2365. doi: 10.1128/JVI.06373-11

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

Whelan, S., Barr, J., and Wertz, G. (2004). "Transcription and replication of non-segmented negative-strand RNA viruses," in *Biology of Negative Strand RNA Viruses: The Power of Reverse Genetics*, ed Y. Kawaoka (Berlin; Heidelberg: Springer), 61–119.

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

Wylie, K. M., Truty, R. M., Sharpton, T. J., Mihindukulasuriya, K. A., Zhou, Y., Gao, H., et al. (2012). Novel bacterial taxa in the human microbiome. *PLoS ONE* 7:e35294. doi: 10.1371/journal.pone.0035294

Conflict of Interest Statement: Drs. Albert D. M. E. Osterhaus and Saskia L. Smits are 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.

Received: 17 September 2014; paper pending published: 20 October 2014; accepted: 30 November 2014; published online: 18 December 2014.

Citation: Smits SL, Bodewes R, Ruiz-Gonzalez A, Baumgärtner W, Koopmans MP, Osterhaus ADME and Schürch AC (2014) Assembly of viral genomes from metagenomes. Front. Microbiol. ***5****:714. doi: 10.3389/fmicb.2014.00714*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.