VIRUS EVOLUTION

Virus Evolution, 2020, 6(2): veaa091

doi: 10.1093/ve/veaa091
Resources

Novel NGS pipeline for virus discovery from a wide spectrum of hosts and sample types

Ilya Plyusnin,[1,2,*,†] Ravi Kant,[2,3] Anne J. Jääskeläinen,[4] Tarja Sironen,[2,3] Liisa Holm,[1,5] Olli Vapalahti,[2,3,4] and Teemu Smura[3,4]

[1]Institute of Biotechnology, University of Helsinki, Helsinki 00014, Finland, [2]Department of Veterinary Bioscience, University of Helsinki, Helsinki 00014, Finland, [3]Department of Virology, University of Helsinki, Helsinki 00014, Finland, [4]Department of Virology and Immunology, University of Helsinki and Helsinki University Hospital, Helsinki 00014, Finland and [5]Organismal and Evolutionary Biology Research Program, University of Helsinki, Helsinki 00014, Finland

*Corresponding author: E-mail: ilja.pljusnin@helsinki.fi

†https://orcid.org/0000-0001-5988-0901

Abstract

The study of the microbiome data holds great potential for elucidating the biological and metabolic functioning of living organisms and their role in the environment. Metagenomic analyses have shown that humans, along with for example, domestic animals, wildlife and arthropods, are colonized by an immense community of viruses. The current Coronavirus pandemic (COVID-19) heightens the need to rapidly detect previously unknown viruses in an unbiased way. The increasing availability of metagenomic data in this era of next-generation sequencing (NGS), along with increasingly affordable sequencing technologies, highlight the need for reliable and comprehensive methods to manage such data. In this article, we present a novel bioinformatics pipeline called LAZYPIPE for identifying both previously known and novel viruses in host associated or environmental samples and give examples of virus discovery based on it. LAZYPIPE is a Unix-based pipeline for automated assembling and taxonomic profiling of NGS libraries implemented as a collection of C++, Perl, and R scripts.

Key words: bioinformatics pipeline; viral metagenomics; virus discovery; NGS data analysis; virome; taxonomic profiling.

1. Introduction

Our ability to produce sequence data in the rapidly growing field of genomics has surpassed our ability to extract meaningful information from it. Analyzing viral data are particularly challenging given the considerable variability in viruses and the low coverage of viral diversity in current databases. It is estimated that a vast majority of virus taxa are yet to be described and classified (Geoghegan and Holmes 2017). This challenge is further complicated by the high evolutionary rate of viruses leading to emergence of new virus lineages and the relative scarcity of viral genetic material in metagenomic samples (Rose et al. 2016). Next-generation sequencing (NGS) is a high throughput, impartial technology with numerous attractive features compared with established diagnostic methods for virus detection (Mokili, Rohwer, and Dutilh 2012). NGS-based studies have improved our understanding of viral diversity (Cantalupo et al. 2011). There is considerable interest within virology to explore the use of metagenomics techniques, specifically in the detection of viruses that cannot be cultured (Smits et al. 2015; Graf et al. 2016). Metagenomics can also be used to diagnose patients with rare or unknown disease aetiologies that would otherwise require multiple targeted tests (Pallen 2014) or emerging infections for which tests are yet to be developed.

In recent years, the role of the bacterial microbiome in health and disease has been acknowledged and studied extensively (Biedermann and Rogler 2015; KataOka 2016). Nonetheless, the influence of the viral constituent of the microbiome (i.e. virome) has received considerably less attention. Recent research has indicated that both pathogenic and commensal viral species can modulate host immune responses and thereby either prevent or induce diseases (Lim et al. 2015; Neil and Cadwell 2018). Additionally, recent research has revealed modifications in the virome that are related to diseases such as acquired immunodeficiency syndrome and inflammatory bowel disease (Norman et al. 2015). Accordingly, there is a need to identify novel viruses that may be established pathogens and to define wider links of the virome with health and disease. Beyond humans, the veterinary, wildlife, arthropod, and environmental viromes have large implications in for example, animal health, zoonotic emergence, and ecosystem research that require new tools to understand and study the virosphere.

Bioinformatics pipelines and algorithms designed for the analysis of NGS microbiome data can be separated into three groups. The first group includes pipelines for virome composition analysis. These pipelines mine the relative abundance and types of viruses present in a given sample. These pipelines include VirusSeeker (Zhao et al. 2017), Viral Informatics Resource for Metagenome Exploration (Wommack et al. 2012), viGEN (Bhuvaneshwar et al. 2018), the Viral MetaGenome Annotation Pipeline (Lorenzi et al. 2011), and MetaVir (Roux et al. 2014). The second group includes pipelines that are designed for bacterial composition analysis, such as MG-RAST (Meyer et al. 2008). Pipelines in the third group, such as MetaPhlan2 (Truong et al. 2015), Kraken2 (Wood, Lu, and Langmead 2019) and Centrifuge (Kim et al. 2016), can perform composition analysis for all known taxa. There are also a number of tools, pipelines, and algorithms for virus discovery, including Genome Detective (Vilsker et al. 2019), VIP (Li et al. 2016), PathSeq (Kostic et al. 2011), SURPI (Naccache et al. 2014), READSCAN (Naeem, Rashid, and Pain 2013), VirusFinder (Wang, Jia, and Zhongming 2013), and MetaShot (Fosso et al. 2017). Most of these pipelines are based on homology searches in the nucleotide (nt) or amino acid (aa) space against a local database of reference sequences. Notably, homology searches in the aa space are expected to retrieve more distant homologs, which is key for detecting divergent novel viral sequences.

The availability of robust bioinformatics pipelines for virome detection and annotation from NGS data continues to be one of the critical steps in many research projects. Pipelines are needed to efficiently detect viral sequences present in a complex mixture of host, bacterial, and other microbial sequences. The discovery of viral sequences depends on sequence alignment with other viral sequences in databases, as, in contrast to bacteria where 16S RNA is present in all taxa, viruses lack 'explanatory genes' found in all taxa.

Lazypipe offers several advantages to the existing methods for taxonomic profiling of viral NGS data. Lazypipe outsources homology search to a separate server eliminating the need to install and update local sequence databases. This is helpful in both reducing the workload on the user and ensuring that all the latest viral sequences are covered by the homology search. Additionally, this feature can significantly reduce the threshold for employing Lazypipe by the less technically savvy researchers. Lazypipe uses SANSparallel (Somervuo and Holm 2015) to search for aa homologs in the UniProtKB database. Searching for homologs in the protein space is expected to retrieve more distant viral homologs than searches with nucleotide sequences (Zhao et al. 2017). Furthermore, SANSparallel is ∼100 times faster compared with the BLASTP search (Somervuo and Holm 2015), which is the default search engine employed by nearly all other annotation pipelines that search the protein space. Lazypipe assembles and annotates viral contigs, thus reducing the workload on the downstream analysis. This function is supported by some other pipelines, such as Metavir (Roux et al. 2014) and virMine (Garretto, Hatzopoulos, and Putonti 2019), but both Metavir and virMine use a much slower BLASTP search and require installation of local sequence databases. Lazypipe implements a flexible stepwise architecture that allows re-execution of individual steps or parts of the analysis. This architecture addresses the increased risk of execution failure that is inherent to the analysis of large NGS libraries. Lazypipe supports data formats that can be used both by human researchers and automated tools. Results are output in the form of intuitive excel tables and interactive graphs, but also, in the form of standardized taxonomic profiles that can be integrated with automated workflows.

Lazypipe does not perform direct taxonomic binning of reads, but instead links these to database sequences via the assembled contigs. This approach results in a very high accuracy of taxon retrieval (see Section 3); however, this may come at the cost of lower accuracy for read binning. The accuracy of read binning was not accessed in this work since the main objective was to construct a highly accurate taxonomic profiler. Still, we provide the option to retrieve reads linked to any reported taxon or contig.

2. Materials and methods

2.1 Laboratory procedures and Samples

The faecal sample from diarrhoeic American mink (*Neovison vison*) was collected in September 2015 from a fur production farm in Finland, as described previously (Smura et al. 2016). The processing and sequencing of the sample were conducted using a protocol described in Conceição-Neto et al. (2015).

The human patient samples were derived from the diagnostic unit of Helsinki University Hospital Laboratory and stored in −80°C. In this study, RNA was extracted using either QIAamp Viral RNA kit (Qiaqen Inc., Valencia, USA) or EasyMag (bioMerieux) according to the manufacturer's instructions, followed by real time polymerase chain reaction (PCR) detection described in Kuivanen et al. (2019) for tick-borne encephalitis virus (TBEV), in Haveri et al. (2020) for severe acute respiratory syndrome coronovirus 2 (SARS-CoV-2), and in Mäki-Tanila et al. (2016) for entero- and parechoviruses. This study was done according to research permits HUS/32/2018 §16 for project TYH2018322 and HUS/44/2019 §13 for projects TYH2018322 and M1023TK001.

Prior to sequencing, samples were treated with DNase I (Thermo Fisher, Waltham, USA) and purified with Agencourt RNA Clean XP magnetic beads (Beckman Life sciences, Indianapolis, USA). Ribosomal RNA was removed using a NEBNext rRNA depletion kit (New England BioLabs, Ipswich, USA) according to the manufacturers protocol. The sequencing library was prepared using a NEBNext Ultra II RNA library prep kit (New England BioLabs).

Libraries were quantified using NEBNext Library Quant kit for Illumina (New England BioLabs). Pooled libraries were sequenced on an Illumina MiSeq platform, using either MiSeq v2 reagent kit with 150 base pair (bp) paired-end reads or a MiSeq v3 reagent kit with 300 bp paired-end reads.

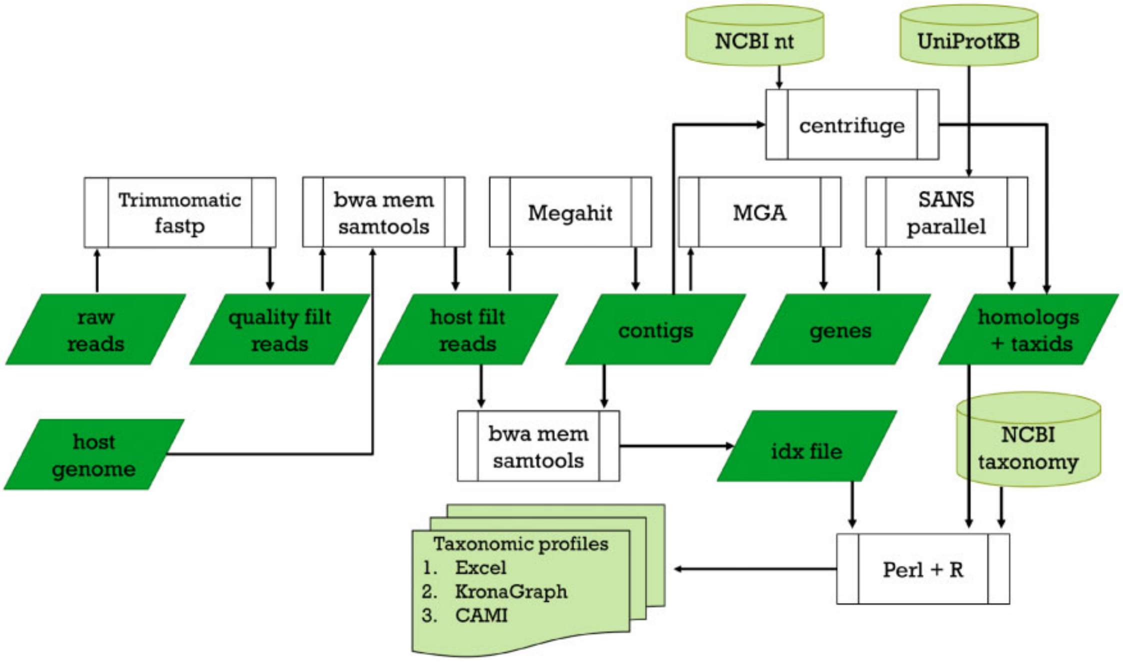

Figure 1. Lazypipe flowchart. Binaries and scripts are displayed in white, input and output files in green.

2.2 Unix pipeline for assembly, taxonomic profiling and binning of NGS data

We implemented a UNIX pipeline for automated assembly and taxonomic profiling of NGS libraries. The pipeline also performs taxonomic binning of the assembled contigs. The workflow of our pipeline is illustrated in Fig. 1. Our pipeline was implemented as a collection of Perl, C++, and R programs with a command-line user interface written in Perl. Our implementation allows for execution of the whole pipeline with a single command or performing each analysis step separately. This allows for great flexibility when working with large NGS libraries. Lazypipe user manual and source code are freely available from project's website (https://www.helsinki.fi/en/projects/lazypipe) and git repository (https://bitbucket.org/plyusnin/lazypipe/.) Each pipeline step is described in more detail below.

Paired-end libraries in FASTQ format (Cock et al. 2010) serve as input. First, primers, short reads and low-quality reads are removed with Trimmomatic (Bolger, Lohse, and Usadel 2014) or fastp (Chen et al. 2018). Then host reads are filtered by aligning reads against the host genome with BWA-MEM (Li 2013) and removing reads with high scoring alignments with SAMtools (Li et al. 2009). A threshold of 50 was selected by comparing the pre-assembly mapping of reads to the host genome to the post-assembly taxonomic binning of reads (without the host genome filtering). Comparing these for several samples showed that a threshold of 50 removes between 90 and 94 per cent of reads that are assigned to *Eukaryota* in post-assembly binning while removing only 1–10 per cent of reads that are assigned to Viruses (data omitted). Decreasing this threshold to 30 resulted in removal of only 63–64 per cent of reads assigned to *Eukaryota* and increased removal of reads assigned to Viruses (18–29%). Increasing threshold to 100 again decreased the number of filtered eukaryotic reads (to 74–86%) with only slight improvement on the number of retained virus reads (0–9%). For the simulated metagenome (Fosso et al. 2017), setting the threshold to 50 results in filtering 99.93 per cent of host reads and only 0.05 per cent of viral reads (excluding the endogenous retroviral reads, which are filtered to a large extent). Thus, we selected fifty as a working threshold although we recognize that a more robust optimization can be performed.

In the next step, reads are assembled with MEGAHIT (Li et al. 2015) or Velvet (Zerbino and Birney 2008). MEGAHIT is used by default as this was the overall best assembler in the CAMI competition (Sczyrba et al. 2017). The pipeline then scans for gene-like regions in the assembled contigs with MetaGeneAnnotator (Noguchi, Taniguchi, and Itoh 2008) (default) or MetaGeneMark (Zhu, Lomsadze, and Borodovsky 2010) and translates these to aa sequences using BioPerl (Stajich et al. 2002). Extracted aa sequences are queried against UniProtKB using the SANSparallel (Somervuo and Holm 2015) server. Top hits that pass a bitscore threshold value are used to assign contigs to the NCBI taxonomy ids. Note that contigs with several genes can be assigned to several taxonomy ids. We also support an alternative strategy of mapping contigs directly against the NCBI nucleotide collection database (NCBI nt). This is done by querying contigs with Centrifuge against NCBI nt and using alignments that pass a threshold value for the alignment score to assign contigs to taxonomy ids. We refer to this alternative version of our pipeline as the Lazypipe-nt.

Reads that passed host genome filtering are realigned to contigs using BWA-MEM top hits that pass a pre-set threshold on the alignment scores. Read distribution tables are generated using SAMtools (Li et al. 2009).

Next, taxonomy links generated by SANSparallel and read distribution tables are processed into an abundance table, which summarizes the number of contigs and reads binned to each taxon. Contigs that are mapped to two or more taxa may contribute different fractions of reads to different taxa according to the selected weighting model. In the *taxacount* model reads are distributed equally between all taxa linked to the contig. In the *bitscore* model reads are distributed according to the sum of bitscores from the database hits. Here, for each taxon linked to the contig the weight is the ratio of two sums: the sum of bitscores for the contig and taxon, and the sum of all bitscores for the taxon. In our benchmarking we used the bitscore model, which showed a slightly better performance. The raw abundance table is converted to an Excel file (using R) with several spreadsheets, each providing a different view of the acquired data. These views include the abundance of virus taxa (excluding bacteriophages), bacteriophages, bacteria, eukaryotes and, optionally, other high-level domains. For each of these groups, abundances are reported at three taxonomic levels (family, genus, and species). This arrangement allows for a rapid overview of NGS results and convenient ‘zooming in’ on the taxa of interest. Taxonomic abundances are also presented as an interactive Krona graph (Ondov, Bergman, and Phillippy 2011), which supports dynamic exploration of abundancies across different taxa. We also convert taxonomic abundances to CAMI Profiling Output Format (Sczyrba et al. 2017). By providing standardized output we support benchmarking of our pipeline by unbiased third-party evaluation initiatives such as CAMI (Sczyrba et al. 2017). Standardized output also aims to support simple and stable integration in automated workflows. To simplify accessibility, contigs for different taxa are sorted into a directory structure that follows the taxonomic hierarchy. A summary table is printed that lists all contigs for viruses, bacteriophages, bacteria and eukaryotes along with hits from SANSparallel or Centrifuge search.

In taxonomic profiling reads and contigs from the least abundant taxa have the highest risk of being misclassified. The organizers of the first CAMI competition addressed this problem by removing the last percentile of read distributions assigned by the compared taxonomic profilers (Sczyrba et al. 2017). We implement a similar strategy by assigning each taxon a cumulative frequency distribution value (*csum*), which sums read frequencies mapped to that taxon and the more abundant taxa. We also assign confidence scores based on the csum score: the [0.95%] interval is assigned Confidence 1, the [95%, 99%] Confidence 2 and the tail values [99%, 100%] are assigned Confidence 3. For a typical NGS library taxa with confidence Score 1 will be true positives, those with Score 3 (i.e. the last percentile) will be false positives and those with Score 2 will represent borderline cases.

As an additional feature, Lazypipe offers an option to create interactive graphical reports that display the location and variation in viral contigs relative to reference viral genomes. This requires installation of a local database of viral reference genomes, which is then searched for taxa matching virus taxa found by the homology search. Contigs are aligned against the matching reference genomes with BWA-MEM and the resulting alignments are displayed with Integrative Genomics Viewer (Thorvaldsdóttir, Robinson, and Mesirov 2013) in an internet browser.

In the last step, we turn to quality control by generating graphical reports. The quality of the original library and assembly are monitored with histograms and key statistics. We also present the number of reads retained at consecutive pipeline steps: after quality filtering with Trimmomatic, after host genome filtering, after assembling, and after gene detection. These are summarized as the survival-rate-plots.

2.3 Benchmarking performance

We evaluated our pipeline on the following two sets of data: a simulated metagenome from the MetaShot project (Fosso et al. 2017) and a mock-virome and bacterial mock-community data (SRA reference SRR3458569; Conceição-Neto et al. 2015).

The MetaShot metagenome is a 20.5 M PE 2 × 150 Illumina library simulated with ART [13]. We mapped reads in this library using accession numbers in read id-fields to 107 viral taxids, 99 prokaryote taxids and the human genome (94.5% of all reads). Strain taxids were further mapped using NCBI taxonomy to species, genus, family, order, and superkingdom taxids resulting in eighty-four species and forty-six genera of viruses, seventy-one species, and forty-two genera of bacteria. Based on this mapping we constructed a CAMI taxonomic profile (Sczyrba et al. 2017), which was then used as the gold standard in pipeline evaluation.

The mock-virome and bacterial mock-community is composed from nine virus cultures (*Porcine circovirus 2, Feline panleukopenia virus*, *BK virus*, *Pepino Mosaic virus*, *Rotavirus A*, *Feline infectious peritonitis virus, Bovine herpesvirus 1, Dickeya solani LIMEstone bacteriophage, and Acanthamoeba polyphaga mimivirus*) and four bacterial cultures (Conceição-Neto et al. 2015). The NGS library contains 12.4M PE Illumina HiSeq reads.

We compared the performance of Lazypipe on the MetaShot benchmark against Kraken2 (Wood and Salzberg 2014), MetaPhlan2 (Truong et al. 2015), and Centrifuge (Kim et al. 2016). Lazypipe was run with SANSparallel (referred to as *Lazypipe*) and Centrifuge (*Lazipipe-nt*) search engines. Kraken2, MetaPhlan2, and Centrifuge were run with default settings. For Centrifuge we used the NCBI nt database; alignments with <60 nt match were removed to improve precision. Classification results were converted to CAMI taxonomic profiles and evaluated against the golden standard using OPAL (Meyer et al. 2019), a CAMI (Sczyrba et al. 2017) spinoff project implementing CAMI metrics for metagenomic profilers. We also performed the precision–recall analysis using ROCR (Sing et al. 2005). For the simulated metagenome, we separately evaluated the entire taxonomic profile output by each of the pipelines and subprofiles limited to virus taxa.

3. Results

3.1 Excellent recall and precision for both simulated and real datasets

Results for OPAL evaluation on the MetaShot benchmark are available from the project’s website (https://www.helsinki.fi/en/projects/lazypipe).

Precision, recall (syn. sensitivity), and F1-score (harmonic mean of precision and recall) for predicted virus taxa and for all predictions are listed in Tables 1 and 2, respectively. For predicted virus taxa both Lazypipe variants have very high precision and recall at both genus and species level (Table 1). Lazypipe-nt has clearly the best balance between precision and recall, which is reflected in the highest F1-scores among the compared tools (Table 1). In the comparison of all predictions Lazypipe has the highest F1-score at the genus levels. Note that in this evaluation all methods have mediocre performance below the genus level. Lazypipe and Centrifuge are challenged with false positives and MetaPhlan2 and Kraken2 with false negatives (Table 2).

To evaluate classification at different cut-off levels we also performed precision–recall analysis for predicted virus taxa and for all predictions (see Fig. 2A and B). In this evaluation, we observe very high precision values for Lazypipe-nt, Lazypipe, and Centrifuge. A slightly better performance of Lazypipe-nt and Centrifuge towards the end of the prediction list (i.e. high recall

Table 1. Accessing accuracy of virus taxon retrieval by different tools.

Tool	Rank	TP	FP	FN	Pr	Rc	F
Lazypipe-nt	Genus	41	2	4	0.953	0.911	0.932
Lazypipe		41	2	4	0.953	0.911	0.932
Centrifuge		45	8	0	0.849	1.000	0.918
MetaPhlan2		32	4	13	0.889	0.711	0.790
Kraken2		21	1	24	0.955	0.467	0.627
Lazypipe-nt	Species	69	2	15	0.972	0.821	0.890
Lazypipe		72	8	12	0.900	0.857	0.878
Centrifuge		80	47	4	0.630	0.952	0.758
MetaPhlan2		38	7	46	0.844	0.452	0.589
Kraken2		16	1	68	0.941	0.190	0.317

Compared tools are ordered by the descending F1-score for virus taxa predicted for simulated metagenome (Fosso et al. 2017).
TP, true positives; FP, false positives; FN, false negatives; Pr, precision; Rc, recall; F, F1-score.

Table 2. Accessing accuracy of viral and bacterial taxon retrieval by different tools.

Tool	Rank	TP	FP	FN	Pr	Rc	F
Lazypipe	Genus	84	22	4	0.792	0.955	0.866
MetaPhlan2		70	7	18	0.909	0.795	0.848
Lazypipe-nt		64	12	24	0.842	0.727	0.780
Kraken2		50	3	38	0.943	0.568	0.709
Centrifuge		82	162	6	0.336	0.932	0.494
MetaPhlan2	Species	105	10	51	0.913	0.673	0.775
Lazypipe		143	94	13	0.603	0.917	0.728
Lazypipe-nt		100	40	56	0.714	0.641	0.676
Kraken2		52	22	104	0.703	0.333	0.452
Centrifuge		126	471	30	0.211	0.808	0.335

Compared tools are ordered by the descending F1-score for all predictions for simulated metagenome (Fosso et al. 2017).
TP, true positives, FP, false positives, FN, false negatives, Pr, precision, Rc, recall, F, F1-score.

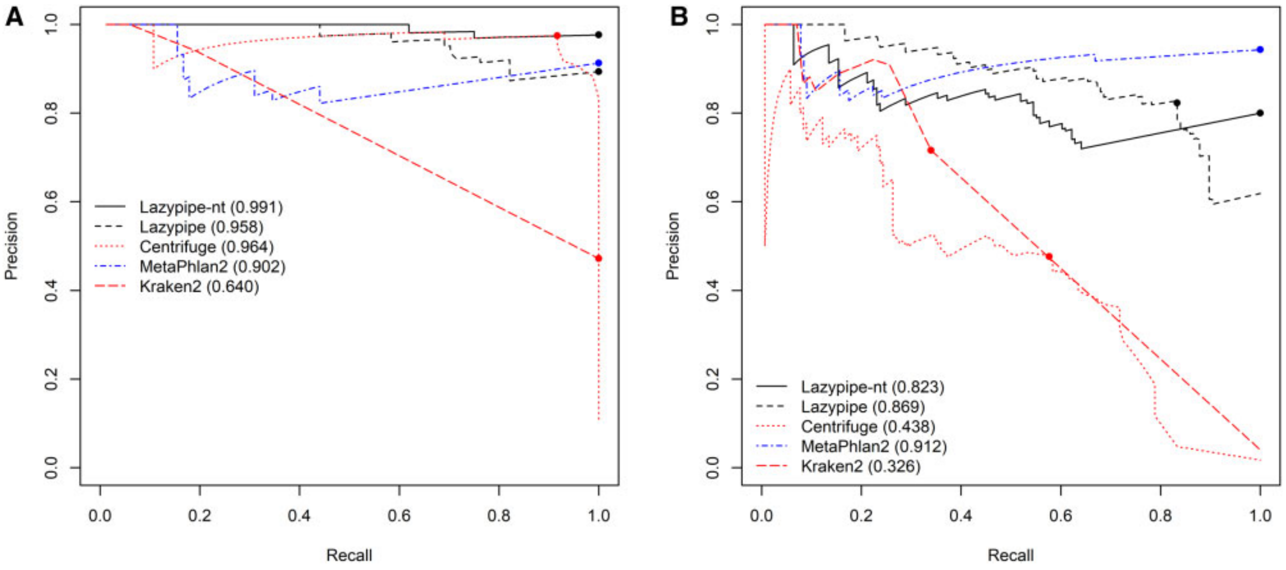

Figure 2. Accessing the classification accuracy with precision–recall analysis. (A) Precision–recall curves for reported virus taxa. (B) Precision–recall curves for all reported taxa. Area under the precision–recall curves is displayed after the tool's name in the figure legend. The dot on each curve corresponds to the maximum F1 value (F_{max}).

values) indicates that nt search may be more efficient than the aa search in retrieving known viral taxa (Fig. 2A). Also, mapping of the assembled contigs (Lazypipe-nt) appears to be more accurate then mapping of shorter read sequences (Centrifuge).

To evaluate the performance of Lazypipe on real data, we ran the Lazypipe analysis with default settings on the mock-community data (for results please see project's webpage). Recovery of the nine mock-community viral taxa was evaluated by manual inspection of Lazypipe summary tables. Lazypipe recovered all seven eukaryotic viruses included in the mock-virome. Moreover, the correct eukaryotic viruses were the only eukaryotic viruses predicted for this data with acceptable confidence scores (Scores 1 and 2; excluding Score 3, which has a high risk of being false positive). Thus, we had 100 per cent sensitivity and 100 per cent precision for the eukaryotic viruses at the species level. Lazypipe also reported the *Dickeya LIMEstone* virus, but did not report the *Acanthamoeba polyphaga mimivirus*.

3.2 Benchmarking time efficiency

We compared the execution time of Lazypipe, Kraken2, MetaPhlan2, and Centrifuge on the MetaShot simulated metagenome on a GNU/Linux machine with sixty-four 2,300 MHz CPUs. All programs were run with sixteen threads. The wall clock time in the order from the fastest to the slowest was: Kraken2 (2 min 30 s), Centrifuge (21 min 43 s), MetaPhlan2 (2 h 21 min 21 s), and Lazypipe (4 h 31 min 51 s). Comparing this order to Tables 1 and 2 we see a trade-off between accuracy and speed. The fastest tools (Kraken2 and Centrifuge) are the least accurate, and the most accurate tools (Lazypipe and MetaPhlan2) are the slowest. Although Lazypipe is about twice as slow as MetaPhlan2, it is more accurate and creates annotated assembly, which is not done by any of the compared tools. We also note that key subprograms employed by Lazypipe (i.e. BWA, Megahit, SANSparallel, and SAMtools) have parallel implementation and are expected to have good scalability.

3.3 Novel virome sequences from mink faecal samples

In addition to the mock-community data we tested the performance of Lazypipe using real data from different sample types (cerebrospinal fluid [CSF], serum, faeces, and tissue samples) derived from various host species.

Since the pipeline is designed also for the detection of unknown viruses, we explored various sources for virus discovery with a by default unknown viral diversity. As an example of searching for the causative agents of veterinary disease, we analyzed sequence data derived from a faecal sample of a mink with gastroenteritis manifesting as diarrhoea. Altogether, Lazypipe detected multiple contigs that indicated the presence of virus genomes (see Table 3). Notably, one contig contained a large open reading frame (ORF) that most likely represents a novel picorna-like virus (order *Picornavirales*) with only 30 per cent aa identity to the closest match. In addition, partial genomes of a toti-like virus with 29–32 per cent aa identity to Beihai toti-like virus 4 (contig length 3,792) and with 38–48 per cent aa identity to Hubei unio douglasiae virus 1 (contig length 3,219) were detected together with smaller fragments of other yet unclassified viruses (see Table 3).

In addition to the above, virus groups with well-known association to the gastrointestinal system were detected. These included members of family *Caliciviridae* and *Parvoviridae*. Of the family *Caliciviridae*, six norovirus and six sapovirus contigs were detected. More thorough examination suggested that the norovirus contigs constitute a complete genome of a new representative of noroviruses with 89 per cent aa identity in ORF1 (nonstructural polyprotein) to norovirus genotypes IV and VI found in cats and dogs (Ford-Siltz et al. 2019), 63 per cent aa identity in ORF2 (VP1) protein to Genotype II found in pigs and 57 per cent aa identity to Genotype II in ORF3 (VP2).

The sapovirus contigs constituted a complete genome with 80–81 per cent aa identity in ORF1 (including VP1 72–73 per cent aa identity) and 45 per cent aa identity in ORF2 (minor capsid protein VP2) to sapovirus GXII previously detected in minks (Guo, Evermann, and Saif 2001; Oka et al. 2016).

Table 3. Virus contigs retrieved by Lazypipe for the mink fecal sample.

Order	Family	Genus	Length (nt)	Closest match	Gene	Identity (%)
Picornavirales			8,990	Kilifi virus		30
NA	Caliciviridae	*Norovirus*	8,006	Norovirus GIV and GVI	ORF1	89
				GII	ORF2	63
					ORF3	57
		Sapovirus	7,511	Sapovirus genotype XII	ORF1	81
					VP1	73
					ORF2	45
NA	Parvoviridae	*Chapparvovirus*	3,069	Chicken chapparvovirus 1	NS	96
				Chicken chapparvovirus 2	VP1	35
		Amdoparvovirus/ Protoparvovirus	2,448	Chiropteran protoparvovirus 1	NS	44
				Carnivore amdoparvovirus 1	VP1	39
Unclassified	Toti-like viruses		3,792	Beihai toti-like virus 4		29-32
			3,219	Hubei unio douglasiae virus 1		38-48
	Bicobirna-like viruses		1,346	Beihai picobirna-like virus 11		81
	Noda-like viruses		1,250	Beihai barnacle virus 11		53
			1,070	Wenzhou noda-like virus 2		46
			857	Wenzhou noda-like virus 2		78
			785	Wenling noda-like virus 1		72
			943	Wuhan pillworm virus 4		42
	Circo-like virus		2,377	uncultured marine virus		34
			919	Anguilla anguilla circovirus		60
			692	Dromedary stool-associated circular ssDNA virus		55
			537	Hermit crab-associated circular genome		52

Displaying contigs exceeding 500 nt in length. Length (nt), contig nt length, Identity (%), aa identity to the closest match.

Table 4. Lazypipe summary for various sample types with known human pathogenic viruses.

Host	Sample type	Genus	Length (nt)	Closest match	Identity (%)
Human	CSF	*Enterovirus*	7,384	Coxsackievirus B5	99
	Serum		7,375	Coxsackievirus A6	100
	Serum	*Parechovirus*	7,321	Human parechovirus 3	99
	Brain (cerebellum)	*Flavivirus*	10,681	TBEV	100
	Nasopharyngeal swab	*Betacoronavirus*	29,806	SARS-coronavirus-2	100
		Mastadenovirus	333–702	Human mastadenovirus C	96–100
I. ricinus	Tick homogenate	*Flavivirus*	11,090	TBEV	99
			2,696–3,014[a]	Alongshan virus	96–99

[a]Segmented genome. Length (nt), contig nt length, Identity (%), aa identity to the closest match.

In addition to these, short low coverage contigs matching to Atlantic salmon calicivirus (78–100% aa identity) were detected. Most likely, these are derived from the feed.

Of the family *Parvoviridae*, the largest contig (3,069 nt) matched to chicken chapparvovirus 2 spanning from 3′ end of the 5′ end of VP1, whereas another large contig (2,448 nt) contained 3′ end of NS protein with 44 per cent aa similarity to Chiropterian protoparvovirus and the 5′ end of VP1 protein with 39 per cent aa identity to Aleutian mink disease virus (amdoparvovirus). In addition to these, small fragments of mink bocaparvovirus were detected.

3.4 Human clinical samples

As an example of testing the suitability of Lazypipe for human clinical samples and exploring its use for detection of viral pathogens in humans, we used sequence data derived from human CSF, serum, brain tissue and nasopharyngeal swab samples that were previously tested positive for entero-, entero/parecho-, tick-borne encephalitis, and SARS-coronavirus-2 viruses, respectively (Table 4). From the CSF sample, a complete genome with 99 per cent sequence identity with Coxsacievirus B5 (a member of Enterovirus B species) strains AU17EV1 and AU17EV2 (Queensland, Australia; Huang et al. 2017) was retrieved. From the two serum samples complete genomes of Coxsackievirus A6 (Enterovirus A species) and Human Parechovirus 3 (Parechovirus A species) were retrieved. From the cerebellum sample a complete genome of TBEV was retrieved. From the nasopharyngeal swab sample originated from the first case of Coronavirus disease 2019 (COVID-19) in Finland (Haveri et al. 2020), a nearly complete SARS-coronavirus-2 (SARS-CoV-2) genome and fragments of *Human mastadenovirus C* sequences were retrieved (see Table 4).

3.5 Arthropod samples

We also analyzed samples of arthropod vectors. From an *Ixodes ricinus* tick sample collected from the Kotka archipelago in 2011, complete genomes of both Siberian subtype TBEV and a novel Alongshan virus (Kuivanen et al. 2019) were obtained (Table 4).

3.6 Sars-CoV-2 Patient Samples from China

We analyzed public Illumina HiSeq/MiSeq libraries sequenced from bronchoalveolar lavage fluid from five patients with pneumonia at the early stage of the COVID-19 outbreak in Wuhan, China. Nine public NGS libraries were collected from NCBI SRA database (BioProject PRJNA605983) and analyzed with Lazypipe. By applying default settings, we intentionally recreated a scenario, in which NGS data from SARS patients would be analyzed prior to identifying the causative agent. *SARS-CoV* was identified by Lazypipe in all patients and in eight out of nine NGS libraries (Table 5). Lazypipe also identified co-infection with Influenza A in two out of five patients (Table 5).

4. Discussion

The availability of robust bioinformatics pipelines for viral metagenomics continues to be one of the critical steps in many research projects. Many of the existing pipelines are hindered by one or several limitations including large locally installed reference databases, slow homology search engines employed, low sensitivity for novel divergent sequences, low precision/recall performance for viral taxa or the lack of benchmarking for viral taxon retrieval, and the lack of assembling and contig annotation steps in the analysis. These limitations slow down the use of the unbiased sequencing approaches for rapid detection of novel emerging viruses.

In this publication we present Lazypipe, a novel bioinformatics pipeline that addresses the limitations typically encountered in viral metagenomics. Lazypipe avoids installation of large reference databases by delegating homology search to an external server. This frees the user from the need to install, index and update local reference databases, which can pose serious technical and resource constrains due to the sheer size of the modern sequence databases. By using SANSparallel (Somervuo and Holm 2015) we also make Lazypipe considerably faster than pipelines based on BLASTP, and, simultaneously, render Lazypipe sensitive to highly divergent sequences, because viral peptides tend to be more conservative than nucleotide sequences. Other fast alternatives to BLASTP, such as DIAMOND (Buchfink, Xie, and Huson 2015), show comparable sensitivity and speed (Medlar and Holm 2018), but unlike SANSparallel, require installation of local databases.

Taxonomic profiling by Lazypipe is done by querying assembled contigs instead of the reads, which translates into highly accurate taxonomic profiling of viral taxa. Benchmarking on simulated data showed that Lazypipe was clearly the most accurate taxonomic profiler for viral taxa among the four software packages compared. Testing on real mock community data demonstrated precision and recall nearing 100 per cent for eukaryotic viruses. The detection of multiple novel viruses from various environmental and clinical samples reported here and in previous studies that used Lazypipe analysis (Forbes et al. 2019; Kuivanen et al. 2019) demonstrates that Lazypipe is also well suited for the detection and characterization of novel and highly divergent viral

Table 5. Lazypipe summary for SARS-CoV-2 clinical samples from Wuhan, China.

Accession	Library	Virus	Taxid	Readn	Readn%	Csum	Contign	Length (nt)
SRR11092058	WIV02	SARS-related coronavirus	694009	36	0.517	3	9	362–568
SRR11092063	WIV02-2	SARS-related coronavirus	694009	559	0.368	1	23	305–2,169
SRR11092057	WIV04	SARS-related coronavirus	694009	732	13.088	1	15	393–4,488
SRR11092062	WIV04-2	SARS-related coronavirus	694009	5,918	3.003	1	1	29,850
SRR11092062	WIV04-2	Influenza A virus	11320	274	0.139	1	2	1,065–4,609
SRR11092062	WIV04-2	Autographa californica multiple nucleopolyhedrovirus	307456	205	0.104	1	2	397–440
SRR11092061	WIV05	SARS-related coronavirus	694009	234	0.051	1	20	315–2,044
SRR11092061	WIV05	Saccharomyces 20S RNA narnavirus	186772	135	0.029	2	1	2,378
SRR11092060	WIV06-2	SARS-related coronavirus	694009	525	0.142	1	22	305–2,530
SRR11092060	WIV06-2	Spodoptera frugiperda rhabdovirus	1481139	165	0.045	1	1	468
SRR11092060	WIV06-2	Saccharomyces 20S RNA narnavirus	186772	103	0.028	2	3	439–1,423
SRR11092064	WIV07	Influenza A virus	11320	9,063	0.097	1	4	399–4,765
SRR11092064	WIV07	Saccharomyces 20S RNA narnavirus	186772	3,386	0.036	1	1	2,440
SRR11092064	WIV07	SARS-related coronavirus	694009	819	0.009	2	16	408–5,727
SRR11092064	WIV07	Bamboo mosaic virus	35286	325	0.003	2	1	355
SRR11092064	WIV07	Spodoptera frugiperda rhabdovirus	1481139	168	0.002	2	1	442
SRR11092059	WIV07-2	Saccharomyces 20S RNA narnavirus	186772	1,693	0.019	2	1	2,440
SRR11092059	WIV07-2	SARS-related coronavirus	694009	467	0.005	2	5	1,531–5,727

Viral taxa identified by Lazypipe from public Illumina libraries sequenced from five patients with pneumonia at the early stage of the COVID-19 outbreak in Wuhan, China. Lazypipe correctly identified SARS-CoV in eight out of nine samples. Additionally, two samples were identified with Influenza A and one sample with human mastadenovirus C coinfection. Accession, NCBI SRA accession, Library, library identifier, Virus, name of the viral taxon, Readn, reads assigned to the taxon, Contign, contigs assigned to the taxon, Csum, csum confidence score (see text), Length (nt), contig nt length.

genomes. Reflecting on the SARS-CoV-2 pandemic situation (April 2020) we tested SARS-CoV-2 positive Illumina libraries with Lazypipe and confirmed that the pipeline detected SARS-CoV in nine out of ten libraries with default settings and without SARS-CoV-2 reference genome. This demonstrates the utility of Lazypipe for scenarios in which novel zoonotic viral agents emerge and can be quickly detected by NGS sequencing from clinical samples.

Previously, we have published two examples of novel and potentially zoonotic viral agents that were identified with Lazypipe from wild animals that can serve as vectors. A new ebolavirus was identified from faeces and organ samples of *Mops condylurus* bats in Kenya (Forbes et al. 2019), and a new tick-borne pathogen Alongshan virus from ticks in Northeast Europe (Kuivanen et al. 2019). These examples demonstrate the efficacy of Lazypipe data analysis for NGS libraries with very different DNA/RNA backgrounds, ranging from mammalian tissues to pooled and crushed arthropods.

The current pandemic highlights the need for an efficient and unbiased way to screen 1, for previously unknown viruses from either wildlife and arthropods for potential viral diversity that may emerge as human or animal pathogens, or 2, from individuals or human populations, production animals or companion animals manifesting with a disease of unknown aetiology for previously unknown or atypical causative agents. We showed here that Lazypipe can contribute to both of these important efforts and that it was able to detect the causative agent of the current pandemic without prior information.

Data availability

Lazypipe user manual, analyzed data and other resources are hosted at the project's website (https://www.helsinki.fi/en/projects/lazypipe). Lazypipe source code is freely available from git repository (https://bitbucket.org/plyusnin/lazypipe/).

Funding

This work was supported by the VEO - European Union's Horizon 2020 [grant number 874735]; the Academy of Finland [grant numbers 316264 and 329323]; Helsinki University Hospital Funds [grant numbers TYH2018322 and M1023TK001]; and the Jane and Aatos Erkko Foundation. Funding for open access charge: VEO - European Union's Horizon 2020 [grant number 874735].

Conflict of interest

None declared.

References

Bhuvaneshwar, K. et al. (2018) 'viGEN: An Open Source Pipeline for the Detection and Quantification of Viral RNA in Human Tumors', *Frontiers in Microbiology*, 9: 1172.

Biedermann, L., and Rogler, G. (2015) 'The Intestinal Microbiota: Its Role in Health and Disease', *European Journal of Pediatrics*, 174: 151–67.

Bolger, A. M., Lohse, M., and Usadel, B. (2014) 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data', *Bioinformatics*, 30: 2114–20.

Buchfink, B., Xie, C., and Huson, D. H. (2015) 'Fast and Sensitive Protein Alignment Using DIAMOND', *Nature Methods*, 12: 59–60.

Cantalupo, P. G. et al. (2011) 'Raw Sewage Harbors Diverse Viral Populations', *mBio*, 2: e00180–11.

Chen, S. et al. (2018) 'Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor', *Bioinformatics*, 34: i884–90.

Cock, P. J. et al. (2010) 'The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants', *Nucleic Acids Research*, 38: 1767–71.

Conceição-Neto, N. et al. (2015) 'Modular Approach to Customise Sample Preparation Procedures for Viral Metagenomics: A

Reproducible Protocol for Virome Analysis', *Scientific Reports*, 5: 16532.

Forbes, K. M. et al. (2019) 'Bombali Virus in *Mops condylurus* Bat, Kenya', *Emerging Infectious Diseases*, 25: 955–7.

Ford-Siltz, L. A. et al. (2019) 'Genomics Analyses of GIV and GVI Noroviruses Reveal the Distinct Clustering of Human and Animal Viruses', *Viruses*, 11: 204.

Fosso, B. et al. (2017) 'MetaShot: An Accurate Workflow for Taxon Classification of Host-Associated Microbiome from Shotgun Metagenomic Data', *Bioinformatics (Oxford, England)*, 33: 1730–2.

Garretto, A., Hatzopoulos, T., and Putonti, C. (2019) 'virMine: Automated Detection of Viral Sequences from Complex Metagenomic Samples', *PeerJ*, 7: e6695.

Geoghegan, J. L. and Holmes, E. C. (2017) 'Predicting Virus Emergence amid Evolutionary Noise', *Open Biology*, 7: 170189.

Graf, E. H. et al. (2016) 'Unbiased Detection of Respiratory Viruses by Use of RNA Sequencing-Based Metagenomics: A Systematic Comparison to a Commercial PCR Panel', *Journal of Clinical Microbiology*, 54: 1000–7.

Guo, M., Evermann, J. F., and Saif, L. J. (2001) 'Detection and Molecular Characterization of Cultivable Caliciviruses from Clinically Normal Mink and Enteric Caliciviruses Associated with Diarrhea in Mink', *Archives of Virology*, 146: 479–93.

Haveri, A. et al. (2020) 'Serological and Molecular Findings during SARS-CoV-2 Infection: The First Case Study in Finland, January to February 2020', *Eurosurveillance*, 25: 2000266.

Huang, B. et al. (2017) 'Genome Sequences of Coxsackievirus B5 Isolates from Two Children with Meningitis in Australia', *Genome Announcements*, 5: e01125–17.

Kataoka, K. (2016) 'The Intestinal Microbiota and Its Role in Human Health and Disease', *The Journal of Medical Investigation*, 63: 27–37.

Kim, D. et al. (2016) 'Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences', *Genome Research*, 26: 1721–9.

Kostic, A. D. et al. (2011) 'PathSeq: Software to Identify or Discover Microbes by Deep Sequencing of Human Tissue', *Nature Biotechnology*, 29: 393–6.

Kuivanen, S. et al. (2019) 'Detection of Novel Tick-Borne Pathogen, Alongshan Virus, in *Ixodes ricinus* Ticks, South-Eastern Finland, 2019', *Eurosurveillance*, 24: 1900394

Li, D. et al. (2015) 'MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph', *Bioinformatics*, 31: 1674–6.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Prepr. ArXiv1303.3997*.

——, et al.; 1000 Genome Project Data Processing Subgroup. (2009) 'The Sequence Alignment/Map Format and SAMtools', *Bioinformatics*, 25: 2078–9.

Li, Y. et al. (2016) 'VIP: An Integrated Pipeline for Metagenomics of Virus Identification and Discovery', *Scientific Reports*, 6: 23774.

Lim, E. S. et al. (2015) 'Early Life Dynamics of the Human Gut Virome and Bacterial Microbiome in Infants', *Nature Medicine*, 21: 1228–34.

Lorenzi, H. A. et al. (2011) 'TheViral MetaGenome Annotation Pipeline (VMGAP):an Automated Tool for the Functional Annotation of Viral Metagenomic Shotgun Sequencing Data', *Standards in Genomic Sciences*, 4: 418–29.

Mäki-Tanila, A. V. et al. (2016) Proceedings of the XIth International Scientific Congress in Fur Animal Production. Scientifur, no. 3/4, vol. 40, International Fur Animal Scientific Association.

Medlar, A., and Holm, L. (2018) 'TOPAZ: Asymmetric Suffix Array Neighbourhood Search for Massive Protein Databases', *BMC Bioinformatics*, 19: 278.

Meyer, F. et al. (2019) 'Assessing Taxonomic Metagenome Profilers with OPAL', *Genome Biology*, 20: 51.

—— et al. (2008) 'The Metagenomics RAST Server - A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes', *BMC Bioinformatics*, 9: 386.

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012) 'Metagenomics and Future Perspectives in Virus Discovery', *Current Opinion in Virology*, 2: 63–77.

Naccache, S. N. et al. (2014) 'A Cloud-Compatible Bioinformatics Pipeline for Ultrarapid Pathogen Identification from Next-Generation Sequencing of Clinical Samples', *Genome Research*, 24: 1180–92.

Naeem, R., Rashid, M., and Pain, A. (2013) 'READSCAN: A Fast and Scalable Pathogen Discovery Program with Accurate Genome Relative Abundance Estimation', *Bioinformatics*, 29: 391–2.

Neil, J. A., and Cadwell, K. (2018) 'The Intestinal Virome and Immunity', *The Journal of Immunology*, 201: 1615–24.

Noguchi, H., Taniguchi, T., and Itoh T. (2008) 'MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes', *DNA Research: An International Journal Rapid Publication of Reports on Genes and Genomes*, 15: 387–96.

Norman, J. M. et al. (2015) 'Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease', *Cell*, 160: 447–60.

Oka, T. et al. (2016) 'Genetic Characterization and Classification of Human and Animal Sapoviruses', *PLoS One*, 11: e0156373.

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011) 'Interactive Metagenomic Visualization in a Web Browser', *BMC Bioinformatics*, 12: 385.

Pallen, M. J. (2014) 'Diagnostic Metagenomics: Potential Applications to Bacterial, Viral and Parasitic Infections', *Parasitology*, 141: 1856–62.

Rose, R. et al. (2016) 'Challenges in the Analysis of Viral Metagenomes', *Virus Evolution*, 2: vew022.

Roux, S. et al. (2014) 'Metavir 2: New Tools for Viral Metagenome Comparison and Assembled Virome Analysis', *BMC Bioinformatics*, 15: 76.

Sczyrba, A. et al. (2017) 'Critical Assessment of Metagenome Interpretation—A Benchmark of Metagenomics Software', *Nature Methods*, 14: 1063–71.

Sing, T. et al. (2005) 'ROCR: Visualizing Classifier Performance in R', *Bioinformatics*, 21: 3940–1.

Smits, S. L. et al. (2015) 'Recovering Full-Length Viral Genomes from Metagenomes', *Frontiers in Microbiology*, 6: 1069.

Smura, T. et al. (2016) 'Fecal Microbiota of Healthy and Diarrheic Farmed Arctic Foxes (*Vulpes lagopus*) and American Mink (*Neovison vison*)–A Case-Control Study', in *XIth International Scientific Congress in Fur Animal Production*, p. 17. Scientifur, no. 3/4, vol. 40, International Fur Animal Scientific Association.

Somervuo, P., and Holm, L. (2015) 'SANSparallel: Interactive Homology Search against Uniprot', *Nucleic Acids Research*, 43: W24–29.

Stajich, J. E. (2002) 'The Bioperl Toolkit: Perl Modules for the Life Sciences', *Genome Research*, 12: 1611–8.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov J. P. (2013) 'Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration', *Briefings in Bioinformatics*, 14: 178–92.

Truong, D. T. et al. (2015) 'MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling', *Nature Methods*, 12: 902–3.

Vilsker, M. et al. (2019) 'Genome Detective: An Automated System for Virus Identification from High-Throughput Sequencing Data', *Bioinformatics*, 35: 871–3.

Wang, Q., Jia, P., and Zhongming, Z. (2013) 'VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data', *PLoS One*, 8: e64465.

Wommack, K. E. et al. (2012) 'VIROME: A Standard Operating Procedure for Analysis of Viral Metagenome Sequences', *Standards in Genomic Sciences*, 6: 427–39.

Wood, D. E., Lu J., and Langmead B. (2019) 'Improved Metagenomic Analysis with Kraken 2', *Genome Biology*, 20: 257.

——, and Salzberg, S. L. (2014) 'Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments', *Genome Biology*, 15: R46.

Zerbino D. R., and Birney E. (2008) 'Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs', *Genome Research*, 18: 821–9.

Zhao, G. et al. (2017) 'VirusSeeker, a Computational Pipeline for Virus Discovery and Virome Composition Analysis', *Virology*, 503: 21–30.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010) 'Ab Initio Gene Identification in Metagenomic Sequences', *Nucleic Acids Research*, 38: e132.