Infection, Genetics and Evolution 66 (2018) 180–187

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Virus detection in high-throughput sequencing data without a reference genome of the host

Jochen Kruppa[a,*], Wendy K. Jo[b], Erhard van der Vries[c], Martin Ludlow[b], Albert Osterhaus[b], Wolfgang Baumgaertner[d], Klaus Jung[a]

[a] *Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17p, Hannover 30559, Germany*
[b] *Research Center for Emerging Infections and Zoonoses, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17, Hannover 30559, Germany*
[c] *Department of Immunity and Infection, University of Utrecht, Yalelaan 1, Utrecht, CN 3584, the Netherlands*
[d] *Department of Pathology, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17, Hannover 30559, Germany*

ARTICLE INFO

Keywords:
Decoy database
Metagenomics
Read mapping
Reference genome
Virus discovery

ABSTRACT

Discovery of novel viruses in host samples is a multidisciplinary process which relies increasingly on next-generation sequencing (NGS) followed by computational analysis. A crucial step in this analysis is to separate host sequence reads from the sequence reads of the virus to be discovered. This becomes especially difficult if no reference genome of the host is available. Furthermore, if the total number of viral reads in a sample is low, de novo assembly of a virus which is a requirement for most existing pipelines is hard to realize.

We present a new modular, computational pipeline for discovery of novel viruses in host samples. While existing pipelines rely on the availability of the hosts reference genome for filtering sequence reads, our new pipeline can also cope with cases for which no reference genome is available. As a further novelty of our method a decoy module is used to assess false classification rates in the discovery process. Additionally, viruses with a low read coverage can be identified and visually reviewed. We validate our pipeline on simulated data as well as two experimental samples with known virus content. For the experimental samples, we were able to reproduce the laboratory findings.

Our newly developed pipeline is applicable for virus detection in a wide range of host species. The three modules we present can either be incorporated individually in other pipelines or be used as a stand-alone pipeline. We are the first to present a decoy approach within a virus detection pipeline that can be used to assess error rates so that the quality of the final result can be judged. We provide an implementation of our modules via Github. However, the principle of the modules can easily be re-implemented by other researchers.

1. Introduction

Samples from humans and animals suspected of a virus infection on clinical grounds, are usually analyzed by classical and modern molecular virological assays, when applicable supported by histo-pathology data. Meanwhile, the advent of next-generation sequencing (NGS) has provided us with the opportunity of reading all sequence information in a biological sample, therefore becoming an important tool for virus discovery which will undoubtedly find its way into routine virus diagnostic practice. However, virus detectionusing NGS data is by no means a straightforward task, but should involve close communication between the clinician, the virologist, the pathologist and the bioinformatician (Smits et al., 2015; Smits and Osterhaus, 2013).

The overall problem of virus identification in NGS data from a host sample is to identify all sequences that don't originate from the host itself. While most sequencing reads will usually belong to the host or other non-relevant microorganisms only a small proportion of reads will belong to the virus to be discovered. The assignment of sequencing reads to the host and other non-relevant organisms and viruses relies on reference genomes available in databases. Here, we present a new bioinformatics pipeline for virus metagenomics that is also applicable if no reference genome data from the host is available.

Currently available bioinformatics pipelines or software solutions for virus sequence detection in NGS data rely on approaches that can be divided into two categories. Category I involve approaches, that first remove all host reads from the sample and map or align the remaining reads to a viral database. In this case, the host's reference genome sequence must be available in a sufficient quality to make sure that all

[*] Corresponding author.
E-mail address: jochen.kruppa@charite.de (J. Kruppa).

https://doi.org/10.1016/j.meegid.2018.09.026
Received 2 August 2018; Received in revised form 25 September 2018; Accepted 27 September 2018
Available online 03 October 2018

host reads are removed and only non-host sequences are among the unmapped reads. Such approaches work for example well with samples from humans or mice and other species for which reference genome data is available at the Ensemble database (Ensemble Database, n.d.). Approaches from the category II first assemble the raw sequencing reads (or only the unmapped reads) to larger contigs, which are further used in the analysis pipeline. Larger contigs allow for a higher mapping accuracy than short reads, and including an assembly step into a detection pipeline is therefore advantageous. Nevertheless, to achieve large contigs - that are longer than the single reads - the coverage of the single viral strains of the sequencing reads must be high. If there are not enough viral reads of a single strain in the sample, the gaps between the reads are too large and contigs cannot be built preventing virus identification. A common element of the approaches in both categories is that reads or contigs are aligned to a given virus sequence database, and a sorted list of detected viruses (or at least taxonomic groups) is returned. The approach we present here belongs to category I, i.e. raw reads instead of contigs are mapped against reference genomes. In the following, we provide a brief summary of other existing pipelines and their usability in the case of samples generated with low sequencing coverage and a non-availability of a host reference.

Among category II pipelines, Iterative Virus Assembler (IVA) (Hunt et al., 2015) uses its own de novo assembler to generate contigs from the raw or host-free reads. Generated contigs can afterwards be mapped to a virus database using SMALT (SMALT, n.d.) and Kraken (Wood and Salzberg, 2014) to determine the viral strain. IVA reports only the virus strain that appears most frequently with quality information, whereas the report produced by Kraken gives the user more information on the identified taxa. Thus, the limitaion of IVA consists on de novo assembly of contigs, which is not possible when the overall sequence coverage is low and the generation of only a single viral strain. Another pipeline, called RIEMS (Scheuch et al., 2015), also first assembles the raw reads to contigs, which are afterwards mapped to a virus database using the NCBI BLAST software suite (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+). Assigned reads are then classified taxonomicaly. As an additional feature, the RIEMS pipeline can also translate the assembled sequences to amino acid sequences and use these sequences for further detection on the protein level.

Among category I pipelines (i.e., direct mapping of sequencing reads), the approach by Petty et al. (Petty et al., 2014) describes the standard procedure in a human pilot study. The raw reads are first mapped to the human reference genome, and the unmapped reads are then mapped to a virus database. The removal of the host reads is a crucial step and has been implemented in VirusFinder (Wang et al., 2013), VirusHunter (Zhao et al., 2013), VirusSeq (Chen et al., 2012), and Vy-PER (Forster et al., 2015). Mostly, these pipelines have been demonstrated on example of human samples. In general, these pipelines demand for a known host reference genome of sufficient quality to remove all non-viral reads from the downstream analysis. First, host reads are removed to obtain data cleaned from host sequences. If host reads are kept in the data, false positive mapping to virus reference sequences may occur. Second, further noise is removed from the data to improve the mapping accuracy. By removing the host reads it is assumed that only viral reads remain. Nielsen et al. (Nielsen et al., 2014) describe a reference free identification and assembly without the implementation as a software solution or ready to use pipeline. In the following the mentioned pipelines of category I are described in more detail.

VirusFinder (Wang et al., 2013) performs first a preprocessing step, in which the raw sequencing reads are mapped to the human reference genome using Bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2009). Then, unmapped reads are extracted and aligned to a viral database using BLAT (Kent, 2002). Finally, the reads are assembled using Trinity (Grabherr et al., 2011). VirusFinder assumes a high sequencing coverage so that good assembly results can be obtained, and is mainly developed to detect virus integrations sites in the human genome. The examples presented in the VirusFinder article have a sequencing coverage ranging from 31.7× to 121.2×. The assembled contigs are then used for the generation of phylogenetic trees and the estimation of relationship to each other. VirusHunter (Zhao et al., 2013) uses BLASTn to filter first the reads belonging to the host after some quality assessment. The host-free reads are then classified using BLASTn and BLASTx into taxonomic groups. Therefore, VirusHunter needs a good host reference genome to filter the reads into host-free and host reads. In addition, the repeated BLAST runs to process the reads are also time consumable. VirusSeq (Chen et al., 2012) focuses on the identification of viral strains in human cancer tissue. First all human sequence reads are removed by mapping to the human reference genome. The remaining human-free reads are then aligned to a viral reference database using the MOSAIK aligner in both steps (Lee et al., 2014). VirusSeq uses the overall count number of matched reads to identify the viral strain. Nevertheless, the threshold is set to 1.000 reads per virus regarding an overall 30× coverage of the whole-genome sequencing data. This threshold can be modified, but VirusSeq is developed for sample with a high read coverage. Therefore, it cannot be used to analyze low coverage datasets. Vy-PER (Forster et al., 2015) uses in the first step the human reference genome to remove all host sequence reads. Reads, which are not mapped to the human reference are then filtered and aligned to the NCBI viral genome database using BLAT (Kent, 2002). The described example on leukemia samples is done with a very high coverage (80× cases and 40× controls), which is not a requirement, but precondition for the elimination of false positives.

All mentioned pipelines of category II make use of an alignment or a mapping software such as Blast or Bowtie2. A broader and more comprehensive overview on available mappers is given by Fonseca et al. (Fonseca et al., 2012), in which the authors discuss the mapper characteristics and the problems of comparing different mappers. In the case of virus detection some specific issues play a role: 1) high heterogeneity of the genomes, 2) mutation rates, 3) insertions of whole genomic areas, and 4) infection of new hosts with adaptions of the viral genome. Hence, problems occur when dealing with samples from a high variety of potential virus infected species. First, a fully assembled reference genome is only available for a small number of animals. Furthermore, the quality of the reference genomes can differ as only the human and mouse genome are of sufficient quality, whereas there is no good reference genome available for many animals. Second, the number of viral sequence reads in a biological sample depends on the production circle of the virus, the time point of infection, and the selection of the correct tissue type to get most of the virus out of the sample. Therefore, the number of possible detectable viral sequences might be low. For building contigs by an assembly process many viral sequence reads must be available, i.e. a good coverage of viral reads must be given, and these reads should not be contaminated with sequences from the host organism. Both issues are not applicable to studies which don't focus on human or mice samples with a small area of possible viral infection.

In this study, three bioinformatics modules for virus detection and two example data samples are described. Module I allows to evaluate the false positive findings by a decoy database approach, module II shows the host-free mapping of DNA sequencing reads to an artificial viral reference genome, and module III describes the mapping of the translated DNA sequence reads to a artificial amino acid viral reference genome. In the results section we demonstrate the results of a simulation study using the decoy database (comparing different mapping softwares within our pipeline) and present the analysis of two example data sets. We close this article with some conclusions. Moreover, we describe the combination of all three modules into a virus detection pipeline in the supplementary material.

2. Methods

In this section, we describe in detail modules that form our new virus detection pipeline. We chose a modular composition of our pipeline, so that individual modules can also be incorporated in other existing pipelines. Module I uses a DNA mapper to detect virus species. To run the mapper, an artificial genome is built in which all virus sequences available in a database are stringed together as a fasta-file. Module II also uses an artificial genome based on the translation of DNA virus sequences into amino acid sequences. Module III allows the estimation of false positive rates by employing a decoy sequence database. In addition, we describe visualization tools for the examination of the mapping results of modules I and II in the supplementary material. The raw information on the mapped reads is not sufficient to decide if a host is infected by a viral strain. Therefore, we visualize our findings (Tukey, 1977) and determine quality values for the judgment of a relevant viral detection (Supplementary Fig. 1).

2.1. Module I: decoy database

The decoy database works independent from the mapping algorithm and can be used for the evaluation of a DNA or RNA mapping run in a virus detection pipeline. After the sequence read mapping, the question arises on how correct the single sequence reads were mapped, particularly because Bowtie2 and Star, as an example, allow reads to map multiple times to the reference genome. Therefore, we adopt an approach by Reidegeld et al. (2008) (Reidegeld et al., 2008), who proposed a decoy database to determine the false discovery rate in automated protein identification. Three decoy strategies were proposed: 1) 'reverse', meaning that the original reference genome is reversed in the sequence order, 2) 'shuffle', where the original reference genome is shuffled randomly to get another base pair order, 3) 'random', where the protein mass was hold constant. The last one is not possible in the case of DNA sequences. The reverse decoy strategy will not work, because the mappers are too sensitive. Thus, we shuffled the reference genome to obtain a decoy reference database. The shuffling was done by setting different k-mer distributions. Now, the question arises, which k-mer distribution should be shuffled? The shuffling with $k = 1$, i.e. only shuffling the bases A, C, G, and T, will deliver no hit to the decoy reference because the mappers are too sensitive. Hence, we used uShuffle (Jiang et al., 2008) to test different shuffled k-mers to find the best k for the shuffled database. The program uShuffle allows a given genome and keep the specified k-mer distribution fix. However, uShuffle can only handle sequences of length $9 \cdot 10^6$. Therefore, we built blocks of this limited size and shuffled them accordingly to each $k = \{10,\ldots,20\}$.

Fig. 1 shows the setting for the decoy database. First, the artificial viral genome is duplicated and then shuffled with a given k, while the k-mer distribution of the artificial reference is kept fix. Second, from a given virus sequence, 50 paired reads of different lengths ($r_l = \{75,150,300\}$) are drawn and combined with 5.000 paired reads drawn from the decoy database with the same read length. Overall, 100 viral and 10.000 decoy reads were combined into one.fastq-file. The quality of each base was set to $Q = 40$. In the third step the reads were mapped back to the artificial genome of true viral sequences and to the decoy reference genome. In this whole process, a read can be mapped to six possible regions, and the mapping result can be summarized in a 2 × 3 table. A true virus read can either be mapped to the correct virus, to a false virus, or to the decoy database. The same can happen to a decoy read. We found that case (d), in which the decoy read maps to the true virus is very rare. In the simulation study below, we can determine the quality of the three mappers given a shuffle k. To summarize the section, the first aim is to determine the best k for the shuffling process to generate the database and second to use the generated decoy database as a measure of goodness for the viral detection.

Table 1
Contingency table of the possible outcomes by the decoy approach.

		Reference			
		True virus	False virus	Decoy	
Read	Virus	a	b	c	$n_{1.}$
	Decoy	d	e	f	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$n_{.3}$	n

From the mapping results of the simulation we generated a 2 × 3 contingency table (Table 1). For the evaluation of the simulation results we define four classification statistics. First, the true positive rate for the virus reads by $tpr_v = a/(a + b + c)$, the true positive rate for the decoy reads by $tpr_d = f/(d + e + f)$, the overall true positive rate by $tpr = \frac{1}{2}(tpr_d + tpr_v)$, and the false positive rate by $fpr = (b + c + d + e)/n$. We found for all three tested mapper a increase of the false positive rate (fpr) if the shuffling k is increased. The true positive rate is decreasing with an increase of k for all mapper. The true positive rate for the viral reads has a higher variability than the true positive rate for the decoy reads. A longer read length increases the overall true positive rate as expected (Fig. 2).

To determine the quality of the DNA mapping of a biological sample, we generated a decoy database with the same length as the artificial viral genome using a fixed k of 15. We decided to use $k = 15$ after consulting the evaluation results presented in Fig. 2. For $k > 15$ the classification rate is dropping and therefore we use this $k = 15$ for the analysis of the biological samples. It must be mentioned that the value of k depends on the particular virus database being used. Thus, a different value might be appropriate when using other database settings. The shuffling of the artificial genome could not be done in one step, therefore we did the shuffling by 396 artificial chromosomes each of the length of roughly 9 million bp. While running the pipeline, we draw $n_{rd} = 1000$ decoy reads from one random chromosome and added these decoy reads to the sample reads. After the DNA mapping we could determine, if all decoy reads are mapped back to the decoy reference: the true positive rate for the decoy reads (tpr_d) and how many reads are mapped to the wrong reference: the false positive rate (fpr). In the case of a biological sample, we can not distinguish between the true virus (a, d) and the false virus (b, e). Therefore, the contingency table is reduced to a 2 × 2 table. Finally, we can estimate the multi mapping rate $mmr = (d + e + f)/n_{rd}$. Hence, the mmr is the number of mapped decoy reads ($d + e + f$) divided by the predefined number of generated decoy

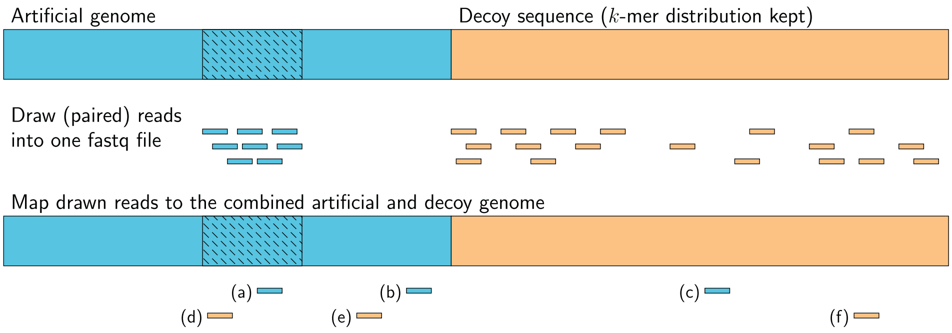
Artificial genome
Decoy sequence (k-mer distribution kept)
Draw (paired) reads into one fastq file
Map drawn reads to the combined artificial and decoy genome
(a)
(b)
(c)
(d)
(e)
(f)

Fig. 1. Generation principle of the decoy framework. First, the artificial genome of the viral strains is duplicated and then shuffled. During the shuffling the original k-mer distribution of the true genome is kept. Afterwards, paired reads from one virus (shaded area) and random decoy reads are drawn. Finally, the drawn reads are mapped to the combined reference genome consisting of the original viral sequences and the shuffled ones. Each single read can be mapped to three areas: i) to the original virus, ii) to a different virus or iii) to the decoy genome.

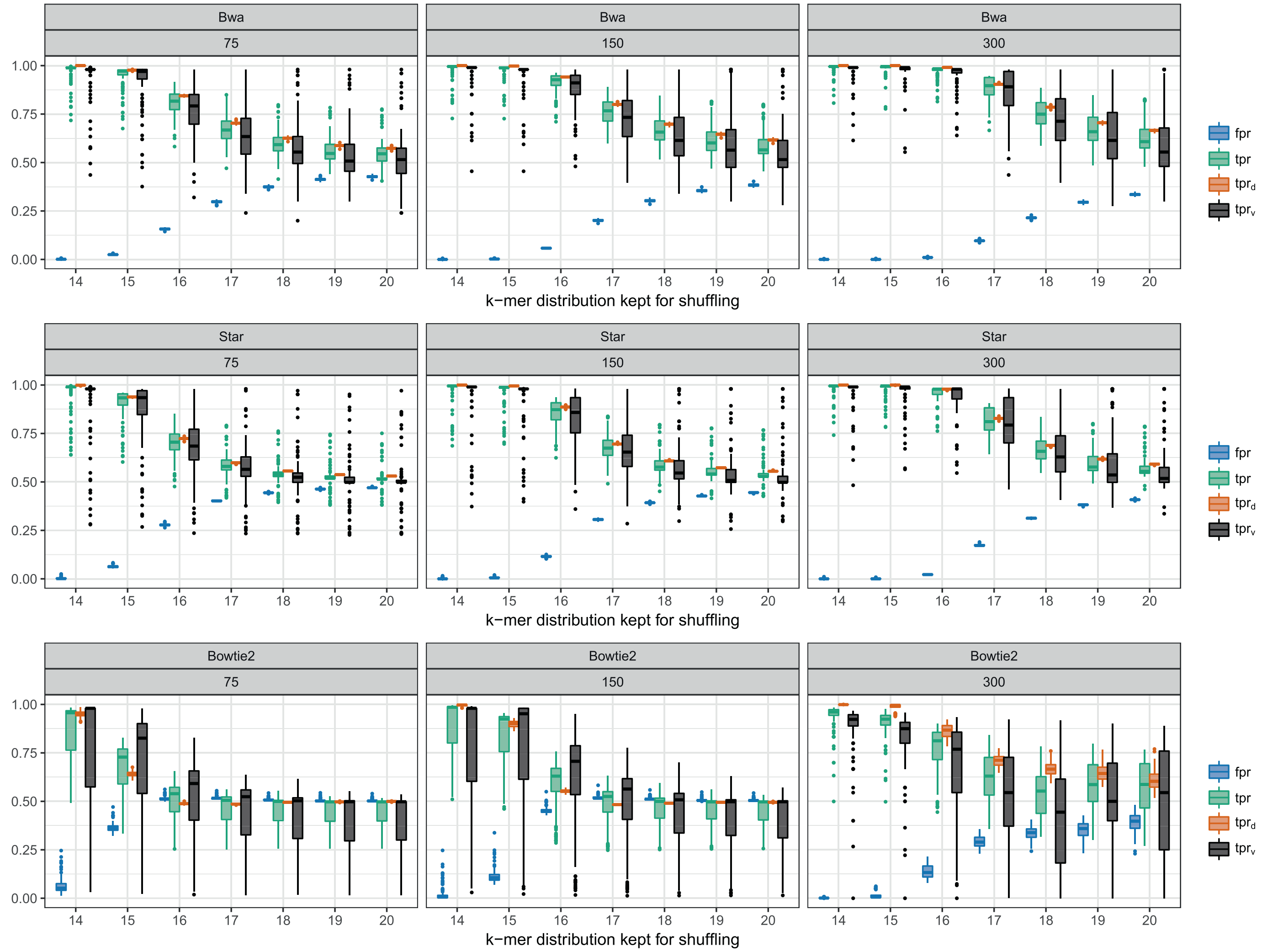

Fig. 2. Mapping results by different DNA sequence mappers: Bwa, Star, and Bowtie2. On the y-axis the classification rates of the mapped viral and decoy reads is shown. On the x-axis the *k* that is kept for the shuffling to generate the decoy reference is given. Three different read lengths were compared in our simulation study: 75, 150, and 300. The boxplots include the results of 100 runs with different virus strains: 100 reads were drawn from a virus strain and 10.000 reads were drawn from the decoy reference.

reads n_{rd}. Again, the membership of the reads to *d* and *e* can not be distinguish and is therefore pooled.

2.2. Module II: DNA read mapping

NGS produces high amount of short sequencing reads, which can be mapped to given reference genomes. The working principle of many virus detection pipelines is based on sample from a human host. In this regards, human sequence reads are removed by mapping all reads to the human reference genome. Unmapped reads are then considered to originate potentially from a virus. However, many practical virus detection problems emerge when analyzing an exotic host species, such as the tinamou, whale, or giraffe. For such species, no reference genome of sufficient quality is available, and very often no reference genome is available at all. However, existing pipelines of the category I can also be run without filtering the host reads. In the case of virus detection, with a low coverage of reads or a mutation of the virus strain, the approach will work with a mapping error. We have introduced the decoy database before, to judge these mapping errors and to bring the falsely mapped reads errors into perspective to the true mapping rate.

To circumvent the problem of having no reference genome available, we omit the step of removing host reads. Instead, all reads are directly mapped to all available viral sequences from the NCBI Genbank combined as one “artificial virus reference genome” fasta file. We downloaded approx. 2.4 million DNA sequences and approx. 3.3 million amino acid sequences. To speed up the process, we used standard DNA mappers such as BWA (Li and Durbin, 2009), STAR (Dobin et al., 2013) or Bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2009) to map the short DNA sequencing reads to the artificial viral genome. All viral sequences, consisting of full viral genomes coding domain sequences or only viral fragments, are stringed together to build an “artificial” genome consisting only of viral sequences in a single *.fasta file. This artificial genome, where the single viral sequences are regarded as chromosomes, is used as a reference genome for the mapping process. We named each ‘chromosome’ by the NCBI GenBank accession number for the related virus and the aligned read counts are reported by the mapper. Since only virus sequences are stringed together, unmapped reads can be considered to belong to the host or other unknown sources. A large number of multiple read mappings occur in this mapping process because sequences can overlap between related virus strains. Normally, the so called ‘multi-maps’ are not desired, but here we are able to detect families of different viral strains.

For this module, we evaluated three different mappers, which are able to handle a high amount of chromosomes. The criterion used for the selection consisted in choosing mappers that are widely used and are still under maintenance. Therefore, we decided to evaluate the mappers BWA, STAR, and Bowtie2. A comprehensive analysis of the three mappers and their properties can be found in the results section. The mapper Bowtie2 was run with its default options. The number of allowed read matched was risen to $K = 10$. Therefore, a single read was allowed to map to up to 10 viral sequences. The STAR and the BWA mapper were run with default values.

In general the DNA mapping step must be seen as an exploratory data analysis (Tukey, 1977). The mapping of DNA reads against a reference fasta file is not uncommon. In the case of virus detection multi maps are wanted because multi maps can describe a infection of a virus family or a group of RNA viruses. However, the question remains open, which of the DNA mapper delivers reliable results. Therefore, we build up a decoy database to judge the mapping qualities in our specific setting of the detection of low abundances of viral reads and exotic hosts.

2.3. Module III: amino acid mapping

In module I, we dropped the gene information of each single virus sequence by collapsing the sequences into one artificial genome. To circumvent this disadvantage, we added a second layer of evidence. To do this, we used the amino acid (AA) mapping as an additional module in our virus detection pipeline. The overall number of approx. 2.4 million DNA sequences is complemented with approx. 3.3 million amino acid sequences. Each amino acid sequence represents one 'chromosome' from the artificial DNA genome. Some virus strains have only one or even no gene reported in their connected Embl-file. Therefore, the number of amino acid sequences does not exceed drastically the number of DNA sequences. All amino acid sequences were combined into one artificial amino acid reference genome. Each 'chromosome' represents one amino acid sequence from a virus strain. The information was stored into an SQL database to enable faster access in the following steps of the pipeline. Details are described below in the visualization module section in the Supplementary material. The generation of the artificial amino acid reference genome allows the mapping of translated DNA reads. We used the already implemented approach Pauda (Huson and Xie, 2013), which uses the DNA aligner Bowtie2 for protein alignment to a reference database (For a short description see the Supplementary material). Several changes to the Pauda implementation were necessary to fulfill our requirements. Supplementary Table 13 shows all changes with the connected program line. Pauda only allows the mapping of single end reads. Therefore, we used PANDAseq (Masella et al., 2012) to combine paired reads into single end reads where necessary. The Pauda output is a BlastX-file, which can be parsed and matching positions can be extracted. In addition, the raw amino read counts mapped to each viral gen is reported. The amino acid mapping is heavily based on an good annotation. The user can improve the results, if good curated annotation database is available by its own.

3. Results

In this section, we first show the results of a simulation for the evaluation of our analysis pipeline (Module I - III), in particular the evaluation of the decoy approach (Module I). Furthermore, we show the results of applying our virus detection pipeline to two example sample files. A visualization of the findings can be found in the supplementary material.

3.1. Evaluation of the decoy database

For the usage of the decoy database in a detection pipeline the k for the generation of the decoy sequences must be determined. A small k will cause no hits to the decoy database, while a large k causes an equal decoy database in comparison to the sampled original reference. Fig. 2 and Supplementary Fig. 7 show the simulation results of 100 virus strains with 100 reads combined with 10,000 decoy reads of read lengths: 75, 150, and 300. Three different mapping tools were compared: Bwa, Star, and Bowtie2. On the x-axis the used k of the shuffling procedure for the generation of the decoy sequences is plotted and shows the classification rates on the y-axis. Supplementary Fig. 7 shows the percentage of remapped reads, i.e. the mappings to the decoy database. Since the Bwa mapper does not allow multi maps the y-axis for the Bwa results shows the percentage of mapped reads, while the other two mappers allow multi maps, hence the y-axis shows mappings.

The Bwa mapper does not allow multi-mapping reads. Hence, with an increase of k, the true virus reads that map to the true virus will decrease (Supplementary Fig. 7). This is also true for the proportion of the decoy reads mapped to the decoy region of the reference. The number of virus reads mapped to the decoy reference is also increased for higher values of k. This tendency can be lowered, if a longer read length is provided. Further, the Bwa mapper shows a slower increase of the *fpr* rate caused by the single mapping of each read. Therefore, it can be concluded, that the Bwa mapper has problems to map a read consistently if a high diversity of virus strains are in the sample. Even more if the viral strains in the sample are closely related. Moreover, this strong relationship occurs also frequently in our artificial virus database, because many families and coding DNA sequences (CDS) regions are very similar and connected. Therefore, we decided to remove the Bwa mapper for the analysis of the biological samples below.

The Star aligner shows a nearly constant detection rate of about 100% of the reads for all virus reads mapped to the true virus. In contrast to the Bowtie2 mapper the multi mapping rate of the true virus to the true genome is lower (Supplementary Fig. 7). For higher k the multi-mapping can be seen on the decoy reads mapped to the decoy database. The percentage of falsely mapped virus reads to the decoy genome increases with higher values of k but can be lowered with a longer read length. The Star mapper shows the lowest variance of the classification rates of the three mappers. Due to the fact that all virus reads are correctly mapped, we consider STAR as the mapper for the analysis of the biological samples.

Finally, we tested the Bowtie2 mapper. First high multi-mapping can be observed (Supplementary Fig. 7). The true virus reads are mapped multiple times to the true virus strain, as well as the decoy reads to the decoy reference. In addition, the true virus reads are often mapped to other viruses. This false mapping can only be observed with the Bowtie2 mapper. In addition, the decoy reads are also mapped to the virus reference. In this setting, the Bowtie2 mapper has problems with the mapping of read length of 300. In this case, the performance is lowered in all mappings. This property cannot be seen in Fig. 2 as clearly as in Supplementary Fig. 7. Overall, less reads are mapped to the reference with a read length of 300. Hence, we recommend to use both types of Figures for the judgment of the mapper. Finally, Bowtie2 has the highest overall variance of the classification rates. As an advantage, the high amount of multi-mapped reads can allow the detection of quasi species or families of related virus strains. Therefore, it is possible to draw conclusions about the infectious virus strain even if its sequence is not in the database. These cases can occur frequently, especially if the host is an animal without reference genome available, as presented in the following analyseis of two biological samples.

3.2. Evaluating the pipeline on the biological samples

In this work, we demonstrate the combination of the single modules into one virus detection pipeline on recently published sequenced samples. With our pipeline, we were able to rediscover the virus strains identified in the original publications of the samples. To evaluate the proposed modules on two biological data sets, we show the re-analysis of two published high-throughput samples: a tinamou sample (NGS-16007) (Jo et al., 2017a) and a fin whale sample (NGS-16021) (Jo et al., 2017b). Fresh frozen liver tissue and formalin-fixed paraffin-embedded (FFPE) brain tissue were used as starting material for the tinamou and fin whale samples, respectively. For both, a DNA and RNA extraction was performed and were run as separate samples. Therefore, four samples had been analyzed: NGS-16007-DNA, NGS-16007-RNA, NGS-16021-DNA, and NGS-160021-RNA. For both species, no reference genome was available. In general, reference genomes for birds and water animals are rare. All raw reads were preprocessed as quality control using Trimmomatic (Bolger et al., 2014). Trimmomatic removed reads that were shorter than 50 bases and clipped nucleic bases at the beginning and the end of the reads with a quality lower than 10. We also removed potential left over Ilumina adapters. The sample NGS-16007-DNA included 621,465 reads, 570,134 (91.74%) after quality control, NGS-16007-RNA had 599,077 read, 539,977 (90.13%) after quality control, NGS-16007-DNA consists of 208,281 reads, 158,625 (76.16%) after quality control, and NGS-16021-RNA consists of 675,244 reads, 583,798 (86.46%) after quality control. For the DNA read mapping we used the Bowtie2 and the Star mapper. The reference genome was the artificial genome consisting of approx. 2.4 million viral sequences. The amino acid mapping was done by Pauda using Bowtie2 as mapper. We report a virus species as detected, if more than five read are mapped to the species reference. The data examples used for our pipeline evaluation have the advantage, that a PCR and corresponding primers have confirmed the existence of the virus strains in the samples as detailed by the related publications. To try out our pipeline, we refer the reader to other example data sets available at Sequence Read Archives (Sequence Read Archives, n.d.).

The results for all four samples are shown in the Supplementary Tables 1 to 8. Table 2 reports the overall number of detected species by DNA read mapping. Due to the higher amount of multi-read mappings by Bowtie2, it produced much more detected species than Star. Supplementary Fig. 2 shows the scatter plot of the mapped read counts by Bowtie2 and Star. It can be clearly seen, that Bowtie2 maps more reads to the reference sequences than Star. We know from the already published results that the sample NGS-16007 is infected with a DNA virus (avian hepadnavirus: GenBank accession numbers KY977506 and KY977507) and the sample NGS-160021 with an RNA virus (dolphin morbillivirus: GenBank accession numbers KR337460, KY681807, andKP835991). We are able to detect the avian hepadnavirus with rank 1 and 2 using Bowtie2, NGS-16007-DNA with 858,208 (KY977506) and 857,901 (KY977507) mapped reads, as well as the Star mapper, NGS-16007-DNA with 907,899 (KY977506) and 598,367 (KY977507). Both avian hepadnavirus strains are also ranked at position 1 and 2 on the amino acid sequence mapping. Using Bowtie2, a full coverage of the reference can be reached, 99% percent of the assembled read consensus sequence is equal to the reference, and the mean base frequency is above 91.1%. The numbers for the Star mapper are nearly the same only the mean base frequency being lower at approx. 81%. Overall, there is no difference between the usage of the DNA or RNA extraction sample.

The true virus is also known for the NGS-160021 sample, which was infected by a dolphin morbillivirus. The morbillivirus was not detected using the NGS-160021-DNA sample. The findings for the NGS-160021-DNA sample in the Supplementary Table 7 show only a low number of DNA and AA hits, as well as worse coverage an mean base frequencies. We will therefore judge these findings as false positives. The *fpr* of 0.5 (STAR) and 0.45 (Bowtie2) is also higher in this sample, which also gives evidence towards a false positive sample without any infection of a viral strain. On the other hand, the morbillivirus was identified using the NGS- 16021-RNA sample. The top ranked hit can be seen in Supplementary Fig. 1. The top 20 of the detected viral strains are mainly connected to the dolphin morbillivirus (Supplementary Table 8). This ranking would not be achieved by concentrating only on the DNA sequence reads. The ranking is mainly driven by the amino acid mapping, where the top hits are all connected to the dolphin morbillivirus. This example shows the strength of the combination of the DNA read sequence and the amino acid mapping. Moreover, Bowtie2 is able to detect all three species while Star is only able to detect two. Star looses one of the three species due to the lower amount of multi-read mappings.

We evaluated also the coverage of Bowtie2 and Star. Supplementary Figs. 7 to 9 show the read coverage of the reference. Both mapping tools detect the true virus strain in the NGS-16007 sample with 100% coverage (orange points). The difference of the multi-mapping and the coverage can be seen in the NGS-16021-RNA sample. Bowtie2 generates more hits and is therefore able to achieve a higher coverage of the true findings (orange points). This behavior is also true for the percentage of equal bases between the reference and the reads as well as the mean base frequency of the mapped reads.

To judge the results of the artificial genome mapping and the 'pseudo' assembly of the DNA sequence reads, we used the assembler (IVA) (Hunt et al., 2015) for accurate de novo assembly of RNA virus genomes. IVA has an internal virus detection approach using Kraken (Wood and Salzberg, 2014). Table 3 shows the lowest branches of the Kraken report. Furthermore, we have blasted all assembled contigs build by IVA using the NCBI blastn suite (Supplementary Tables 9 to 12). The Kraken results show the same tendency. The tinamou sample (NGS-16007) is infected with a bird hepatitis B virus. The tinamou hepatitis B virus was not in the Kraken database. Therefore, the hit produced was the nearest possible one. The blastn results of the contigs identified the tinamou hepatitis B virus. The fin whale sample of the RNA extraction was also correctly classified as a Dolphin morbillivirus. The blastn results could also detect parts of the dolphin morbillivirus, but was not able to find the correct GenBank accession number mentioned in Jo et al. (2017) (Jo et al., 2017b). Overall, only a small proportion of the virus reference genomes could be covered by the DNA sequence reads.

4. Discussion

We have presented a new approach for the detection of viral sequences and families using high-throughput sequencing data for scenarios where no host reference genome is available. The three presented modules can be used separately or as a combination in a complete virus detection pipeline. The use of the artificial genome allows to overcome the missing of the host reference genome, while the decoy approach is designed to judge the potential proportion of false positives by the used mapper. The visualization of the mapped reads to

Table 2
Number of detected virus strains n_{virus} with more than five DNA read counts, true positive rate of the decoy reads tpr_d, false positive rate *fpr*, and multi mapping rate *mmr*.

Bowtie2	n_{virus}	tpr_d	*fpr*	*mmr*
NGS-16007-DNA	130	0.74	0.28	2.32
NGS-16007-RNA	372	0.63	0.30	5.10
NGS-160021-DNA	421	0.72	0.45	2.89
NGS-160021-RNA	1351	0.82	0.44	7.63
STAR	n_{virus}	tpr_d	*fpr*	*mmr*
NGS-16007-DNA	7	0.99	0.20	1.04
NGS-16007-RNA	18	0.96	0.20	1.11
NGS-160021-DNA	40	0.94	0.50	1.02
NGS-160021-RNA	161	0.96	0.39	1.40

Table 3
Kraken report of the contigs generated by IVA.

	Kraken	
NGS-16007-DNA	Parrot hepatitis B virus	Parrot hepatitis B virus
NGS-16007-RNA	Parrot hepatitis B virus	Parrot hepatitis B virus
NGS-160021-DNA	Parrot hepatitis B virus	Parrot hepatitis B virus
NGS-160021-RNA	Dolphin morbillivirus	Duck adenovirus 2

the reference removes false findings of badly mapped or uninformative reads at the edges of the reference.

A clear advantage of our detection pipeline is its speed. With our approach, the processing of a single sample takes 30 min on average on a Linux-cluster with 40 cores. The speed is mainly achieved by mapping the sample fastq-file against the artificial genome fasta-file, wherein chromosomes represent viruses or parts of them. This approach is especially useful when sequencing is performed with a low coverage which does not allow for the assembly of larger contigs. In an assembly based pipeline, contigs could be aligned to virus reference genomes using blast tools. Each contig would then get a list of potential origins ordered by the *E*-value. Depending on the length of the contigs the calculations would still be time-consuming. As determined by Scheuch et al. (Scheuch et al., 2015), aligning 250.000 reads by the blastn program takes approx. 128 h. This approach is therefore only reasonable when host reads can be removed before the assembly procedure, and when sequencing was performed with sufficient coverage for assembly. Thus, our pipeline makes use of read mapping which is much faster than read alignment. The higher speed comes with some uncertainty in the mapping results. Therefore, we added a decoy, a amino mapping and visualization layer to the judgment of the detection findings.

Although the processing of an individual sample with our pipeline is relatively fast, the building of the artificial genome is still time-consuming and may be limited by the available working memory. To overcome the memory burden, different chunks or blocks of the artificial reference genome could be built and then analyzed sequentially. Next, the final results from the chunks must be merged. As an asset, the artificial genome has to be built once to make the pipeline run.

A further useful characteristic of our detection pipeline is that multi-maps can occur, i.e. reads that map to multiple positions. This can eventually be helpful to identify quasispecies or different strains of the virus. As mentioned by Domingo et al. (Domingo et al., 1985; Domingo and Schuster, 2016), RNA virus strains tend to build up quasispecies in a biological sample: an extremely heterogeneous population of one viral RNA strain with many mutation in its genome. Therefore, many variations of one RNA strain may exists in the host. This is also a common problem with the research on new zoonoses (Woolhouse et al., 2016). Although the infectious RNA virus strain might not be detected, because the strain is not very similar to the original NCBI GenBank entry, single fragments of the virus might be identified and be part of the artificial genome used in the pipeline. In this case, the multi-maps produced by Bowtie2 are not a problematic feature but a possible solution to detect all subtypes of the infectious RNA strain. To detect such quasispecies our approach helps by using a 'pseudo' assembly of the DNA sequence reads guided by the mapping positions on the reference genome. In contrast, if the reads are directly assembled and compared to the reference the problem of host polluted reads would occur. In addition, the heterogeneous quasispecies and a majority of other naturally occurring viral strains would drastically lower the assembly quality of the generated contigs.

While multi-maps are useful for identifying quasispecies, their occurence must still be controlled. Therefore, we tested different DNA mappers on a decoy database to get on idea of the behavior of the mappers. Especially, the awareness of false positive mapped reads in the case of multi-mapping must be sharpened. Therefore, viral strains with only a small number of mapped DNA sequence reads must be handled with care and reproduced in the wet lab. The BWA mapper allows no multiple read mappings, which turned out as a big drawback in real experimental data and for the answering of many biological questions.

In examples with real biological data, we were able to detect the virus strains with our computational analyses. However, it is important to take into account that the success of the analyses can also be dependent on the sample material (frozen tissue/FFPE), the virus characteristics (DNA/RNA), and stage of infection (acute/chronic). In the case of the tinamou sample, which was a frozen tissue, avian hepadnavirus (a DNA virus) was detected using datasets generated from both DNA and RNA extractions. Whereas in the case of the fin whale sample, which was an FFPE tissue, dolphin morbillivirus (an RNA virus) was only detected using the dataset generated by RNA extraction.

In general, a computational virus detection pipeline can not be fully automated and results remain uncertain. Therefore, we implemented the decoy modul to assess different error rates. Finally, the pipeline produces a sorted list of possible hits which must then be validated by further laboratory assays. Thus, NGS data allows to get a starting point for the validation in the wet lab. The DNA sequence reads and the assembled contigs allow to design primers for later diagnostic stages. In order to not omit a virus for further laboratory evaluation, we accept a higher rate of false positives to have the power to detect all possible infectious virus strains.

5. Conclusion

Current bioinformatics pipelines for virus discovery mostly assume that reference genome data of the host is available or that the sequencing coverage is sufficiently large to assemble contigs. Our newly developed pipeline is applicable if both requirements are not fulfilled. The three modules we presented can either be incorporated individually in other pipelines or be used as a full pipeline. Further, we evaluated the practicability of three different mappers for the detection of viral reads. From this evaluation, we recommend the STAR or Bowtie2 mapper which allow for possible multi mapping of reads. Thus, quasi species and virus of the same family can be discovered. In constrast, the BWA mapper does not allow multi mapped reads and will therefore spread reads over the family. We are the first to present a decoy approach within a virus detection pipeline that can be used to assess error rates so that the quality of the final result can be judged. We provide an implementation of our modules via Github (`https://github.com/jkruppa/virDisco`). However, the principle of these modules can easily be re-implemented by other researchers.

Supplementary data to this article can be found online.

Funding

This work was supported by the Niedersachsen-Research Network on Neuroinfectiology (N-RENNT) of the Ministry of Science and Culture of Lower Saxony. The funding body was not involved in the design or conclusion of our study.

This study was supported by the COMPARE project and received funding from the European Union's Horizon 2020 research and innovation program COMPARE (grant agreement no. 643476).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2018.09.026.

References

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30 (15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Chen, Y., Yao, H., Thompson, E.J., Tannir, N.M., Weinstein, J.N., Su, X., 2012. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 29 (2), 266–267. https://doi.org/10.1093/bioinformatics/bts665.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. Star: ultrafast universal rna-seq aligner. Bioinformatics 29 (1), 15–21.

Domingo, E., Schuster, P., 2016. Quasispecies: from theory to experimental systems. In: Current Topics in Microbiology and Immunology. Vol. 392 Springer.

Domingo, E., Martnez-Salas, E., Sobrino, F., de la Torre, J.C., Portela, A., Ortn, J., López-Galindez, C., Pérez-Breña, P., Villanueva, N., Nájera, R., et al., 1985. The quasispecies (extremely heterogeneous) nature of viral rna genome populations: biological relevance—a review. Gene 40 (1), 1–8.

Ensemble Database https://www.ensembl.org/info/data/ftp/index.html.

Fonseca, N.A., Rung, J., Brazma, A., Marioni, J.C., 2012. Tools for mapping high-throughput sequencing data. Bioinformatics 28 (24), 3169–3177. https://doi.org/10.1093/bioinformatics/bts605.

Forster, M., Szymczak, S., Ellinghaus, D., Hemmrich, G., Rühlemann, M., Kraemer, L., Mucha, S., Wienbrandt, L., Stanulla, M., Franke, A., et al., 2015. Vy-per: eliminating false positive detection of virus integration events in next generation sequencing data. Sci. Rep. 5, 11534.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29 (7), 644–652. https://doi.org/10.1038/nbt.1883.

Hunt, M., Gall, A., Ong, S.H., Brener, J., Ferns, B., Goulder, P., Nastouli, E., Keane, J.A., Kellam, P., Otto, T.D., 2015. Iva: accurate de novo assembly of rna virus genomes. Bioinformatics 31 (14), 2374–2376.

Huson, D.H., Xie, C., 2013. A poor man's BLASTX–high-throughput metagenomic protein database search using PAUDA. Bioinformatics 30 (1), 38–39. https://doi.org/10.1093/bioinformatics/btt254.

Jiang, M., Anderson, J., Gillespie, J., Mayne, M., 2008. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinf. 9 (1), 192. https://doi.org/10.1186/1471-2105-9-192.

Jo, W.K., Pfankuche, V.M., Petersen, H., Frei, S., Kummrow, M., Lorenzen, S., Ludlow, M., Metzger, J., Baumgaertner, W., Osterhaus, A., van der Vries, E., 2017a. New avian hepadnavirus in palaeognathous bird, germany. Emerg. Infect. Dis. 23 (1), 2089–2091.

Jo, W.K., Grilo, M.L., Wohlsein, P., Andersen-Ranberg, E.U., Hansen, M.S., Kinze, C.C., Hjulsager, C.K., Olsen, M.T., Lehnert, K., Prenger-Berninghoff, E., et al., 2017b. Dolphin morbillivirus in a fin whale (balaenoptera physalus) in Denmark, 2016. J. Wildl. Dis. 53 (4), 921–924.

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12 (4), 656–664. https://doi.org/10.1101/gr.229202.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with bowtie 2. Nat. Methods 9 (4), 357–359. https://doi.org/10.1038/nmeth.1923.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (3), R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Lee, W.-P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., Marth, G.T., 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9 (3), e90581. https://doi.org/10.1371/journal.pone.0090581.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics 25 (14), 1754–1760.

Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G., Neufeld, J.D., 2012. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinf. 13 (1), 31. https://doi.org/10.1186/1471-2105-13-31.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Chatelier, E.L., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.-M., dos Santos, M.B.Q., Blom, N., Borruel, N., Burgdorf, K.S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R.S., Kennedy, S., Kristiansen, K., Kultima, J.R., Léonard, P., Levenez, F., Lund, O., Moumen, B., Paslier, D.L., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sørensen, S., Tap, J., Tims, S., Ussery, D.W., Yamada, T., Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Chatelier, E.L., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.-M., dos Santos, M.B.Q., Blom, N., Borruel, N., Burgdorf, K.S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R.S., Kennedy, S., Kristiansen, K., Kultima, J.R., Leonard, P., Levenez, F., Lund, O., Moumen, B., Paslier, D.L., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sørensen, S., Tap, J., Tims, S., Ussery, D.W., Yamada, T., Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., Ehrlich, S.D., Jamet, A., Mérieux, A., Cultrone, A., Torrejon, A., Quinquis, B., Brechot, C., Delorme, C., M'Rini, C., de Vos, W.M., Maguin, E., Varela, E., Guedon, E., Gwen, F., Haimet, F., Artiguenave, F., Vandemeulebrouck, G., Denariaz, G., Khaci, G., Blottière, H., Knol, J., Weissenbach, J., van Hylckama Vlieg, J.E.T., Torben, J., Parkhill, J., Turner, K., van de Guchte, M., Antolin, M., Rescigno, M., Kleerebezem, M., Derrien, M., Galleron, N., Sanchez, N., Grarup, N., Veiga, P., Oozeer, R., Dervyn, R., Layec, S., Bruls, T., Winogradski, Y., Renault, Z.E.G.P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., Ehrlich, S.D., 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32 (8), 822–828. https://doi.org/10.1038/nbt.2939.

Petty, T.J., Cordey, S., Padioleau, I., Docquier, M., Turin, L., Preynat-Seauve, O., Zdobnov, E.M., Kaiser, L., 2014. Comprehensive human virus screening using high-throughput sequencing with a user-friendly representation of bioinformatics analysis: a pilot study. J. Clin. Microbiol. 52 (9), 3351–3361. https://doi.org/10.1128/jcm.01389-14.

Reidegeld, K.A., Eisenacher, M., Kohl, M., Chamrad, D., Koerting, G., Blueggel, M., Meyer, H.E., Stephan, C., 2008. An easy-to-use decoy database builder software tool, implementing different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications. Proteomics 8 (6), 1129–1137. https://doi.org/10.1002/pmic.200701073.

Scheuch, M., Höper, D., Beer, M., 2015. Riems: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC Bioinf. 16 (1), 69.

Sequence Read Archives https://www.ncbi.nlm.nih.gov/sra/docs/.

SMALT http://www.sanger.ac.uk/science/tools/smalt-0.

Smits, S.L., Osterhaus, A.D., 2013. Virus discovery: one step beyond. Curr. Opin. Virol. 3 (2), e1–e6.

Smits, S.L., Bodewes, R., Ruiz-González, A., Baumgärtner, W., Koopmans, M.P., Osterhaus, A.D., Schürch, A.C., 2015. Recovering full-length viral genomes from metagenomes. Front. Microbiol. 6, 1069.

Tukey, J.W., 1977. Exploratory Data Analysis. Vol. 2 (Reading, Mass).

Wang, Q., Jia, P., Zhao, 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 (5), e64465. https://doi.org/10.1371/journal.pone.0064465.

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15 (3), R46. https://doi.org/10.1186/gb-2014-15-3-r46.

Woolhouse, M.E., Brierley, L., McCaffery, C., Lycett, S., 2016. Assessing the epidemic potential of rna and dna viruses. Emerg. Infect. Dis. 22 (12), 2037.

Zhao, G., Krishnamurthy, S., Cai, Z., Popov, V.L., da Rosa, A.P.T., Guzman, H., Cao, S., Virgin, H.W., Tesh, R.B., Wang, D., 2013. Identification of novel viruses using VirusHunter—an automated data analysis pipeline. PLoS One 8 (10), e78470. https://doi.org/10.1371/journal.pone.0078470.