RESEARCH ARTICLE

ZWA: Viral genome assembly and characterization hindrances from virus-host chimeric reads; a refining approach

Nikolas Dovrolis*, Katerina Kassela, Konstantinos Konstantinidis, Adamantia Kouvela, Stavroula Veletza, Ioannis Karakasiliotis*

Laboratory of Biology, Department of Medicine, Democritus University of Thrace, Alexandroupolis, Greece

* ndovroli@med.duth.gr (ND); ioakarak@med.duth.gr (IK)

Abstract

Viral metagenomics, also known as virome studies, have yielded an unprecedented number of novel sequences, essential in recognizing and characterizing the etiological agent and the origin of emerging infectious diseases. Several tools and pipelines have been developed, to date, for the identification and assembly of viral genomes. Assembly pipelines often result in viral genomes contaminated with host genetic material, some of which are currently deposited into public databases. In the current report, we present a group of deposited sequences that encompass ribosomal RNA (rRNA) contamination. We highlight the detrimental role of chimeric next generation sequencing reads, between host rRNA sequences and viral sequences, in virus genome assembly and we present the hindrances these reads may pose to current methodologies. We have further developed a refining pipeline, the Zero Waste Algorithm (ZWA) that assists in the assembly of low abundance viral genomes. ZWA performs context-depended trimming of chimeric reads, precisely removing their rRNA moiety. These, otherwise discarded, reads were fed to the assembly pipeline and assisted in the construction of larger and cleaner contigs making a substantial impact on current assembly methodologies. ZWA pipeline may significantly enhance virus genome assembly from low abundance samples and virus metagenomics approaches in which a small number of reads determine genome quality and integrity.

Author summary

For years now the study of viruses and their genetic composition has been important in their identification and classification. Especially in these times of the pandemic turmoil, accurate knowledge of a virus' exact genetic composition can help identify its strengths and weaknesses allowing us to track its evolution and assist in the development of vaccines and antiviral agents. The reconstruction of these genomic sequences is called the assembly process, a bioinformatics approach which can be complicated and full of pitfalls. This work identifies one such issue, concerning artifacts introduced in viral genomes from the new technologies of nucleic acid sequencing. The proposed algorithm helps alleviate this problem by tentatively removing these problematic regions while keeping the vast

OPEN ACCESS

Citation: Dovrolis N, Kassela K, Konstantinidis K, Kouvela A, Veletza S, Karakasiliotis I (2021) ZWA: Viral genome assembly and characterization hindrances from virus-host chimeric reads; a refining approach. PLoS Comput Biol 17(8): e1009304. https://doi.org/10.1371/journal.pcbi.1009304

Editor: Christos A. Ouzounis, CPERI, GREECE

Received: March 19, 2021

Accepted: July 24, 2021

Published: August 9, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1009304

Copyright: © 2021 Dovrolis et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data underlying this article are available in BioSample repository with the following accession STUDY: PRJNA681030, SAMPLE: SAMN17358770, EXPERIMENT: SRX9862216, RUN: sacha.fastq.gz (SRR13449040 - https://www.ncbi.nlm.nih.gov/sra/SRR13449040/). All the associated scripts can be found at https://github.com/ndovro/ZWA and https://hub.docker.com/repository/docker/ndovroli/zwa/.

Funding: All authors are co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH/CREATE/INNOVATE (project code: T1EDK/5000). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

majority of the genetic information required to produce a more complete viral genome. This work is anticipated to assist in the submission of higher integrity and accuracy viral genomes in public databases used for novel virus identification and characterization.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Identification of viruses through next generation sequencing (NGS) relies on the use of well curated viral databases [1]. Recent advances and the broad application of viral genome assemblies, from unbiased virus screening, have yielded an unprecedented amount of novel complete or partial genomes [2]. To date, NGS has been essential in the discovery of novel viruses and the analysis of complex samples in which multiple viruses are present. RNA-seq analysis and virus genome assembly from biological and environmental materials, containing novel human and animal viruses such as SARS-cov2, has greatly assisted their prompt and unbiased characterization [3]. Full genome assembly of SARS-cov2 was crucial for understanding the virus' biology and its origin along with the design of efficient molecular diagnosis methodologies and the molecular tracing of the pandemic [4]. In parallel, NGS analysis of complex clinical or environmental samples has introduced the "virome" [5–7], a new concept of the metagenomics era.

As viruses are a highly divergent form of life with multiple evolutionary origins [8,9], metagenomics approaches rely on unbiased DNA/RNA sequencing rather than targeted PCR based sequencing [10]. Alignment of the yielded DNA/RNA sequencing reads or further processed assemblies against virus databases are critical for the identification of viral components in a sample. Curated databases contain a very small portion of the currently known viral sequences. Open databases, such as GenBank, while they contain most of the current sequences, are poorly curated regarding the integrity and the accuracy of the submitted sequences. Indeed, various reports have highlighted contamination of sequences with bacterial moieties that were erroneously incorporated in the final assembly [11–13]. Such errors are especially important as they result in false positive identification of viruses that happen to contain in their proposed viral genomes parts of host DNA or RNA.

De novo assembly of novel viruses is based on a few thousand reads, or less, since environmental and clinical samples are usually not enriched [14]. Artifacts may arise, in part due simultaneous presence of multiple forms of the same RNA; positive and negative sense genomic RNAs or mRNAs, splicing variants and the presence of chimeric reads. Chimeric reads originate from chimeric DNA fragments during library preparation and have, in the past, been proven as a significant impediment in *de novo* transcriptome assemblies [15,16]. This is especially important during viral genome assembly from samples poor in viral sequences, where the complete removal of chimeric reads may result in partial genomes.

Contaminant sequences from viruses, bacteria and fungi have been an issue in DNA and RNA sequencing for years with a variety of approaches identifying and filtering them [17–20]. Previous works have highlighted the impact of metagenomic contaminants in genome or transcriptome assemblies and the role of NGS library preparations in the generation of contaminant reads that encompass exogenous nucleic acids [21]. In 2017, Dittami et al [12] recognizing the presence of contaminants in genomes, designed a bioinformatics tool which analyzed published *Saccharina japonica* genomes and detected the contaminants proposing their tool as a part of standards bioinformatic analysis for the specific kelp. This tool focused on a breakdown of the genome into smaller contigs and the usage of BLASTn [22] to identify their origin. On the same year Fierst et al. [23] proposed a Machine Learning approach which, using nucleotide sequence similarities (adapted from microbial research techniques), filtered reads which were detected as contaminants, prior to final genome assembly. On the same principle, a variety of tools have been designed to identify and remove host-derived reads from metagenomic samples to assist in genome assembly. In fact, all current approaches and applied pipelines detect and curate genomes and transcriptomes post assembly or by excluding contaminating (chimeric) reads from the assembly. GenCoF [24], RNA-QC-chain [25] and FastqPuri [26], for example, are designed specifically as rRNA removal pipelines which implement a variety of tools for quality-control of sequences (based on sequencing/mapping quality) and complete removal of rRNA reads from the samples. The authors of MCSC [27] have also recognized the issue of contaminated assemblies and have developed a hierarchical cluster algorithm which distinguishes reads as contaminants and then tries to align and remove them. rRNAFilter [28], as the name suggests, is also a removal tool for rRNA reads with the added benefit of being alignment-free and based on read *k-mers* for the identification of rRNA. Fastq Screen [29] and SortMeRNA [30] adopt a similar approach to each other which involves mapping reads onto multiple genomes in order to identify and separate contaminant sequences. These latter approaches are often used in metagenomic studies, which aim to keep the rRNA reads in order to identify micro-organisms and discard host reads. For example during the identification of the new SARS-CoV-2 such an approach was utilized [31]. Testament to the fact that rRNA contamination remains a problem to this day, is the current development of new tools to leverage the issue.

In the present study, we sought to assess the impact of host-virus chimeras in the integrity of GenBank database and the role of rRNA-virus chimeric reads in the emergence of rRNA-contaminated virus genome contigs during *de novo* assembly. Furthermore, we developed an algorithm that enhanced the *de novo* assembly of viral genomes by upcycling chimeric reads, discarded by current pipelines.

Materials and methods

Reference databases

Two reference databases were developed based on sequences found on NCBI's Nucleotide database (NUCCORE). The first one, the Ribosomal Database (RiDB) contained 28S and 18S ribosomal RNA (rRNA) sequences from selected species that have been extensively studied in the past for the identification and assembly of viruses. For *Gallus gallus* (XR_003078040.1 and AF173612.1), for *Bos Taurus* (NR_036644.1 and NR_036642.1), for *Ixodes simplex* (KY457499.1 and KY457499.1), for *Anopheles sacharovi* (L22060.1 and X57172.1), for *Canis lupus* (XR_003129431.1 and XR_003129430.1), for *Ovis aries* (XR_003587871.1 and KY129860.1), for *Mus musculus* (NR_003279.1 and NR_003278.3) and for *Homo sapiens* (M11167.1 and X03205.1). The second database, the Viral Database (ViDB), was constructed using NCBI's e-utilities [32] with the keyword "viruses [ORGN]" and contains 3,197,815 virus sequence entries (on November 27 2019).

Ribosomal contamination detection and identification

The RiDB was used to create, through a custom Python script (https://github.com/ndovro/ZWA/blob/main/mockup.py), mock fastq files that contained 4 million random "reads" (90-150bp) of Phred 30 quality for each of the host rRNA. These synthetic datasets were validated by aligning them on the original sequences, in order to assess even distribution of reads, high depth and full coverage (**S1 Fig**). These mock files simulate sequencing runs and were aligned on the ViDB using 3 different aligners after creating the appropriate indices: BBMAP v 38.50 [33], STAR v. 020201 [34] and BOWTIE2 v. 2.3.4.1 [35]. All aligning runs were configured to keep reads that mapped on various contigs of the reference (different viral sequences of ViDB) and to ignore splicing events. This way we were able to identify sequences in Genbank that contain these chimeric rRNA-virus assemblies. The percentile (relative position) on each ViDB sequence where the rRNA contamination was detected was calculated and visualized all using Graphpad Prism 8 [36], Venny v.2.0.1 [37] and the R statistical programming language [38].

RNA-seq

A pool of five *Anopheles sacharovi* individuals were homogenized and total RNA was extracted by TRIzol reagent (Thermo Fischer Scientific) according to the manufacturer's protocol. Whole transcriptome libraries were prepared from 500 ng of RNA extract, using the Ion Total RNA-Seq v2 Core Kit (#4479789, ThermoFisher Scientific) according to the manufacturer's instruction. In brief, the RNA library preparation involved RNA fragmentation, adapter ligation, reverse transcription and 14 cycles of PCR amplification using Ion Xpress RNA-Seq Barcode 1–16 Kit (#4475485, ThermoFisher Scientific). Quantification of the library was performed using Qubit Fluorometer high-sensitivity kit (ThermoFisher Scientific) and its median size was determined in LabChip GX Touch 24 (PerkinElmer). The libraries were loaded onto an Ion 540 chip, using Ion Chef (Thermo Fisher Scientific) and sequencing was carried out on an Ion GeneStudio S5 sequencer (ThermoFisher Scientific). Ion GeneStudio S5 sequencer returns the reads already quality trimmed.

Sequence assembly and annotation

Trinity v2.8.5 [39] was used to *de novo* assemble contigs from the RNA-seq reads. The contigs produced by the assembler were analyzed with batch BLASTX (organism: viruses [taxid:10239]) for the identification of sequences that corresponded to viral genomes. Contigs corresponded to the Xanthi rhabdovirus (MW520377). The above contigs were used as reference for mapping the mock read file of RiDB in order to identify contigs encompassing rRNA moieties. Bam files of selected rRNA "contaminated" contigs were visualized using the Integrative Genomics Viewer (IGV) v. 2.5.3 [40] to identify the chimeric reads.

In order to identify the percentage of chimeric reads that corresponded to both viral genomes and rRNAs we used two reference files. One contained the rRNA genes' sequences for *Anopheles sacharovi* (MT808434, MT808462) and the other containing the Xanthi rhabdovirus sequence. BBMAP was used to map reads on each of these references. Using samtools v.1.7 [41] we identified the reads which mapped on pairings (virus vs rRNA) of the reference files. Venny v.2.0.1 was used to create Venn diagrams.

Contamination cleanup and assembly optimization

After identifying the existence of chimeric reads as a part of the sequencing process, we set out to create a method that computationally removed ribosomal contamination from RNAseq reads without having to discard the entire read. We constructed and implemented a pipeline, the Zero Waste Algorithm (https://github.com/ndovro/ZWA/), which performs context-based read trimming by identifying the exact bases on each read where the rRNA has been fused and cutting them out leaving us with a "cured" read.

ZWA implements the following steps using well established bioinformatics tools:

1. The run is initialized by providing the raw RNAseq reads in fastq format and an rRNA reference file appropriate to our sample.

2. BWA mem v. 0.7.17-r1188 [42] is used to separate the reads that align to the rRNA reference (ribo.fastq) and those that don't (other.fastq)

3. Using samtools [41] we identify the reads that contain soft-clipping in the ribo.fastq file, convert it to fasta using BBmap's [33] reformat script and discard the pure ribosomal sequences using faSomeRecords (https://github.com/santiagosnchez/faSomeRecords).

4. On the soft-clipped sequences from the previous step we perform a local BLASTn [22] v.2.9.0 query versus the ribosomal reference to locate the start and end bases of their chimeric part.

5. In this step we excise the ribosomal part from the sequences.

6. Finally, we combine the reads that did not align to the ribosomal reference in step 2 with the "cured" sequences of step 5 and perform a Trinity [39] assembly on the composite fasta file.

The concept behind this pipeline is depicted in **Fig 1.** Additional information regarding the implementation of this pipeline is stated both on the github page and as comments in the ZWA script. In addition, a Docker image based on Ubuntu 20.04 has been created containing all the appropriate prerequisite tools and the ZWA script. It can be found at https://hub.docker.com/repository/docker/ndovroli/zwa/ and can be deployed on an available server or cloud, as is, maximizing performance.

Results

Building a pipeline for the unbiased identification of viruses from various species using the GenBank database we often encountered recurrent hits of specific entries regardless of the type or origin of the specimen. Using Basic Local Alignment Search Tool (BLAST) [22] these entries were found to encompass host RNA sequences, the vast majority of which were annotated as host ribosomal RNAs.

Chimeric virus-rRNA entries in GenBank

Screening GenBank for chimeric assemblies that encompassed rRNA stretches derived from various host organisms we identified 38 erroneous entries. These 38 entries were matched using three different aligners with either 28S or 18S sequences in our Ribosomal database (RiDB), constructed using the rRNAs from medically or veterinary important animal species. BBMAP (38 hits) and STAR (32 hits) appeared to be the most sensitive, while, BOWTIE2 was less efficient (9 hits) (**Fig 2**). All the above sequences were manually (BLASTn [22]) validated to encompass an rRNA stretch. These results appear to be independent of the sequencing platform used.

As chimeric genomes may arise due to the presence of chimeric reads during assembly it was hypothesized that these contaminating sequences would be the cause of premature assembly termination, and thus they were expected to map at the extremities of these "contaminated" genomes. In order to assess the position of rRNA contaminants we mapped their relative position on the sequence of the respective Genbank virus entry. As expected, the majority of the rRNA sequences were located at the extremities while, to our surprise, a significant portion was located in the middle of the viral sequences (**Fig 3**). **S1 Table** also provides a complete list of these sequences along with the starting position and length of contamination for future reference.

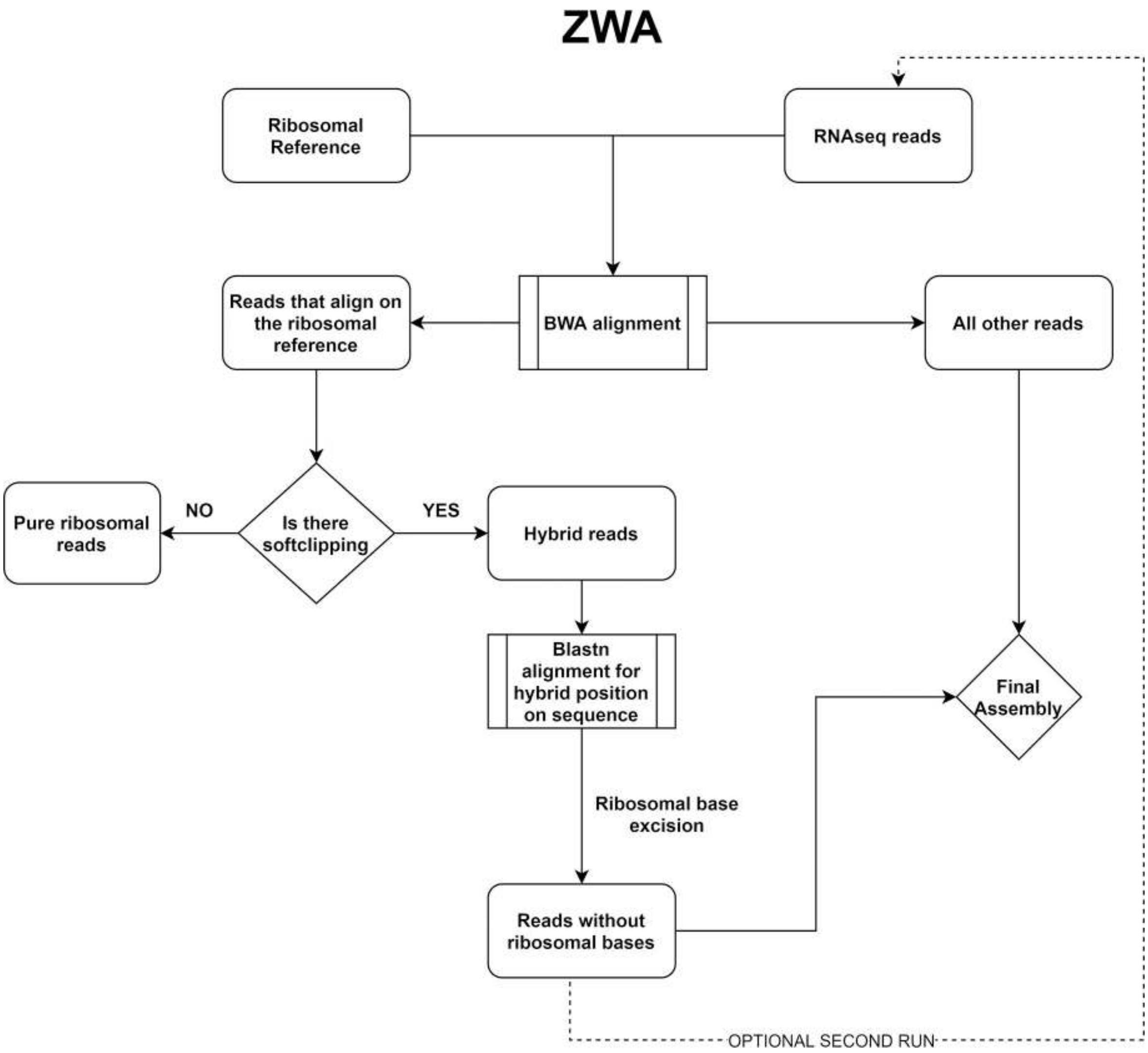

Fig 1. Flowchart of the chimeric reads clean-up pipeline.

https://doi.org/10.1371/journal.pcbi.1009304.g001

Chimeric reads as inherent problem of RNA-seq

In order to analyze the impact of chimeric reads in the *de novo* assembly of RNA viruses we used the mosquito *Anopheles sacharovi*, a known vector of human pathogens, as a paradigm. In a pool of such mosquitoes (environmental samples) we performed RNA-seq analysis of total RNA. *De novo* assembly using TRINITY resulted in a number of contigs that corresponded to viral sequences, as identified by batch BLASTx [22] analysis. The virus contigs identified, corresponded to Xanthi rhabdovirus genome (MW520377). The vast majority of the viral sequences encompassed at least one (up to 4) moiety corresponding to ribosomal RNA, usually at the extremities of the contigs. Analysis of individual reads that had passed all quality controls and mapped on chimeric contigs, revealed that chimeric reads served as bridges in the assembly of chimeric contigs (**Fig 4**). These reads were singletons and were only presented multiple times as PCR duplicates (same sequence and size products). Either in the extremities or in the middle of the contigs and even if such chimeric reads were a small fraction of the total mapped reads on any particular chimeric contig, they were detrimental to the

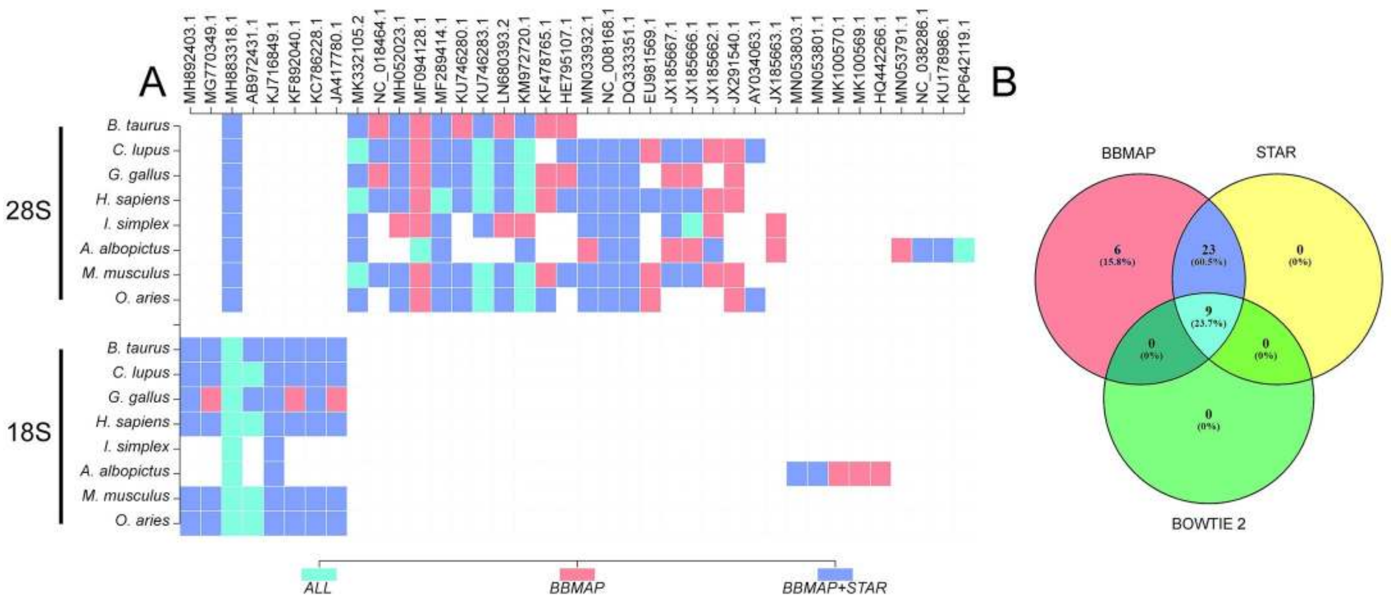

Fig 2. A) Match matrix of ribosomal hits on viral sequences. X-axis contains all the sequences from the ViDB that are chimeric and Y-axis shows the 28S and 18S rRNA sequences from our RiDB. Colours represent the aligners used to identify these chimeric (other combinations than those depicted resulted in null). B) Venn diagram of the chimeric hits identified by the 3 aligners for all host species.

https://doi.org/10.1371/journal.pcbi.1009304.g002

achievement of a correct assembly. One such chimeric contig is presented in Fig 4 depicting the effect of chimeric reads on the emergence of such a contig. The stochastic methodology of the assemblers resulted in the emergence of various such chimeric solutions in every test of the pipeline.

To identify the percentage of chimeric reads between rRNAs and a viral genome in an NGS reaction we aligned total raw reads against *Anopheles sacharovi* rRNAs and Xanthi rhabdovirus. A significant proportion of the reads that mapped to the viral genome mapped also to the rRNAs. From a total of ~3.5 million reads (quality trimmed), Xanthi rhabdovirus was represented by 1434 reads of which 836 (58%) were chimeric according to SortMeRNA or 807 (56%) according to BWA (Fig 5).

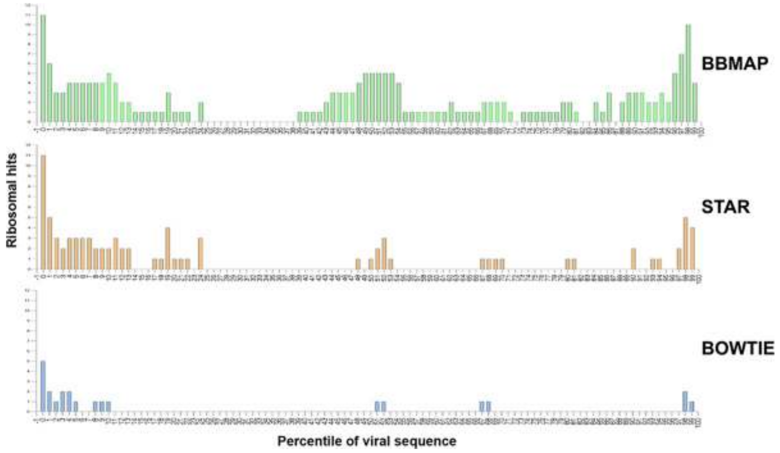

Fig 3. Alignment of sequences from the RiDB on the ViDB by all three aligners. Bar chart representation of the number of rRNA hits on viral sequence GenBank entries per percentile of the sequence. X-axis represents the percentile of viral sequence in which the ribosomal sequences start.

https://doi.org/10.1371/journal.pcbi.1009304.g003

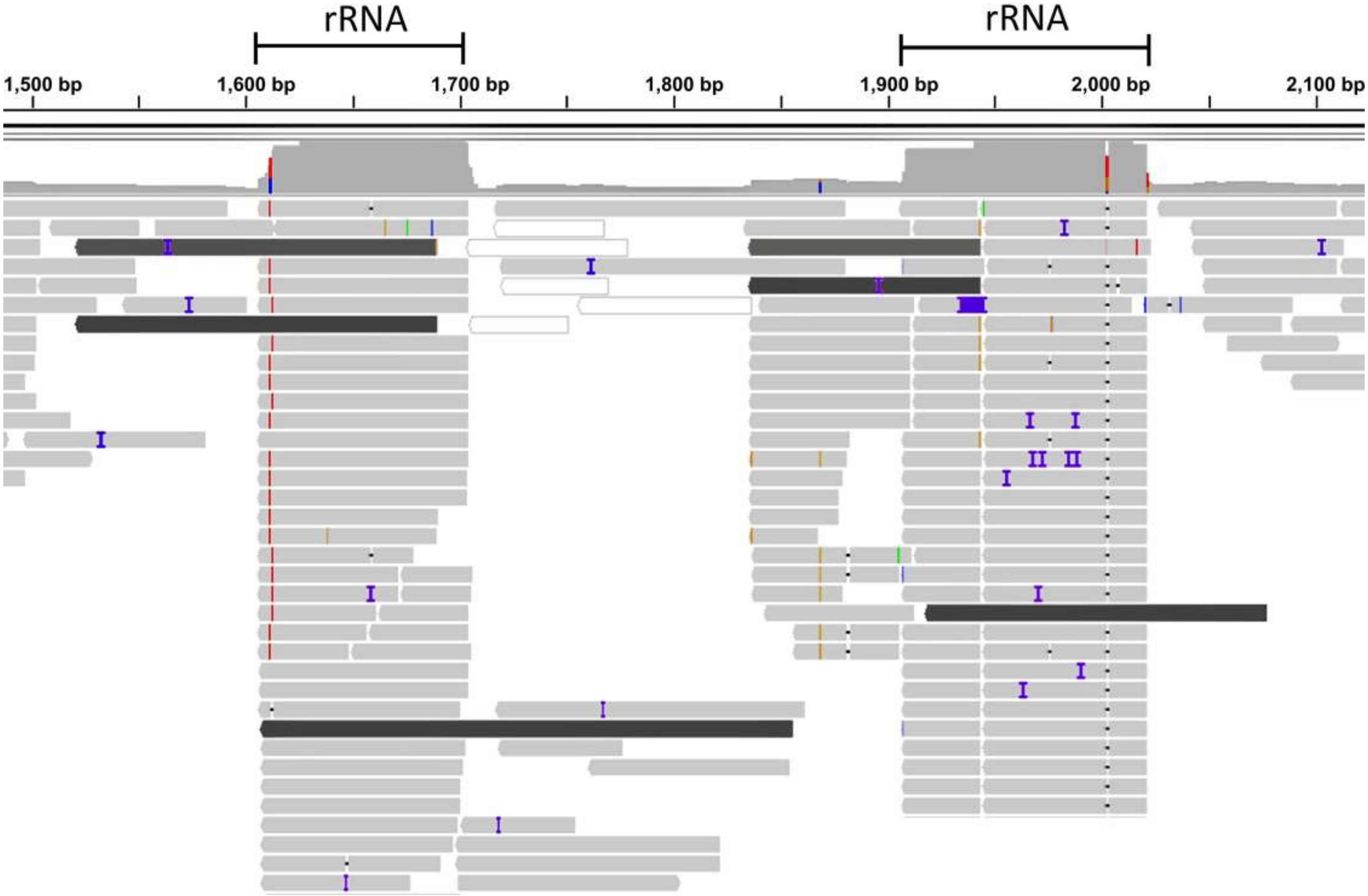

Fig 4. Example of visualization, in Integrative Genomics Viewer, of sequencing reads aligned on rRNA/viral chimeric contigs generated via de-novo assembly. Dark grey reads indicate chimeric that serve as bridges in the assembly between rRNA and viral sequences. Coloured lines represent gaps and mutations against reference sequence. Reads have been reduced for illustration purposes.

https://doi.org/10.1371/journal.pcbi.1009304.g004

Improving assembly of viruses through specific trimming of chimeric reads

De-novo assembly of viral genomes directly from raw reads of total-RNA seq very often results in the chimeric contigs such as the one presented in Fig 4 that often terminate prematurely the alignment, resulting in both contaminated and aberrant contigs. Alignment of all the contigs that derived from direct *de-novo* assembly on the Xanthi rhabdovirus genome revealed the extent of aberrancy and contamination of these contigs. Although the contigs cover a large extent of the viral genome, the incorporation of rRNA contamination is often unavoidable as the vast majority of the contigs are flanked by rRNA moieties (**Fig 6**, colored lines). The reduced mean contig length (804.3 bp) and the reduced mean coverage of the contigs (84.00%, percent of the contig that maps to the virus genome) were indicative of the detrimental effect of the contaminants in virus genome assembly (**Table 1**).

Frequent incorporation of ribosomal stretches in the *de novo* assemblies from total RNA-seq indicated the need of a pipeline that improves assemblies by removing rRNA contaminations. Current methodologies using read-filtering algorithms such as SortMeRNA resulted in the significant reduction of reads that map to the viral genome (598 reads from 1434 total) (**Table 1**). RRNA-virus chimeric reads were binned with the rest of the rRNA-containing

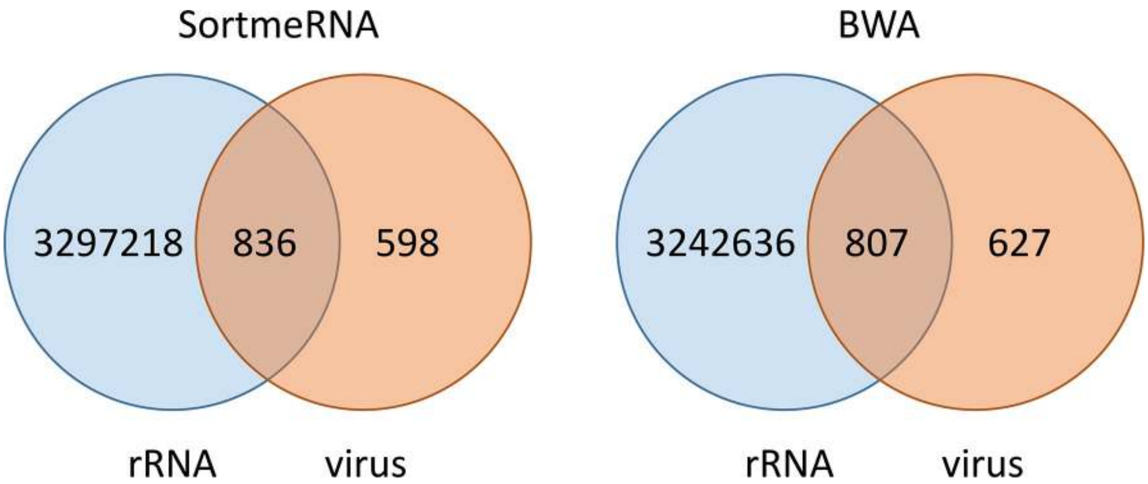

Fig 5. Venn diagrams highlighting the chimeric rRNA/viral reads produced by the NGS process on Xanthi rhabdovirus from Anopheles sacharovi sample.

https://doi.org/10.1371/journal.pcbi.1009304.g005

reads and were not used during *de-novo* assembly. On the other hand a significant number (836 using SortMeRNA and 807 using BWA) of rRNA-containing reads could partially align (non-rRNA moiety) on the viral genome signifying a great loss of genetic information. Such a loss of information was detrimental in the assembly, as although it assisted in less incorporation of rRNA moieties (89% mean coverage of contigs) the mean contig length was not improved (**Table 1**).

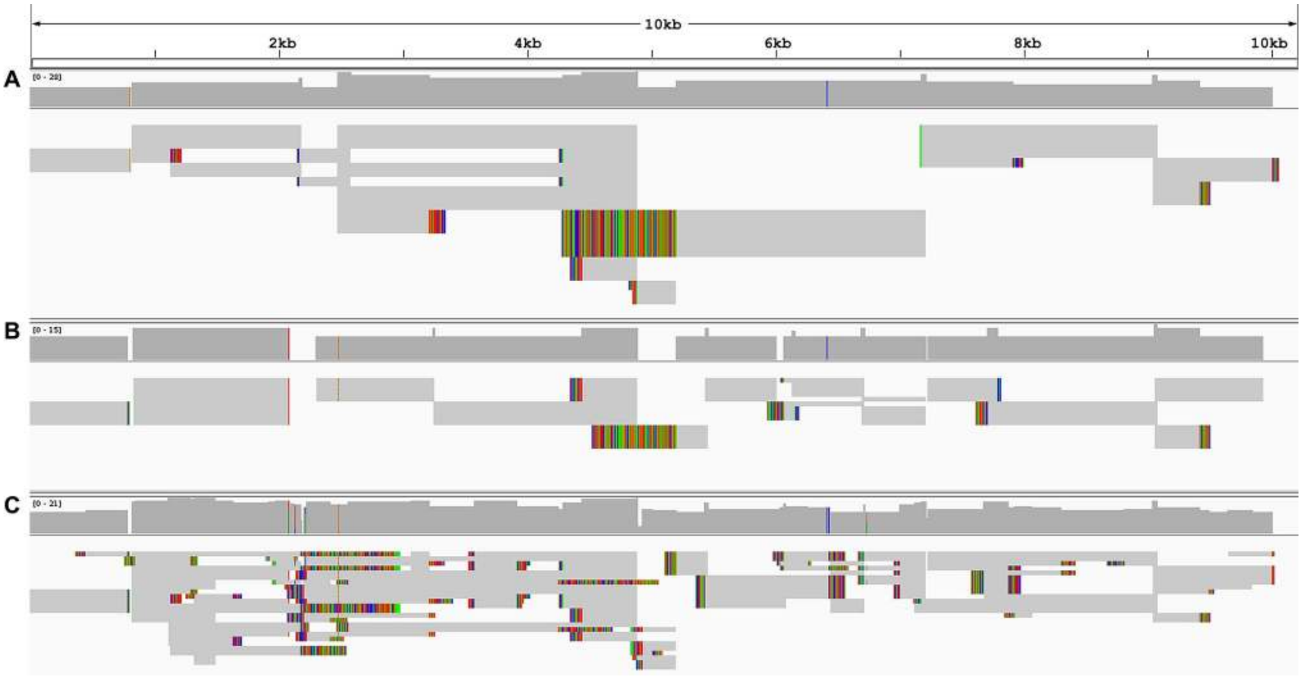

Fig 6. IGV representation of chimeric contigs between rRNA and viral sequences mapped on the Xanthi rhabdovirus genome after A) context-based trimming of reads for rRNA contaminants, B) SortMeRNA rRNA reads filtering and C) raw data assembly. Coloured lines are softclips of the contigs that do not align to the viral genome and represent chimeric moieties of the contigs.

https://doi.org/10.1371/journal.pcbi.1009304.g006

Table 1. Characteristics of the assembled contigs after ZWA, SortMeRNA and without treatment (raw reads). Mean coverage, genome coverage and the respective used reads after mapping on virus genome (± standard deviation), *n = 5 Trinity assemblies **5 Trinity assemblies.

		MAPPED ON VIRUS GENOME		
	MEAN CONTIG LENGTH* (bases)	MEAN COVERAGE OF CONTIGS* (%)	VIRUS GENOME COVERAGE** (%)	USED READS FOR ASSEMBLY**
Assembly after ZWA	1209.4 ±43	91.06 ±0.44	97.97	1335
Assembly after SortMeRNA	793.7 ±25	89.00 ±1.49	90.62	598
Assembly using raw reads	804.3 ±118	84.00 ±3.04	97.61	1434

https://doi.org/10.1371/journal.pcbi.1009304.t001

In order to make chimeric reads useful for an efficient *de novo* assembly without risking the introduction of contaminants we developed a novel pipeline; Zero Waste Algorithm, ZWA. Through this pipeline the rRNA portion of rRNA reads that partially (<90%) mapped to the rRNA database (RiDB) was context-base trimmed from these chimeric reads, leaving the non-rRNA portion for use in the assembly pipeline. During this read-cleaning process, 656 reads out of the 807 chimeric reads (BWA process) were cured from their rRNA moiety (Fig 7). However, as chimeric reads are more complicated and may incorporate more than one rRNA moiety the pipeline was run for a second round of read-cleaning during which 52 reads out of the 656 cured in the previous run were further cured from an additional rRNA moiety (Fig 7). Merging the reads that were cured through both rounds of ZWA (cured reads) with the non-rRNA reads during the initial sorting (BWA process) were able to use twice as many reads (1335) as those sorted by SortMeRNA during virus genome assembly (Table 1).

Discussion

Identification of novel viruses in clinical and environmental specimens usually employs pipelines that match RNA-seq *de-novo* assemblies to databases that contain viral sequences such as the NCBI database GenBank. However, the ever-increasing rate of submissions in Genbank and the lack of several steps of sequence integrity validation have resulted into sequence artifacts that interfere with processes such as sequence alignments and contig annotation. In an effort to curate the Genbank database ~8000 viral sequences from the 3,400,000 available viral genomes were included in the RefSeq database [43]. More specialized databases include NCBI

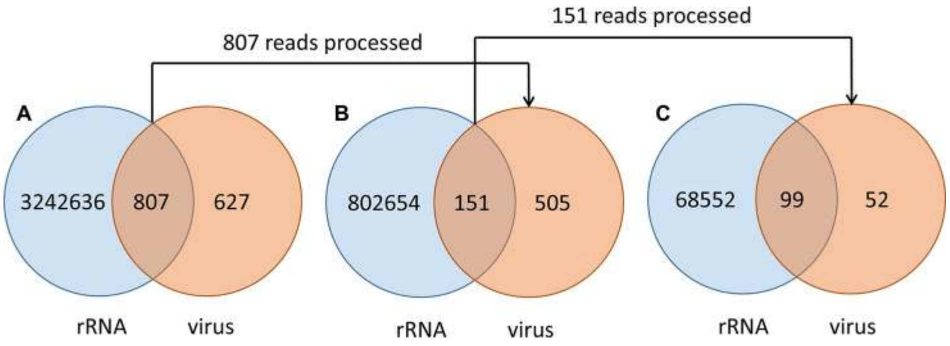

Fig 7. Ribosomal RNA read cleaning process outcome. Venn diagrams depicting (A) the initial sorting of rRNA reads and non-rRNA reads with numbers referring to rRNA reads and virus reads in the two bins. The intersection are rRNA-virus chimeric reads. (B) The rRNA reads recognized as chimeric (<90% rRNA) and the resulted cured virus reads after ZWA. The intersection are the remaining rRNA-virus chimeric reads. (C) The rRNA reads recognized in the second round as chimeric (<90% rRNA) and the resulted cured virus reads after ZWA. The intersection are the discarded remaining rRNA-virus chimeric reads.

https://doi.org/10.1371/journal.pcbi.1009304.g007

virus, EBI metagenomics, IMG/VR v.2.0, Metavir2 and iVirus. However, the vast majority of newly identified viral sequences or partial viral sequences, that are products of *de novo* assemblies through viral metagenomics screening of new hosts, are rarely included in curated databases. Alignments against GenBank are by far the most frequent pipeline when assessing viral metagenomics. Database assisted viral assembly through alignment to GenBank (like what the tool VrAP, Viral Assembly Pipeline does) [44] is also a novel methodology for more accurate virus assembly and identification that relies on GenBank integrity.

During the implementation of a pipeline for unbiased high-throughput identification of novel viruses, we came across GenBank entries that encompass ribosomal RNA moieties and came up as false positive identifications. Screening GenBank for such entries, we identified 38 submissions with significant ribosomal RNA contamination. Such contamination was mainly located at the extremities of the sequences although several entries encompassed the contamination in the middle. Chimeric assembly artifacts may mainly arise due to chimeric reads that emerge as library preparation byproducts. As the virus genome entries are increasing rapidly in sequence databases due to the wide application of viral metagenomics, there is an eminent need to avoid the generation of more contaminated entries from virome approaches. Viral genomes during virus metagenomics are often encountered in a small fraction of individuals or are diluted out in environmental samples. For example, genomes of human pathogens such as West-Nile and Ross River virus are present in less than 1% of vectors such as mosquitoes [45,46], while analysis of just 10 L of sea water in China yielded 4,593 nearly full-length RNA virus polymerase genes [47].

In order to refine the assembly process and reduce the generation of contaminated viral genomes in such approaches we assessed the effect of host rRNA-virus chimeric reads in the assembly of low abundance virus genomes. For this reason, in a working example we used a pool of *Anopheles sacharovi* mosquitoes, an important human pathogen vector [48] sequenced using total RNA-seq, for the identification of novel RNA virus genomes. During the analysis we noted that a significant fraction (half) of reads that mapped to Xanthi rhabdovirus genome, a virus previously identified in this species, contained also an rRNA moiety.

Previous attempts to minimize rRNA contamination during total RNA-seq included poly (A) selection and ribodepletion before library preparation or computational filtering of reads that mapped on ribosomal RNAs [24–30,49]. Poly(A) selection is prohibited during virus metagenomics as several virus families do not contain a poly(A) tail in their genomic RNA [50], while ribodepletion is not always efficient as it is species specific [51] and the organism or pool of organisms (environmental samples) during virome analysis may vary significantly. On the other hand, bioinformatics pipelines efficiently filter reads that map to the host rRNAs removing them from the assembly procedure. Such pipelines are part of current the state-of-the-art methodologies used for RNA virus genome assembly, while in others no rRNA filtering was performed prior to assembly [47,52,53]. Applying such a filtering procedure using SortMeRNA, a dedicated software for ribosomal RNA sorting, we observed that the algorithm would process rRNA-virus chimeric reads together with the rRNA reads, binning them accordingly. Although such a procedure generated cleaner contigs during assembly, the significant reduction of virus reads (half) had detrimental effect in obtaining important contigs for the assembly of the complete virus genome.

In an attempt to feed back the filtered out chimeric reads in the assembly pipeline we developed a novel pipeline that would cure rRNA-containing chimeric reads using a context-based trimming algorithm; as opposed to quality-based trimming. The ZWA algorithm, using a BLASTn-directed procedure, was able to remove accurately the rRNA moieties from the chimeric reads rendering them useful for the enhancement of the *de novo* assembly efficiency. This is contrary to other past algorithms which trim reads using a virus reference guided approach that rely on existing virus databases, rendering such algorithms inefficient in virus metagenomics and the identification of novel virus genomes [54]. The ZWA algorithm returned the majority of the initially discarded reads, without their rRNA bases, into the assembly of the virus genome achieving higher genome coverage with bigger and cleaner contigs. The ZWA approach resulted overall in a smoother and less wasteful assembly procedure superior to current approaches that involve rRNA filtering algorithms.

A limitation of this procedure is the possibility to miss naturally occurring RNA virus recombinant genomes that have incorporated small fragments of ribosomal RNAs. Such a very rare event has been presented twice in the literature reporting a presumable incorporation of 54nt and 15nt fragments of ribosomal RNA (28S rRNA) in influenza virus [55] and poliovirus [56] genomes respectively. Since then, there has been no other similar report. Of note, these 2 genomes are not included in GenBank so our search algorithm did not identify them. However, assembly of such naturally occurring recombinant genomes should present a clear break in the recombination site after the application of the ZWA algorithm, as no other bridging read would be present in the sample.

Conclusions

Growing needs to screen habitats for potential emerging viral pathogens combined with the increasing affordability and the reduced complexity of NGS procedures increase the importance of maintaining the integrity and reliability of virus databases. Our work highlights an aspect in the plethora of NGS artifacts and proposes a pipeline that significantly reduces contaminated and aberrant viral genome assemblies by implementing a precise upcycling of previously discarded chimeric reads.

Supporting information

S1 Table. GenBank Accession IDs of virus sequences containing ribosomal sub-sequences along with the starting position and length of the moieties.
(DOCX)

S1 Fig. Alignment of mock human 28S rRNA reads on the human 28S reference sequence (M11167.1) for validation. X-axis shows the bp positions of the reference and Y-axis shows how many reads of the mock file were aligned on each position.
(DOCX)

Author Contributions

Conceptualization: Ioannis Karakasiliotis.

Data curation: Nikolas Dovrolis.

Formal analysis: Nikolas Dovrolis.

Funding acquisition: Ioannis Karakasiliotis.

Investigation: Katerina Kassela, Konstantinos Konstantinidis, Adamantia Kouvela.

Methodology: Nikolas Dovrolis, Ioannis Karakasiliotis.

Project administration: Ioannis Karakasiliotis.

Resources: Stavroula Veletza.

Software: Nikolas Dovrolis.

Supervision: Ioannis Karakasiliotis.

Visualization: Nikolas Dovrolis, Ioannis Karakasiliotis.

Writing – original draft: Nikolas Dovrolis, Ioannis Karakasiliotis.

Writing – review & editing: Nikolas Dovrolis, Stavroula Veletza, Ioannis Karakasiliotis.

References

1. Bibby K. Metagenomic identification of viral pathogens. Trends in biotechnology. 2013; 31(5):275–9. https://doi.org/10.1016/j.tibtech.2013.01.016 PMID: 23415279
2. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. NUCLEIC ACID RES. 2015; 43(D1):D571–D7. https://doi.org/10.1093/nar/gku1207 PMID: 25428358
3. Woo PC, Lau SK, Chu C-m, Chan K-h, Tsoi H-w, Huang Y, et al. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. Journal of virology. 2005; 79(2):884–95. https://doi.org/10.1128/JVI.79.2.884-895.2005 PMID: 15613317
4. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nature Medicine. 2020 2020/09/01; 26(9):1405–10. https://doi.org/10.1038/s41591-020-0997-y PMID: 32678356
5. Wang D. 5 challenges in understanding the role of the virome in health and disease. PLoS Pathogens. 2020; 16(3):e1008318. https://doi.org/10.1371/journal.ppat.1008318 PMID: 32214384
6. Garmaeva S, Sinha T, Kurilshikov A, Fu J, Wijmenga C, Zhernakova A. Studying the gut virome in the metagenomic era: challenges and perspectives. BMC biology. 2019; 17(1):84. https://doi.org/10.1186/s12915-019-0704-y PMID: 31660953
7. Li B, Si H-R, Zhu Y, Yang X-L, Anderson DE, Shi Z-L, et al. Discovery of bat coronaviruses through surveillance and probe capture-based next-generation sequencing. Msphere. 2020; 5(1).
8. Moreira D, López-García P. Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol. 2009; 7(4):306–11. https://doi.org/10.1038/nrmicro2108 PMID: 19270719
9. Koonin EV, Senkevich TG, Dolja VV. Compelling reasons why viruses are relevant for the origin of cells. Nat Rev Microbiol. 2009; 7(8):615–. https://doi.org/10.1038/nrmicro2108-c5 PMID: 19561624
10. Cantalupo PG, Pipas JM. Detecting viral sequences in NGS data. Current opinion in virology. 2019; 39:41–8. https://doi.org/10.1016/j.coviro.2019.07.010 PMID: 31465960
11. Jha PK, Sarkar S. DNA sequencing and comparative sequence analysis reveal that the Escherichia coli genomic DNA may replace the target DNA during molecular cloning: evidence for the erroneous assembly of E. coli DNA into database sequences. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology. 1997; 118(2):333–9. https://doi.org/10.1016/s0305-0491(97)00175-2 PMID: 9440226
12. Dittami SM, Corre E. Detection of bacterial contaminants and hybrid sequences in the genome of the kelp Saccharina japonica using Taxoblast. PeerJ. 2017; 5:e4073. https://doi.org/10.7717/peerj.4073 PMID: 29158994
13. Pible O, Hartmann EM, Imbert G, Armengaud J. The importance of recognizing and reporting sequence database contamination for proteomics. EuPA Open Proteomics. 2014; 3:246–9.
14. Zolfo M, Pinto F, Asnicar F, Manghi P, Tett A, Bushman FD, et al. Detecting contamination in viromes using ViromeQC. Nature biotechnology. 2019; 37(12):1408–12. https://doi.org/10.1038/s41587-019-0334-5 PMID: 31748692
15. Chu H-T, Hsiao WW, Chen J-C, Yeh T-J, Tsai M -H, Lin H, et al. EBARDenovo: highly accurate de novo assembly of RNA-Seq with efficient chimera-detection. Bioinformatics. 2013; 29(8):1004–10. https://doi.org/10.1093/bioinformatics/btt092 PMID: 23457040
16. Kerkvliet J, de Fouchier A, van Wijk M, Groot AT. The Bellerophon pipeline, improving de novo transcriptomes and removing chimeras. Ecology and evolution. 2019; 9(18):10513–21. https://doi.org/10.1002/ece3.5571 PMID: 31624564
17. Sangiovanni M, Granata I, Thind AS, Guarracino MR. From trash to treasure: detecting unexpected contamination in unmapped NGS data. BMC bioinformatics. 2019; 20(4):1–12. https://doi.org/10.1186/s12859-019-2684-x PMID: 30999839
18. Rachtman E, Balaban M, Bafna V, Mirarab S. The impact of contaminants on the accuracy of genome skimming and the effectiveness of exclusion read filters. Molecular Ecology Resources. 2020; 20(3). https://doi.org/10.1111/1755-0998.13135 PMID: 31943790

19. Goig GA, Blanco S, Garcia-Basteiro AL, Comas I. Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. BMC biology. 2020; 18(1):1–15. https://doi.org/10.1186/s12915-019-0728-3 PMID: 31898513

20. Arroyo Mühr LS, Lagheden C, Hassan SS, Kleppe SN, Hultin E, Dillner J. De novo sequence assembly requires bioinformatic checking of chimeric sequences. Plos one. 2020; 15(8):e0237455. https://doi.org/10.1371/journal.pone.0237455 PMID: 32777809

21. Francois CM, Durand F, Figuet E, Galtier N. Prevalence and implications of contamination in public genomic resources: a case study of 43 reference arthropod assemblies. G3: Genes, Genomes, Genetics. 2020; 10(2):721–30.

22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

23. Fierst JL, Murdock DA. Decontaminating eukaryotic genome assemblies with machine learning. BMC bioinformatics. 2017; 18(1):533. https://doi.org/10.1186/s12859-017-1941-0 PMID: 29191179

24. Czajkowski MD, Vance DP, Frese SA, Casaburi G. GenCoF: a graphical user interface to rapidly remove human genome contaminants from metagenomic datasets. Bioinformatics. 2019; 35(13):2318–9. https://doi.org/10.1093/bioinformatics/bty963 PMID: 30475995

25. Zhou Q, Su X, Jing G, Chen S, Ning K. RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC genomics. 2018; 19(1):1–10. https://doi.org/10.1186/s12864-017-4368-0 PMID: 29291715

26. Pérez-Rubio P, Lottaz C, Engelmann JC. FastqPuri: high-performance preprocessing of RNA-seq data. BMC bioinformatics. 2019; 20(1):1–11. https://doi.org/10.1186/s12859-018-2565-8 PMID: 30606105

27. Lafond-Lapalme J, Duceppe M-O, Wang S, Moffett P, Mimee B. A new method for decontamination of de novo transcriptomes using a hierarchical clustering algorithm. Bioinformatics. 2017; 33(9):1293–300. https://doi.org/10.1093/bioinformatics/btw793 PMID: 28011783

28. Wang Y, Hu H, Li X. rRNAFilter: A Fast Approach for Ribosomal RNA Read Removal Without a Reference Database. Journal of Computational Biology. 2017; 24(4):368–75. https://doi.org/10.1089/cmb.2016.0113 PMID: 27610931

29. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research. 2018; 7. https://doi.org/10.12688/f1000research.15931.2 PMID: 30254741

30. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012; 28(24):3211–7. https://doi.org/10.1093/bioinformatics/bts611 PMID: 23071270

31. Ren L-L, Wang Y-M, Wu Z-Q, Xiang Z-C, Guo L, Xu T, et al. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chinese medical journal. 2020. https://doi.org/10.1097/CM9.0000000000000722 PMID: 32004165

32. Kans J. Entrez direct: E-utilities on the UNIX command line. Entrez Programming Utilities Help [Internet]: National Center for Biotechnology Information (US); 2020.

33. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

35. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

36. Swift ML. GraphPad prism, data analysis, and scientific graphing. Journal of chemical information and computer sciences. 1997; 37(2):411–2.

37. Oliveros J. VENNY. An interactive tool for comparing lists with Venn Diagrams. h ttp. bioinfogp cnb csic es/tools/venny/index html. 2007.

38. Ihaka R, Gentleman R. R: a language for data analysis and graphics. Journal of computational and graphical statistics. 1996; 5(3):299–314.

39. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology. 2011; 29(7):644. https://doi.org/10.1038/nbt.1883 PMID: 21572440

40. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013; 14(2):178–92. https://doi.org/10.1093/bib/bbs017 PMID: 22517427

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

42. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics. 2009; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

43. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. NUCLEIC ACID RES. 2007; 35(suppl_1): D61–D5. https://doi.org/10.1093/nar/gkl842 PMID: 17130148

44. Hölzer M, Marz M. Software dedicated to virus sequence analysis “bioinformatics goes viral”. Advances in Virus Research: Elsevier; 2017. p. 233–57.

45. Batovska J, Lynch S, Cogan N, Brown K, Darbro J, Kho E, et al. Effective mosquito and arbovirus surveillance using metabarcoding. Molecular ecology resources. 2018; 18(1):32–40. https://doi.org/10.1111/1755-0998.12682 PMID: 28417591

46. Rudolf I, Betášová L, Blažejová H, Venclíková K, Straková P, Šebesta O, et al. West Nile virus in overwintering mosquitoes, central Europe. Parasites & vectors. 2017; 10(1):1–4.

47. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nature microbiology. 2020; 5(10):1262–70. https://doi.org/10.1038/s41564-020-0755-4 PMID: 32690954

48. Hertig E. Distribution of Anopheles vectors and potential malaria transmission stability in Europe and the Mediterranean area under future climate change. Parasites & vectors. 2019; 12(1):18. https://doi.org/10.1186/s13071-018-3278-6 PMID: 30621785

49. Fang N, Akinci-Tolun R. Depletion of Ribosomal RNA Sequences from Single-Cell RNA-Sequencing Library. Current protocols in molecular biology. 2016; 115(1):7.27. 1–7. 0. https://doi.org/10.1002/cpmb.11 PMID: 27366895

50. Chaitanya K. Structure and Organization of Virus Genomes. Genome and Genomics: Springer; 2019. p. 1–30.

51. Kim IV, Ross EJ, Dietrich S, Döring K, Alvarado AS, Kuhn C-D. Efficient depletion of ribosomal RNA for RNA sequencing in planarians. BMC genomics. 2019; 20(1):1–12. https://doi.org/10.1186/s12864-018-5379-1 PMID: 30606130

52. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell host & microbe. 2020; 28(5):724–40. e8. https://doi.org/10.1016/j.chom.2020.08.003 PMID: 32841606

53. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, et al. Redefining the invertebrate RNA virosphere. Nature. 2016; 540(7634):539–43. https://doi.org/10.1038/nature20167 PMID: 27880757

54. Berthet N, Descorps-Declère S, Nkili-Meyong AA, Nakouné E, Gessain A, Manuguerra J-C, et al. Improved assembly procedure of viral RNA genomes amplified with Phi29 polymerase from new generation sequencing data. Biological research. 2016; 49(1):1–8. https://doi.org/10.1186/s40659-016-0099-y PMID: 27605096

55. Khatchikian D, Orlich M, Rott R. Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus. Nature. 1989; 340(6229):156–7. https://doi.org/10.1038/340156a0 PMID: 2544809

56. Charini WA, Todd S, Gutman GA, Semler BL. Transduction of a human RNA sequence by poliovirus. Journal of virology. 1994; 68(10):6547–52. https://doi.org/10.1128/JVI.68.10.6547-6552.1994 PMID: 8083991