Chapter 12

Analysis of *Leishbuviridae* from Trypanosomatids

Danyil Grybchuk, Alexei Yu Kostygov, and Vyacheslav Yurchenko

Abstract

Over the last decade, considerable progress has been made in unraveling RNA virus diversity. This has contributed to our understanding of the evolution of these viruses, which include emerging zoonotic human pathogens. Current success has been greatly facilitated by the development of next-generation sequencing platforms instrumental for meta-transcriptomic studies. However, due to the rapid evolution of RNA viruses, there are numerous "blind spots" waiting to be explored; one of those is the RNA virome of unicellular eukaryotes. Here, we present the pipeline, which has been successfully used to characterize various types of RNA viruses, including *Leishbuviridae* (*Bunyaviricetes, Hareavirales*) in the parasitic flagellates of the family Trypanosomatidae. The pipeline relies on axenic in vitro cell culture and double-stranded RNA enrichment, followed by direct RNA-sequencing. A detailed procedure description starting from the initial total RNA preparation to the final assembly of the viral segments is provided.

Key words Trypanosomatidae, *Leishbuviridae*, Protists, Virus Discovery, RNA isolation, dsRNA, NGS

1 Introduction

Transcription and replication of RNA viruses, regardless of genome strandedness and polarity, requires production of both "+" and "−" RNA strands in the cell cytoplasm by viral RNA-dependent RNA polymerase (RdRp) [1–4]. This trait of RNA viruses was intensely studied half-a-century ago as a curious deviation from the Central dogma of molecular biology. At the core of this research were chemical and enzymatic methods allowing to distinguish single-stranded (ss) and double-stranded (ds) RNA species. Differential precipitation of ssRNA and dsRNA with 2 M and 5 M lithium chloride (LiCl), respectively, was first reported by David Baltimore as part of his Nobel Prize winning studies on the

The original version of the chapter has been revised. A correction to this chapter can be found at https://doi.org/10.1007/978-1-0716-4338-9_22

Hani Boshra (ed.), *Bunyaviruses: Methods and Protocols*, Methods in Molecular Biology, vol. 2893,
https://doi.org/10.1007/978-1-0716-4338-9_12,

replication of poliovirus [5]. Soon after that, the ssRNA/ssDNA-specific S1 nuclease was first used to reveal the role of terminal secondary structures in replication of *Mammalian orthoreovirus 3*, MRV-3 [6] and LiCl fractionation of dsRNA was optimized for bulk detection of plant viruses [7]. Both LiCl- and S1 nuclease-based approaches were actively used till the early 2000s for the search of fungal dsRNA viruses fueled, in part, by the interest in potential biocontrol agents for phytopathogenic fungi [8, 9]. For example, LiCl was used to detect a partitivirus and a mitovirus in *Ophiostoma* spp., a causative agent of the Dutch elm disease [10, 11], while S1 nuclease-based approach yielded the discovery of an endornavirus in hypovirulent strain of *Helicobasidium mompa* (violet root rot) [12] and numerous divergent RNA mycoviruses in various fungi [13–16].

Even though negative-sense (−)ssRNA viruses do not accumulate dsRNA in the cell to avoid destruction by cellular RNA interference systems [17], their "+" and "−" RNAs recombine in vitro upon nucleic acid extraction. This phenomenon allows all ssRNA viruses to be detected by gel electrophoresis of S1 nuclease- and/or LiCl-treated total RNA samples, albeit with lower sensitivity as compared to dsRNA viruses. Combined with the observation that protist and fungal RNA viruses are often cryptic and persist in low copy numbers [18–21], the approaches introduced above require high amount of total RNA for analysis.

Individual viral bands can be cut out from the agarose gel and amplified by reverse-transcription (RT)-PCR. There are three methods for the viral RNA amplification, which differ in RNA pretreatment and the type of oligonucleotides used for the first strand cDNA synthesis: (i) no pretreatment and RT-PCR using specific primer with random six nucleotides at the 3′ end; (ii) *Escherichia coli* poly(A) polymerase treatment and RT-PCR using specific primer with 6–8 Thymine nucleotides at the 3′ end [22]; and (iii) attachment of an oligonucleotide using the T4 RNA ligase followed by RT-PCR with a complementary specific primer [23] or ligation of a self-priming hairpin oligonucleotide [24]. Methods (ii) and (iii) are suitable for the recovery of the complete genomic segment(s), which can be then sequenced conventionally by "primer walking" [25]. However, they rely on the free hydroxyl group at the 3′ end and, thus, are not suitable for viruses, which have modifications at this position. Conversely, method (i) does not depend on the 3′ end chemistry but yields only partial sequences.

Nowadays, the wide availability of next-generation sequencing (NGS) methods makes such RNA-amplification methods redundant. Moreover, modern bioinformatic software, such as Trinity [26] and SPAdes [27], allow de novo assembly of viral genomes with no prior knowledge about them. Therefore, these tools are applicable not only for the detection of known pathogens but also for the discovery of novel RNA viruses. Over the last decade, there was a burst of studies reporting thousands of sequences of RNA

viruses in meta-transcriptomes across various biotopes and ecosystems [28–31]. These data significantly expanded our understanding of diversity and evolution of RNA viruses but suffered from one considerable drawback: they provided no information on the viral hosts.

A typical RNA-seq data analysis pipeline includes the following steps: raw read quality control and trimming, mapping and filtering of trimmed reads followed by their assembly (reference-based or de novo) into contigs (transcripts and/or viral genomic segments), identification and evaluation of their relative abundance within and between datasets [32, 33]. Quality control software performs a series of analyses to identify known problems in the read data, while trimming programs, such as Trimmomatic [34], solve these problems by discarding reads with overall high sequencing error rate, deleting low-quality bases from the 3′ termini, and removing adaptor sequences introduced at the 5′ ends of each read as a part of the sequencing procedure. Next, the "clean" reads are mapped onto the reference genome if one is available. Routinely, this is done with Bowtie2 and SAMtools bundle [35, 36] or BBtools [37]. For eukaryotic organisms, this step is crucial to determine exon–intron boundaries and untranscribed regions [38, 39]. The advantage of working with RNA viruses is that their genomes are usually simple and their genes are distinct from those of the hosts [40]. As such, the initial mapping can be useful to set apart viral- from host reads to speed up further assembly and minimize the probability of chimeric sequence appearance. The assembly of viral sequences is usually done in the reference-free (de novo) mode [26, 27], since appropriate reference is absent in most cases. The resulting assembled sequences (contigs) usually contain the complete genomic segment(s) of a virus. Some gaps and fragmentation may occur if viruses are scarce and/or contain secondary structures.

The identification of viral contigs in the databases is done via amino acid sequence homology search. Since the input data are in nucleotides, this requires translating contigs in all six frames and comparing each translation product against the database using BLASTx first realized in BLAST+ software [41]. This procedure is computationally intense and requires both an efficient software specifically designed to work with genomic/transcriptomic data, such as DIAMOND [42], and use of the clustered databases, where searches are performed only against a subset of representative sequences with a given all-to-all amino acid identity threshold, for example, UniRef [43]. In addition, a more sensitive search for viral proteins can be done using the hidden Markov model approach that compares query proteins against a profile build from an amino acid sequence alignment of a specific set of homologous proteins [44, 45]. For RNA viruses, the best candidate is RdRp [46, 47].

Another piece of information retrieved from the assembled data is the coverage of viral contigs (i.e., how many reads went into constructing a given contig). This value may help to estimate

abundance of viral RNAs relative to each other or even to the host transcriptome, if total RNA was sequenced. In order to calculate coverage, the reads are first aligned back to the assembled contigs using the mapping software mentioned above. Then, the number of reads aligned to each contig is counted using a custom script. To facilitate comparison within a single sequencing experiment or between different ones, the read count for each contig is normalized per length of the latter in kilobases and per million reads. For single-end reads, the coverage value is called RPKM (reads per kilobase per million). For paired-end reads, a complementary pair of read mates aligned to the same sequence is counted as one read and the resultant value is called FPKM (fragments per kilobase per million). This allows comparison between single- and paired-end sequencing experiments [32].

In this chapter, we describe a pipeline for molecular characterization of novel viruses of parasitic flagellates belonging to the family Trypanosomatidae. It has been used to document numerous groups of RNA viruses including -ssRNA viruses currently classified as the family *Leishbuviridae* (class Bunyaviricetes) [48–52]. The pipeline starts with an axenic (ideally, clonal) culture of a flagellate, which removes any ambiguity in respect to the host of studied viruses. It proceeds with the viral dsRNA detection by enzymatic (DNase I/ S1 nuclease) enrichment. For this purpose, total RNA is isolated from the late log-phase culture to ensure that (i) viral genome is actively transcribing and replicating to produce comparable amounts of + and − RNA strands and (ii) the host cell density is high enough to maximize the total RNA yield. Finally, we provide a full bioinformatic workflow to assemble and identify RNA viruses of trypanosomatids.

2 Materials

2.1 Total RNA Isolation

1. Phosphate buffer saline, PBS.
2. Insulin syringe (optional).
3. TRIzol (or TRI Reagent) for RNA extraction.
4. Chloroform.
5. iPrOH.
6. 70% EtOH.
7. Deproteination solution (optional, *see* **Note 2**): 50 mM Tris-HCl pH 9.3, 1% 2-mercaptoethanol.
8. Nuclease-free water.
9. Multipurpose refrigerated centrifuge.
10. Refrigerated microcentrifuge.

2.2 dsRNA Enrichment for Screening and NGS

1. DNase I at 2 units/μL.
2. S1 nuclease from *Aspergillus oryzae* at 100 u/μL.
3. 200 mM EDTA.
4. ssRNA precipitating solution (optional, *see* **Note 4**): 3.3 M LiCl, 250 mM NaCl, 25 mM Tris-HCl pH 8.
5. 3 M NaOAc (optional, *see* **Note 4**).
6. 96% EtOH (optional, *see* **Note 4**).
7. Heating block for 1.5-mL Eppendorf tubes.
8. RNA purification kit allowing elution in a small volume.

2.3 Gel Electrophoresis

1. 15–25 cm long 0.8% agarose gel prepared with 1× TAE buffer.
2. Horizontal electrophoresis chamber, power supply.
3. 1× TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA)
4. Ethidium bromide 0.5 μg/mL solution in distilled water (or an alternative fluorescent dye).
5. Gel documentation system.

2.4 NGS Sequencing and Bioinformatic Analysis

1. Hardware: Workstation with 16 CPU cores and 64 GB RAM, Linux OS.
2. Software: FastQC, Trimmomatic (0.40), Bowtie2 (2.4.4), SAMtools (1.13), Trinity (2.4.0), DIAMOND (2.0.14), BLAST+ (2.13.0).

3 Methods

3.1 Total RNA Isolation

1. Harvest at least 10^8 trypanosomatid cells from a late log-phase axenic culture by centrifugation at 2500 × g for 15 min at 4 °C.
2. Resuspend the pellet in sterile 1 × PBS and centrifuge again as above.
3. Resuspend the pellet in 1 mL of TRIzol, transfer the suspension into a 2-mL Eppendorf tube, and break any clumps by pipetting (*see* **Note 1**).
4. Place samples at −80 °C for later processing or proceed to the next step.
5. Add 0.5 mL (1:2 of the added TRIzol volume) of 100% chloroform and shake vigorously.
6. Centrifuge at 20,000 × g for 15 min at 4 °C in a microcentrifuge.
7. Transfer about 700 μL of the upper phase into a new 1.5-mL Eppendorf tube (*see* **Note 2**).
8. Add 1 volume (about 700 μL) of prechilled at −20 °C iPrOH and mix gently by inverting the tube.

9. Incubate at −20 °C for at least 2 h, better overnight.
10. Centrifuge at 20,000 × g for 15 min at 4 °C in a microcentrifuge.
11. Discard the supernatant, add 500 μL of 70% ethanol to the pellet and centrifuge again.
12. Repeat the **step 11** with 150 μL of 70% ethanol.
13. Remove the liquid, dry the pellet, dissolve it in 25 μL of nuclease-free water, and measure the concentration (*see* **Note 3**).
14. Store at −80 °C for up to 1 month or proceed to the next step (Subheading 3.2).

3.2 DsRNA Enrichment for Screening and NGS

1. Take about 50 μg of total RNA in 21.5 μL. If the concentration is substantially lower or more RNA is needed for the analysis to increase sensitivity, split the sample and process several 21.5 μL-aliquots in parallel.
2. Add 1 μL of DNase I and 2.5 μL of 10× DNAse Reaction Buffer.
3. Incubate at 37 °C for 60 min and proceed to the next steps without heat inactivation.
4. Add 0.5 μL of S1 nuclease directly into the DNase mix.
5. Incubate at 37 °C for 60 min (*see* **Note 4**).
6. Add 0.5 μL 200 mM EDTA and incubate at 70 °C for 10 min for the enzyme inactivation.
7. At this stage, the dsRNA can be visualized in agarose gel (Subheading 3.3) or purified for NGS.
8. For NGS, purify dsRNA by the RNA purification kit, elute in 15 μL of nuclease-free water preheated to 50 °C (*see* **Note 5**).
9. Check 1–5 μL of resultant dsRNA prep by agarose gel electrophoresis.
10. Use purified dsRNA for Illumina paired-end RNA sequencing with TruSeq Stranded library prepared using random hexamers.

3.3 DsRNA Visualization

1. Add 0.2–1 μL of 6× loading dye to 1–5 μL of the kit-purified dsRNA prep (Subheading 3.2, **step 8**) or 5 μL of DNase I/ S1 nuclease-treated total RNA (Subheading 3.2, **step 7**), mix, and load onto a 0.8% agarose gel.
2. Run the gel for at least 1.5 h at maximum voltage of 5 V/cm for proper band separation (*see* **Note 6**).
3. Stain the gel in 0.5 μg/mL ethidium bromide solution for 15 min (*see* **Note 7**).

4. Destain the gel in pure distilled water for 15 min (*see* **Note 8**).
5. Analyze the band patterns using a gel documentation system.

3.4 NGS Sequencing Data Processing and Bioinformatic Analyses

1. Sequence dsRNA to the minimal depth of 3 Gb. Check the quality of the reads using FastQC.
2. Trim the reads and sort them into left and right paired/unpaired with Trimmomatic. The software requires the file with adaptor sequences, which are available from the sequencing company (for example, for the Illumina paired-end reads the file is called TruSeq3-PE-2.fa). Trimmomatic can read .gz files, thus, it is not necessary to unpack them beforehand. The typical command line for trimming of paired-end RNA-seq Illumina is:

```
trimmomatic PE -threads 16 -phred33 <yourdata>_1.fastq.gz
<yourdata>_2.fastq.gz -baseout <yourdata>_trimmed ILLUMINA-
CLIP:TruSeq3-PE-2.fa:2:20:10 LEADING:3 TRAILING:3 SLIDINGWIN-
DOW:4:15 MINLEN:50
```

   where **PE** is a specification of paired-end reads, **-threads** is the number of CPU cores, **-phred33** is type of quality score used by the sequencing company, ***_fastq.gz** are left and right reads provided by the sequencing company, **-baseout** is the basename of the output files, of which there will be four: left paired (1P), right paired (2P), left unpaired (1U), where the right mate did not pass the quality control, and right unpaired (2U), where the left mate was dropped (the exact description of the other parameters can be found in the Trimmomatic manual).
3. If the reference genome of the host organism is available, it is highly recommended to decontaminate the dataset from eukaryotic reads before the assembly. This not only speeds up the whole procedure but also often results in cleaner virus assemblies. Otherwise, direct de novo assembly of the entire dataset is also possible. For direct assembly skip to **step 6**.
4. Map the trimmed reads (**step 2**) onto the host genome using Bowtie2. This step requires a reference genome sequence in fasta format, which can be downloaded from the NCBI website by browsing available genomes. First, the reference genome is indexed with a command:

```
bowtie2-build <yourgenome>.fasta <yourgenome>
```

   Then mapping is performed using:

```
bowtie2 -x <yourgenome> -1 <yourdata>_trimmed_1P -2 <yourda-
ta>_trimmed_2P -U <yourdata>_trimmed_1U,<yourdata>_trimmed_2U
```

```
--end-to-end --very-sensitive -p 16 -S reads_to_<yourgenome>.
sam 2> reads_to_<yourgenome>_bowtie2log.txt
```

where option **-x** specifies the base-name of index files created by the previous command in this step, **-1**, **-2**, and **-U** specifies trimmed reads generated in **step 2** (note that unpaired reads are provided under a single parameter and their filenames are separated with a comma without a whitespace), **--end-to-end** is the alignment mode, in which the entire read length is considered as opposed to local alignment, where read terminals can be disregarded, **--very-sensitive** is a preset of alignment parameters, **-S** specifies the name of the output SAM file with read mapping, and ***_bowtie2log.txt** is a run-log with alignment statistics and errors.

5. Extract the unmapped reads with samtools fastq and grep programs using the SAM file generated in **step 4** as an input.

   For the right paired reads run:

```
samtools fastq -@ 16 -f 68 reads_to_<yourgenome>.sam > <your-
data>_decontam_1P;
```

For the left paired reads run:

```
samtools fastq -@ 16 -f 132 reads_to_<yourgenome>.sam >
<yourdata>_decontam_2P;
```

where **-@** is the number of CPU cores and **-f** is a SAM flag (full interactive explanation can be found at the following link https://broadinstitute.github.io/picard/explain-flags.html). Getting the unpaired reads is more complicated and requires first outputting all unmapped mates 1 and 2 separately and then eliminating paired mates already present in 1P and 2P files. For this purpose, the following commands can be used:

```
samtools fastq -@ 16 -f 72 reads_to_<yourgenome>.sam >
tmp_reads72;
grep -vf <(grep '@' <yourdata>_decontam_1P) <(grep '@'
tmp_reads72) | grep -A 3 -f - tmp_reads72 | sed 's/^--$//g' |
sed '/^\s*$/d' > <yourdata>_decontam_1U;
```

and for 2 U:

```
samtools fastq -@ 16 -f 136 <yourdata>_decontam_2P >
tmp_reads136;
grep -vf <(grep '@' <yourdata>_decontam_2P) <(grep '@'
tmp_reads136) | grep -A 3 -f - tmp_reads136 | sed 's/^--$//g'
| sed '/^\s*$/d' > <yourdata>_decontam_2U; rm tmp_reads*
```

6. Assemble reads de novo with Trinity software using paired-end protocol (program options are self-explanatory).

```
Trinity --seqType fq --min_contig_length 200 --max_memory 64G
--CPU 16 --left <yourdata>_decontam_1P,<yourdata>_decontam_1U
--right <yourdata>_decontam_2P,<yourdata>_decontam_2U --out-
put trinity_out_<yourdata>_decontam > trinity_<yourdata>_de-
contam.log 2> trinity_<yourdata>_decontam.errlog;
```

If decontamination against reference eukaryotic genome was not performed, substitute "decontam" with "trimmed" in the command above. The resulting output will be the folder with specified name (note that its name must begin with "trinity_out"), which will contain assembled contigs in the "Trinity.fasta" file. Move this file to the working directory and proceed.

7. The following steps will describe the procedure of mapping reads onto assembled contigs. This is done similarly to **step 4** but with Trinity.fasta and original trimmed reads generated in **step 2**. Run.

```
bowtie2-build Trinity.fasta Trinity
```

then

```
bowtie2 -x Trinity -1 <yourdata>_trimmed_1P -2 <yourdata>_-
trimmed_2P -U <yourdata>_trimmed_1U,<yourdata>_trimmed_2U --
end-to-end --very-sensitive -p 16 -S Trinity.sam 2> Trinity_-
bowtie2log.txt
```

Make sure to save the error-log from Bowtie2 (Trinity_-bowtie2log.txt) containing the information on the number of mapped reads, which is important to calculate per-million factor.

8. Use SAMtools to process the output SAM-file (indexing, sorting and converting to binary). This step ensures efficient search in the mapping database. In addition, the resultant *.bam and *.bai files can be loaded together with the contig file Trinity.fasta into a genome browser if visual inspection and/or graphical representation of contig coverage is desired.

Run the series of commands:

```
samtools faidx Trinity.fasta;
samtools view -S -F 0x4 -b -u -t Trinity.fasta Trinity.sam |
samtools sort –o Trinity_sort;
mv Trinity_sort Trinity_sort.bam;
samtools index Trinity_sort.bam Trinity_sort.bam.bai;
```

9. Now, the sorted *.bam file can be used to extract the read count for an individual contig. For this purpose, a loop over Trinity.fasta.fai can be used. The loop reads in contig name and length and passes them to samtools view for read extraction. Then, samtools fastq is used to convert reads back to the text file, from which the read headers are extracted. If both left and right mates are present among contig's reads, only one is kept in accordance with FPKM calculation strategy. The surviving headers are counted. The loop outputs contig name, length, and read count. The output is passed to awk program, which performs per-million and per-kilobase normalization, and the resultant file with contigs' coverage is recorded as Trinity_FPKMs.txt. First, record the per-million variable from the Bowtie2 log file generated in **step 7** with the following command:

```
num_of_reads=`head -n1 Trinity_bowtie2log.txt | awk '{print
$1/1000000}'`
```

then run the loop (note the "pipe" symbol at the end of the loop before awk):

```
while read c d e; do
 samtools view -S -F 4 Trinity_sort.bam $c -o tmp.bam;
 samtools fastq tmp.bam > tmp.fastq;
 echo $c $d `grep '@' tmp.fastq | sed 's/\/.*$//' | sort -u | wc
-l`;
 rm tmp*;
done < Trinity.fasta.fai |
awk -v b="$num_of_reads" '{print$1"\t"$2"\t"$3*1000/$2/b}' >
Trinity_FPKMs.txt
```

(*see* **Note 9**).

10. The identification of virus contigs is best performed in two steps. Firstly, the contigs are searched against nucleotide database with BLASTn. This search is usually fast and very specific. It helps to sift out noncoding sequences, such as ribosomal RNA, which are rather abundant even if the read decontamination (**step 4**) was performed. Furthermore, it reduces the number of potential target contigs that need to be checked via computationally greedy BLASTx-search. The BLASTn is best performed against a custom database built from the host genome or genomes of closely related organisms, although NCBI "nt" database can be used as well. To build a custom database run the following command.

```
makeblastdb -in <yourgenome(s)> -out <yourgenome(s)>_blastndb
-dbtype nucl
```

then, perform the search

```
blastn -query Trinity.fasta -db <yourgenome(s)>_blastndb -out
Trinity_blastn.txt -evalue 1e-5 -outfmt 6 -num_threads 16
```

where **-query** is the file with assembled contigs, **-db** is the search database (custom or "nt"), **-out** is the output file, **-evalue** is the minimal e-value to record a hit (the lower it is—the higher confidence hits are retrieved, default "10" is too high and can result in false-positive hits) and **-outfmt** is type of the output, for further processing the tabulated output (**6**) is the best.

11. Secondly, the contig sequences that were not identified in the previous step (absent from Trinity_blastn.txt file) are extracted from the Trinity.fasta file and BLASTx search against UniProt50 database is performed using DIAMOND software. For sequence extraction fasta file must be linearized, that is, the nucleotide sequence must occupy exactly one line. This can be done with the following Unix shell commands:

```
sed '/^>/ s/.$/&,,,/' Trinity.fasta | tr -d '\n' | sed $'s/>/\\
\n&/g' | sed $'s/,,,/\\\n/g' | tail -n+2 > Trinity_lin.fasta
```

Then, for sequence extraction, run:

```
grep -vf <(awk -F '\t' '{print$1}' Trinity_blastn.txt | sed 's/
.*$//' | sort -u) <(grep '>' Trinity_lin.fasta | sed 's/ .*$//
') | grep -A 1 -f - Trinity_lin.fasta | sed 's/^--$//g' | sed
'/^\s*$/d' > Trinity_for_blastx.fasta
```

Now, prepare a search database for DIAMOND. Download UniRef50 from https://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50/uniref50.fasta.gz and unpack it with gunzip command in the terminal. Fasta-headers of UniRef50 file contain useful information, such as name and accession of the protein, scientific name, and NCBI taxonomy ID of the organism/group. However, both DIAMOND and BLAST+ disregard this information as it is separated with whitespaces in the fasta-header. To avoid the information loss and other bugs, all fasta headers in uniref50.fasta file can be fixed with the following command:

```
sed 's/ /__/;s/ /_/g;s/:/_/;s/\//_/;s/,/_/;s/;/_/;s/\[/_/;s/
\]//' uniref50.fasta | tr -d '(' | tr -d ')' > uniref50_h.fasta
```

Prepare DIAMOND database from a file with proper fasta-headers:

```
diamond makedb –in uniref50_h.fasta -d uniref50DMND
```

Finally, run DIAMOND BLASTx search:

```
diamond blastx -d uniref50DMND -q Trinity_for_blastx.fasta -o
Trinity_blastx.txt --outfmt 6 qseqid stitle pident bitscore
evalue length qstart qend sstart send --ultra-sensitive -t
DMND_TMP -k 1 -p 16
```

where **-d** is DIAMOND database prepared from UniRef50, **-q** is file with contigs selected for BLASTx, **-o** is the output table with hits, **--outfmt 6** is the specification of tabular output format (the same as in BLASTn, **step 11**), **-t** is path to the temporary file, which must be stored locally if the computation is run on a remote server, **-k** is the number of hits per query, and **-p** is the number of CPUs.

12. The information on contigs' coverages and its BLASTn/BLASTx hits can be integrated for better overview using the first column (i.e., contig name) as a key. At this point, more conserved virus proteins, such as RdRp, can be readily identified based on "Tax=" label of the BLASTx hit. A simple text search with "vir" through integrated results file might already return virus hits. Finding more divergent viral proteins, such as, for example, a leishbuviral glycoprotein, is more complicated and requires multiple lines of evidence. These include: (i) correspondence between the contig length and size of the genomic segment observed in the gel (this condition may be violated in case of read assembly errors); (ii) elevated coverage of putative viral contig compared to host mRNA (*see* **Note 10**). Genomic segments originating from the same virus should have comparable coverage. However, the level of expression of genes from these segments may vary, which will be reflected in FPKM values; (iii) since RNA viruses evolve rapidly and encode their genetic information very efficiently, it is expected that the putative viral contig will contain a large ORF (or several overlapping ORFs) that span 70–95% of contig length but will return null result from BLASTx searches (so-called ORFans). If a contig satisfies all three criteria above, it is highly likely to be of viral origin.
13. All *Bunyaviricetes* have complementary sequences at their genomic segments' termini. These can be used as additional markers of viral contigs and facilitate assessing completeness of the latter. For this purpose, numerous RNA secondary structure prediction web-servers (RNAfold, RNAstructure, IPknot, to name just a few) can be employed. The completeness of the genomic segment can be assessed by the position of the ORF in the contig and presence/absence of start and stop-codons. For

short segments up to 1500 nt, the entire contig sequence can be used for prediction although results may be difficult to interpret. Therefore, it is better to extract the first and the last ~200 nt of the contig and connect them with a stretch of 100 uridines. Alternatively, the two fragments can be input directly to RNAcofold server (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAcofold.cgi). If terminal complementary nucleotides are present, they will pair to form a typical "panhandle" structure with 5′- and 3′-ends brought together. The unpaired nucleotides before and after the panhandle can be removed.

4 Notes

1. If any clumps of cells are visible, use an insulin syringe to break them.
2. If the middle (white) phase containing proteins is too thick, it is advisable to shake the sample again, separate it into multiple tubes, top each tube to 1 mL with TRIzol, and go back to the **step 5**. The resultant RNA can be combined at later steps, if necessary. When it is known in advance that the studied organisms contain abundant surface proteins, this problem can be mitigated by incubating the cells for 2–3 min in 1 mL of freshly prepared deproteination solution (Subheading 2.1, **step 7** (Total RNA Isolation, Deproteination solution)) and subsequent washing with PBS.
3. At this step, RNA concentration is usually too high and the solution is viscous. Therefore, it is advisable to dilute a small aliquot ten-fold before measuring the concentration.
4. LiCl precipitation of ssRNA is an alternative to the S1 nuclease removal of ssRNA. For that, after **step 5** of protocol 3.2 (dsRNA Enrichment for Screening and NGS), inactivate DNase I by adding 0.5 μL 200 mM EDTA and incubating for 10 min at 70 °C. Then, add 37.5 μL of ssRNA precipitating solution (Subheading 2.2, **step 4** (dsRNA Enrichment for Screening and NGS)) to 25 μL of the DNase I-treated sample and keep the mix at 4 °C for at least 16 h. Next, centrifuge the sample at 20,000 × g for 30 min at 4 °C in a microcentrifuge, take supernatant, and proceed with standard nucleic acid precipitation protocol: add 1/10 volume (6.3 μL) of 3 M NaOAc, 2.5 volumes (172 μL) of prechilled 96% EtOH, and place the tube at −20 °C for at least 2 h, followed by **steps 10** and **13** of the protocol 3.1 (Total RNA Isolation). In our experience, the LiCl approach results in sharper bands in gel when compared to the nuclease-treated sample. This is likely due to the absence of salts and other chemicals present in the buffers for enzymatic

digestion. As such, this might enhance weaker bands and improve sensitivity. In addition, the smear of undigested low molecular weight RNA, which sometimes appear with S1 nuclease method, is never observed after precipitation with LiCl. On the other hand, the LiCl treatment does not completely eliminate ribosomal RNA, which can mask viral bands. Of note, LiCl precipitation can be performed after the S1 nuclease treatment. In this case the advantages of both methods are combined.

5. We routinely use Zymoclean Gel RNA Recovery Kit for this purpose.
6. A higher voltage can cause overheating and distortion of the gel. It is advisable to use a large electrophoresis chamber and wide loading pockets for best band visualization.
7. Although ethidium bromide is cheap, in many labs it is not used due to safety concerns. In such a case, an alternative fluorescent dye with high sensitivity to RNA (e. g. EliDNA™ PS Green Plus) should be used according to the manufacturer's instructions.
8. Post-staining ensures even distribution of a signal throughout the whole gel area, which is crucial to achieve maximum sensitivity of the method. Destaining procedure removes the unwanted background.
9. It is also possible to get the read count without looping through a contig list using the following command:

```
samtools idxstats *.bam
```

(index file *.bai should be present in the folder). However, it counts multiple alignments of the same read as well as both left and right mates, thus, overestimating the read count.

10. In our experience, if dsRNA-enriched preparations are used for both gel electrophoresis and RNA-seq, contigs with FPKM of at least 30 are visible on the gel as faint bands. Viral dsRNAs usually have FPKM values in the range of 10^2–10^4, similar coverages are observed for trypanosomatid ribosomal RNA. Host protein-coding genes have FPKM values between 1 and 50.

Acknowledgments

The work on *Leishbuviridae* in Yurchenko laboratory is primarily supported by the Grant Agency of the Czech Republic (24-328 10009S) and the European Union's Operational Program "Just Transition" (LERCO CZ.10.03.01/00/22_003/0000003).

References

1. Rampersad S, Tennant P (2018) Replication and expression strategies of viruses. In: Tennant P, Fermin G, Foster JE (eds) Viruses: molecular biology, host interactions, and applications to biotechnology. Academic Press/ Elsevier, London, pp 55–82
2. Baltimore D (1971) Expression of animal virus genomes. Bacteriol Rev 35(3):235–241
3. Field AK, Lampson GP, Tytell AA, Hilleman MR (1972) Demonstration of double-stranded ribonucleic acid in concentrates of RNA viruses. Proc Soc Exp Biol Med 141(2):440–444
4. Zinzula L, Tramontano E (2013) Strategies of highly pathogenic RNA viruses to block dsRNA detection by RIG-I-like receptors: hide, mask, hit. Antivir Res 100(3):615–635
5. Baltimore D (1966) Purification and properties of poliovirus double-stranded ribonucleic acid. J Mol Biol 18(3):421–428
6. Muthukrishnan S, Shatkin AJ (1975) Reovirus genome RNA segments: resistance to S_1 nuclease. Virology 64(1):96–105
7. Diaz-Ruiz JR, Kaper JM (1978) Isolation of viral double-stranded RNAs using a LiCl fractionation procedure. Prep Biochem 8(1):1–17
8. Ghabrial SA, Castón JR, Jiang D, Nibert ML, Suzuki N (2015) 50-plus years of fungal viruses. Virology 479–480:356–368
9. García-Pedrajas MD, Cañizares MC, Sarmiento-Villamil JL, Jacquat AG, Dambolena JS (2019) Mycoviruses in biological control: from basic research to field implementation. Phytopathology 109(11):1828–1839
10. Hong Y, Cole TE, Brasier CM, Buck KW (1998) Evolutionary relationships among putative RNA-dependent RNA polymerases encoded by a mitochondrial virus-like RNA in the Dutch elm disease fungus, *Ophiostoma novo-ulmi,* by other viruses and virus-like RNAs and by the *Arabidopsis* mitochondrial genome. Virology 246(1):158–169
11. Crawford LJ, Osman TA, Booy FP, Coutts RH, Brasier CM, Buck KW (2006) Molecular characterization of a partitivirus from *Ophiostoma himal-ulmi*. Virus Genes 33(1):33–39
12. Osaki H, Nakamura H, Sasaki A, Matsumoto N, Yoshida K (2006) An endornavirus from a hypovirulent strain of the violet root rot fungus, *Helicobasidium mompa*. Virus Res 118(1–2):143–149
13. Kozlakidis Z, Hacker CV, Bradley D, Jamal A, Phoon X, Webber J, Brasier CM, Buck KW, Coutts RH (2009) Molecular characterisation of two novel double-stranded RNA elements from *Phlebiopsis gigantea*. Virus Genes 39(1):132–136
14. Magae Y (2012) Molecular characterization of a novel mycovirus in the cultivated mushroom, *Lentinula edodes*. Virol J 9:60
15. Kanhayuwa L, Kotta-Loizou I, Ozkan S, Gunning AP, Coutts RH (2015) A novel mycovirus from *Aspergillus fumigatus* contains four unique dsRNAs as its genome and is infectious as dsRNA. Proc Natl Acad Sci USA 112(29):9100–9105
16. Niu Y, Yuan Y, Mao J, Yang Z, Cao Q, Zhang T, Wang S, Liu D (2018) Characterization of two novel mycoviruses from *Penicillium digitatum* and the related fungicide resistance analysis. Sci Rep 8(1):5513
17. Weber F, Wagner V, Rasmussen SB, Hartmann R, Paludan SR (2006) Double-stranded RNA is produced by positive-strand RNA viruses and DNA viruses but not in detectable amounts by negative-strand RNA viruses. J Virol 80(10):5059–5064
18. Pearson MN, Beever RE, Boine B, Arthur K (2009) Mycoviruses of filamentous fungi and their relevance to plant pathology. Mol Plant Pathol 10(1):115–128
19. Khan HA, Nerva L, Bhatti MF (2023) The good, the bad and the cryptic: the multifaceted roles of mycoviruses and their potential applications for a sustainable agriculture. Virology 585:259–269
20. Robinson JI, Beverley SM (2018) Concentration of 2'C-methyladenosine triphosphate by *Leishmania guyanensis* enables specific inhibition of *Leishmania RNA virus 1 via* its RNA polymerase. J Biol Chem 293(17):6460–6469
21. Kuhlmann FM, Robinson JI, Bluemling GR, Ronet C, Fasel N, Beverley SM (2017) Antiviral screening identifies adenosine analogs targeting the endogenous dsRNA *Leishmania RNA virus 1* (LRV1) pathogenicity factor. Proc Natl Acad Sci USA 114(5):E811–E819
22. Cashdollar LW, Esparza J, Hudson GR, Chmelo R, Lee PW, Joklik WK (1982) Cloning the double-stranded RNA genes of reovirus: sequence of the cloned *S2* gene. Proc Natl Acad Sci USA 79(24):7644–7648
23. Imai M, Richardson MA, Ikegami N, Shatkin AJ, Furuichi Y (1983) Molecular cloning of double-stranded RNA virus genomes. Proc Natl Acad Sci USA 80(2):373–377
24. Niu Y, Zhang T, Zhu Y, Yuan Y, Wang S, Liu J, Liu D (2016) Isolation and characterization of

a novel mycovirus from *Penicillium digitatum*. Virology 494:15–22

25. Sterky F, Lundeberg J (2000) Sequence analysis of genes and genomes. J Biotechnol 76(1):1–31
26. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A (2013) *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat Protoc 8(8):1494–1512
27. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19(5):455–477
28. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu J, Holmes EC, Zhang YZ (2015) Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. elife 4
29. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang W, Xu J, Holmes EC, Zhang YZ (2016) Redefining the invertebrate RNA virosphere. Nature 540(7634):539–543
30. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, Holmes EC, Zhang YZ (2018) The evolutionary history of vertebrate RNA viruses. Nature 556(7700):197–202
31. Zayed AA, Wainaina JM, Dominguez-Huerta G, Pelletier E, Guo J, Mohssen M, Tian F, Pratama AA, Bolduc B, Zablocki O, Cronin D, Solden L, Delage E, Alberti A, Aury JM, Carradec Q, da Silva C, Labadie K, Poulain J, Ruscheweyh HJ, Salazar G, Shatoff E, Tara Oceans Coordinatorsdouble, d, Bundschuh R, Fredrick K, Kubatko LS, Chaffron S, Culley AI, Sunagawa S, Kuhn JH, Wincker P, Sullivan MB, Acinas SG, Babin M, Bork P, Boss E, Bowler C, Cochrane G, de Vargas C, Gorsky G, Guidi L, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Kandels S, Karp-Boss L, Karsenti E, Not F, Ogata H, Poulton N, Pesant S, Sardet C, Speich S, Stemmann L, Sullivan MB, Sungawa S, Wincker P (2022) Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. Science 376(6589):156–162
32. Kukurba KR, Montgomery SB (2015) RNA sequencing and analysis. Cold Spring Harb Protoc 2015(11):951–969
33. Yang IS, Kim S (2015) Analysis of whole transcriptome sequencing data: workflow and software. Genomics Inform 13(4):119–125
34. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120
35. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359
36. Ramirez-Gonzalez RH, Bonnal R, Caccamo M, Maclean D (2012) Bio-SAMtools: ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. Source Code Biol Med 7(1):6
37. Bushnell B, Rood J, Singer E (2017) BBMerge – accurate paired shotgun read merging *via* overlap. PLoS One 12(10):e0185056
38. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc 7(3):562–578
39. Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res 38(14):4570–4578
40. Cross ST, Michalski D, Miller MR, Wilusz J (2019) RNA regulatory processes in RNA virus biology. Wiley Interdiscip Rev RNA 10(5):e1536
41. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421
42. Buchfink B, Reuter K, Drost HG (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18(4):366–368
43. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6):926–932
44. Prakash A, Jeffryes M, Bateman A, Finn RD (2017) The HMMER web server for protein sequence similarity search. Curr Protoc Bioinformatics 60: 3.15.11–13.15.23
45. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J (2019) HH-suite3 for fast remote homology

detection and deep protein annotation. BMC Bioinformatics 20(1):473

46. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, Krupovic M, Fire A, Dolja VV, Koonin EV (2020) Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat Microbiol 5(10):1262–1270
47. Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, Chen IM, Ivanova N, Zeigler Allen L, Paez-Espino D, Bryant DA, Bhaya D, Consortium RNAVD, Krupovic M, Dolja VV, Kyrpides NC, Koonin EV, Gophna U (2022) Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell 185(21):4023–4037
48. Grybchuk D, Akopyants NS, Kostygov AY, Konovalovas A, Lye LF, Dobson DE, Zangger H, Fasel N, Butenko A, Frolov AO, Votýpka J, d'Avila-Levy CM, Kulich P, Moravcová J, Plevka P, Rogozin IB, Serva S, Lukeš J, Beverley SM, Yurchenko V (2018) Viral discovery and diversity in trypanosomatid protozoa with a focus on relatives of the human parasite *Leishmania*. Proc Natl Acad Sci USA 115(3):E506–E515
49. Grybchuk D, Kostygov AY, Macedo DH, Votýpka J, Lukeš J, Yurchenko V (2018) RNA viruses in *Blechomonas* (Trypanosomatidae) and evolution of *Leishmaniavirus*. MBio 9(5): e01932–e01918
50. Grybchuk D, Macedo DH, Kleschenko Y, Kraeva N, Lukashev AN, Bates PA, Kulich P, Leštinová T, Volf P, Kostygov AY, Yurchenko V (2020) The first non-LRV RNA virus in *Leishmania*. Viruses 12(2):168
51. Macedo DH, Grybchuk D, Režnarová J, Votýpka J, Klocek D, Yurchenko T, Ševčík J, Magri A, Urda Dolinská M, Záhonová K, Lukeš J, Servienė E, Jászayová A, Serva S, Malysheva MN, Frolov AO, Yurchenko V, Kostygov AY (2023) Diversity of RNA viruses in the cosmopolitan monoxenous trypanosomatid *Leptomonas pyrrhocoris*. BMC Biol 21(1):191
52. Klocek D, Grybchuk D, Macedo DH, Galan A, Votýpka J, Schmid-Hempel R, Schmid-Hempel P, Yurchenko V, Kostygov AY (2023) RNA viruses of *Crithidia bombi*, a parasite of bumblebees. J Invertebr Pathol 201:107991