Detection, quantification, and typing of enteroviruses in wastewater

Albert Carcereny

Enteric virus laboratory. University of Barcelona   https://orcid.org/0000-0002-9615-7576

David Garcia-Pedemonte

Enteric virus laboratory. University of Barcelona   https://orcid.org/0000-0001-8375-8797

Albert Bosch

Enteric virus laboratory. University of Barcelona   https://orcid.org/0000-0002-8111-9059

Rosa M Pintó

rpinto@ub.edu

Enteric virus laboratory. University of Barcelona   https://orcid.org/0000-0003-1382-6648

Susana Guix

susanaguix@ub.edu

Enteric virus laboratory. University of Barcelona   https://orcid.org/0000-0002-1588-3198

Method Article

Abstract

Wastewater analysis serves as a complementary tool to clinical data in monitoring both the quantity and diversity of circulating enteroviruses among the population. While enterovirus infections typically result in asymptomatic or mild disease, certain types can lead to severe symptoms, highlighting the importance of tracking their prevalence and distribution.

Here we present both a RTqPCR assay targeting a conserved region in the 5' end of the enterovirus genome to detect and quantify its presence in sewage along with a sequencing pipeline based in MinION platform (Oxford Nanopore Technologies) to characterize circulating enterovirus types in samples with sufficient virus load (Cq≤33). Two sequencing approaches are possible: one targeting all enterovirus types (1), and the other specifically focusing on Cluster C enterovirus (2), including poliovirus. The obtained data are processed using a designed pipeline to quantify the proportions of different enterovirus types, providing valuable insights into the spread of potentially severe types.

Introduction

Reagents

RT-qPCR

· One Step PrimeScript™ RT-PCR Kit (Perfect Real Time) (RR064A, Takara).

· Twist Synthetic human Enterovirus 68 RNA control (NC_038308.1) (103005, Twist Bioscience).

· Primers and probe (Table 1).

· Nuclease free water.

NGS

· Enzymes and primers proposed by Nix et.al., 2006 (1) with the aim to amplify and sequence all enterovirus
 types.

· Enzymes and primers proposed by Shaw et.al., 2020 (2) with the aim to amplify and sequence cluster C enterovirus.

· Reagents listed in the ONT protocol *Ligation sequencing amplicons - Native Barcoding Kit 24 V14 (SQK-NBD114.24)* (3).

Equipment

· Bio-Rad CFX96 touch real-time PCR detection system.

· Bio-Rad T100 Thermal Cycler.

· MinION Mk1C or Mk1B.

· Mid-range computer.

Procedure

RT-qPCR

Composition of RTqPCR mastermix (final volume of RTqPCR reaction of 20 µl)

2x One Step: 10 µl

Forward primer 10µM: 0.8 µl (400 nm)

Reverse primer 10µM: 0.8 µl (400 nm)

Probe 10µM: 0.4 µl (200 nm)

Takara Ex Taq: 0.4 µl

Prime Script Enzyme: 0.4 µl

H_2O PCR: 2.2 µl

Extracted RNA: 5 µl

<u>Controls to be included in each RTqPCR assay</u>:

· 2 wells with 5 µl of nuclease-free water (RTqPCR negative controls)

· 2 wells with 5 µl of RNA extraction negative control (RNA extraction negative controls)

· 1 well with 5 µl of the synthetic human Enterovirus 68 RNA control at 1000 copies/µL (RTqPCR positive control).

· For each sample, 2 wells of undiluted RNA and 2 wells of 1/10 dilution are analysed to assess inhibition.

For quantification, a standard curve is constructed with a minimum of 5 10-fold dilutions and 3 wells for each dilution, using the synthetic human Enterovirus 68 RNA control as reference material.

<u>Thermocycler conditions</u>

- 10 min at 50ºC (x1)

- 3 min at 95ºC (x1)

- 15 sec at 95ºC and 30 sec at 57ºC (x45)

<u>Interpretation of results</u>

The calculation of enterovirus concentration in genome copies per reaction (gc/rxn) in each well is performed using the standard curve.

Occurrence of inhibition and calculation of mean viral titters are estimated by comparing concentrations obtained from duplicate wells tested for the two RNA dilutions (undiluted RNA and 1/10 dilution), as described in Carcereny et al., 2021 (4). Mean concentration of samples and standard error are calculated using as many data as possible, taking into consideration the following steps:

a) Calculate mean concentration as gc/rxn for each RNA dilution:

    a. Data from RNA dilutions containing "No Cq" or Cq≥40 in both wells are not used for calculation.

        i. When in the analysis of the 2 wells of any RNA dilution, one of the wells has a Cq value <40 and the
        other has "No Cq" or Cq≥40, this last well is assigned a concentration equal to the theoretical limit of
        detection (LoD) of 1 gc/rxn.

b) Calculate mean concentration as gc/rxn for each sample:

    b. When the difference between the concentration estimated from undiluted RNA and 1/10 dilution is < 0.5
        $\log_{10}$, mean concentration of sample is calculated using data from the 4 wells.

        i. When the difference between the concentration estimated from undiluted RNA and 1/10 dilution is ≥
        0.5 $\log_{10}$, inhibition is considered, and mean concentration of sample is calculated using data from
        the 1/10 dilution.

The final concentration can be expressed in gc/L.

The calculated LOD was of 4gc/rxn and the limit of quantification (LOQ) was of 8.5gc/rxn.

NGS

<u>Sample amplification</u>

Samples with a Cq value ≤33 in the undiluted RNA wells are selected for sequencing. While sequencing samples with a Cq value >33 is possible, it may result in a lower success ratio due to potential degradation or lower viral load.

Two amplification approaches are available depending on the aim of the study. To sequence all enterovirus types, present in a sample, the protocol published by Nix et.al., 2006 (1) targeting a region within the VP1 gene (~348-393bp) is followed. Alternatively, if the focus is only Cluster C enterovirus, the protocol published by Shaw et.al., 2020 (2) targeting Cluster C VP1 (~1089bp) is followed.

For both amplification approaches, the *Ligation sequencing amplicons - Native Barcoding Kit 24 V14 (SQK-NBD114.24)* (3) sequencing protocol developed by Oxford Nanopore is followed.

Sequencing parameters

Mk1C and Mk1B are used with R10.4.1 flow cells with the following parameters:

- Run length: 72h

- Active channel selection: On

- Pore scan frequency: 1.5 hours (default)

- Reserved pores: On

- Minimum read length: 20bp

- Read splitting: On

- Override read splitting min score: On and set to 48

- Basecalling: Super-accurate basecalling (SUP) 400 bps 5kHz. In Mk1C live basecalling is deactivated and raw
  data is post-run GPU basecalled in a Workstation with Dorado basecaller.

- Modified basecalling: Off

- Trim barcodes: On

- Barcode both ends: On

- Mid-read barcode filtering: On

- Override minimum barcoding score: On and set to 60 (default)

- Override minimum mid-read barcoding score: On and set to 50 (default)

- Minimum Q score: 8

Raw data is obtained in POD5 format and FASTQ compression is enabled to save disk space.

Sequencing analysis

How to run

A reference database containing 6303 Enterovirus sequences was generated and is provided along with the necessary scripts to run the pipeline https://github.com/susanaguix/WEVTYTO

The pipeline is based on VSEARCH tool (5) allowing high analysis flexibility and adaptability.

The following steps are to be followed to perform the analysis:

1. Raw FASTQ.gz files generated for each barcode need to be merged and renamed with its corresponding
    sample name. Please avoid the use of underscores in your sample names.

2. All renamed FASTQ.gz files should be copied into a folder along with the reference FASTA file
    "ev_reference_sequences.fasta.gz" and the Python script "ev_typing_nix.py", if Nix's protocol was followed,
    or the "ev_typing_shaw.py" if Shaw's protocol was followed.

3. Run the corresponding Python script. The use of a Conda environment with all the required dependencies is
    recommended.

4. A new folder named "results" will be created containing all the generated files and 2 folders, one named
    "Excel_results" and another named "renamed_fasta". More details can be found in the next subsection.

<u>Detailed information</u>

The first step uses VSEARCH filtering options to trim 20 nucleotides from both ends of each read to discard low-quality bases and remove primer sequences. Then, the filtered FASTQ files are converted to FASTA and clustered with VSEARCH "--cluster_fast" option. Clustering step can be customized by changing, adding, or removing options. By default, the following parameters are defined:

- Cluster identity (--id): 95%.

- Cluster consensus sequence output (--consout).

- Minimum and maximum sequence length filter (--minseqlength and --maxseqlength): 250 bp and 600 bp
  respectively for Nix protocol and 800 bp and 1000 bp respectively for Shaw protocol.

- Sequence count added into the cluster consensus sequence header (--sizeout).

- Consider sequence abundance (--sizein).

- Sort clusters consensus sequences by decreasing order (--clusterout_sort).

- Check both strands when clustering (--strand both).

- Number of CPU threads dedicated to the clustering process (--threads): 4. This value can be adjusted to increase the analysis speed.

A FASTA file containing the clusters consensus sequences for each sample is generated in this step and taken into the BLAST search step where the consensus sequences will be aligned against the reference sequence database.

The BLAST search step is performed with VSEARCH ("--usearch_global") option along with the following parameters:

- Minimum BLAST id to consider a match valid (--id): 80%.

- Check both strands when aligning (--strand both).

- Output the obtained alignments (--alnout).

- Generate a BLAST results file (--blast6out).

- Output sample sequences matching and not matching database sequences (--matched and --notmatched).

- Output only the best alignment result between sample sequences and reference sequences (--top_hits_only).

- Discard those alignments in which the number of aligned nucleotides is less than 200 (--mincols 200).

The generated results file is then used to calculate the abundance of the different enterovirus present in the sample. Two (or more if the number of obtained results is very large) Excel files are obtained per each sample and saved into a folder named "Excel_results". The Excel file (or files) named "original" contains all the unaggregated results of a given sample distributed in the following 7 columns:

- Reference: collects all the hits between the cluster consensus sequences and the reference sequences. The
  accession number and the corresponding Enterovirus type are indicated.

- Sequence Count: the number of original sequences of each hit.

- BLAST ID: the identity percentage obtained for each hit.

- Seq BLAST id[SGA1] (80-94.9%): number of original sequences having a BLAST id between 80% and 94.9%
  for each hit.

- Seq BLAST id (95-100%): number of original sequences having a BLAST id between 95% and 100% for each
  hit.

- EV Type: the corresponding enterovirus type to which each cluster consensus sequence corresponds.

- BLAST id * Seq Count: multiplication of the BLAST id by the Sequence Count for each entry. This column is
  used for the calculation of the weighted BLAST id in the aggregated file.

The Excel file named "aggregated" contains a summary of the results distributed in the following 9 columns:

- EV Type: a list of all the enterovirus types found in the sample.

- Total Sequence Count: the sum of all the original sequences for each enterovirus type.

- Percentage: the calculated percentage value for each enterovirus type.

- Min BLAST id: the minimum BLAST id obtained for each enterovirus type.

- Max BLAST id: the maximum BLAST id obtained for each enterovirus type.

- Seq BLAST id (80-94.9%): total number of sequences having a BLAST id between 80% and 94.9% for each
  enterovirus type.

- Seq BLAST id (95-100%): total number of sequences having a BLAST id between 95% and 100% for each
  enterovirus type.

- BLAST id * Seq Count_sum: the sum of the multiplication of the BLAST id by the Sequence Count for each
  enterovirus type.

- Mean Weighted BLAST id: the mean BLAST id for each enterovirus type considering the number of sequences.
  The "BLAST ID * Seq Count_sum" value is divided by the "Total Sequence Count" value for each enterovirus
  type.

Finally, a folder named "renamed_fasta" is created containing the matched sample sequences with its headers renamed including both the corresponding accession number and the enterovirus type. This file can then be used for subsequent analysis allowing sequence traceability.

Troubleshooting

Time Taken

Anticipated Results

References

1. Allan Nix W, Oberste MS, Pallansch MA. Sensitive, seminested PCR amplification of VP1 sequences for direct identification of all enterovirus serotypes from original clinical specimens. J Clin Microbiol. 2006 Aug;44(8):2698–704.

2. Shaw AG, Majumdar M, Troman C, O'Toole Á, Benny B, Abraham D, et al. Rapid and sensitive direct detection and identification of poliovirus from stool and environmental surveillance samples by use of nanopore sequencing. J Clin Microbiol. 2020 Sep 1;58(9).

3. https://community.nanoporetech.com/docs/prepare/library_prep_protocols/ligation-sequencing-amplicons-native-barcoding-v14-sqk-nbd114-24/v/nba_9168_v114_revm_15sep2022

4. Carcereny A, Martínez-Velázquez A, Bosch A, Allende A, Truchado P, Cascales J, et al. Monitoring Emergence of the SARS-CoV-2 B.1.1.7 Variant through the Spanish National SARS-CoV-2 Wastewater Surveillance System (VATar COVID-19). Environ Sci Technol. 2021 Sep 7;55(17):11756–66.

5. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: A versatile open source tool for metagenomics. PeerJ. 2016;2016(10).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.pdf