Long Amplicon Nanopore Sequencing for Dual-Typing RdRp and VP1 Genes of Norovirus Genogroups I and II in Wastewater

George Scott[1], David Ryder[1], Mary Buckley[1], Richard Hill[1], Samantha Treagus[2], Tina Stapleton[1], David Walker[1], James Lowther[1], Frederico Batista[1]

[1]Centre for Environment, Fisheries and Aquaculture Science;
[2]Faculty of Environment, Science and Economy, University of Exeter

George Scott
Centre for Environment, Fisheries and Aquaculture Science

MAR 26, 2024

DISCLAIMER

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

DOI:
dx.doi.org/10.17504/protocols.io.8epv5xpmjg1b/v1

Protocol Citation: George Scott, David Ryder, Mary Buckley, Richard Hill, Samantha Treagus, Tina Stapleton, David Walker, James Lowther, Frederico Batista 2024. Long Amplicon Nanopore Sequencing for Dual-Typing RdRp and VP1 Genes of Norovirus Genogroups I and II in Wastewater. **protocols.io** https://dx.doi.org/10.17504/protocols.io.8epv5xpmjg1b/v1

Protocol status: Working
We use this protocol and it's working

Created: Jan 17, 2024

Last Modified: Mar 26, 2024

PROTOCOL integer ID: 93653

Keywords: Norovirus, Oxford Nanopore Technologies, Long Amplicon, Wastewater, Sequencing

Funders Acknowledgement:
Funded by His Majesty's Treasury Shared Outcome Fund PATH-SAFE programme

ABSTRACT

This protocol outlines the procedures for dual-typing (genotyping and polymerase typing) of norovirus genogroups I and II (GI and GII) from wastewater. It assumes that wastewater sample collection, viral concentration and nucleic acid extraction has already been performed. The wastewater samples used during method development were processed using an ammonium sulphate precipitation (150 mL sample) followed by nucleic acid extraction using Kingfisher Flex™ (Thermo Scientific™, UK) and NucliSENS® reagents (BioMérieux, France).
Adapting this technique for use with differing viral concentration methods and nucleic acid extraction techniques should still be suitable. The analysis of different sample matrices will also likely be possible following sample processing (up to and including nucleic acid extraction) with methods appropriate for that matrix.
This protocol starts with reverse transcription followed by semi-nested PCR amplifying the *RdRp+VP1* region
of norovirus GI and GII independently. GI and GII amplicons are combined into a single library for simultaneous sequencing of the ≈1000 bp PCR products using either the Oxford Nanopore Technologies MinION or GridION.
Consensus sequences are generated using NGSpeciedID, grouped at 95% then dual-typed giving both genotype and polymerase type (e.g.GI.2[P2]). Reads are aligned to consensus sequences and PCR chimeras are filtered using reference, de novo and manual approaches.

PROTOCOL REFERENCES

Primers:
Ollivier, J. *et al.* (2022) *Application of Next Generation Sequencing on Norovirus-contaminated oyster samples*, *EFSA Supporting Publications*. doi: 10.2903/sp.efsa.2022.en-7348.

Yuen, Catton et al, 2001

Kojima et al, 2002

CITATIONS

STEP 1

Harry T Child, Paul A O'Neill, Karen Moore, Hubert Denise, Matthew Loose, Steve Paterson, Ronny van Aerle, Aaron Jeffries. Wastewater Sequencing using the EasySeq™ RC-PCR SARS CoV-2 (Nimagen) V3.0
https://protocols.io/view/wastewater-sequencing-using-the-easyseq-rc-pcr-sar-cih2ub8e

MATERIALS

A	B	C
Material	Manufacturer	Product code
Mag-Bind TotalPure NGS	OmegaBIO-TEK	M1378-00
Ethanol, Molecular Biology Grade	Sigma-Aldrich	1085430250
Water, Molecuar Biology Grade	Sigma-Aldrich	W4502-1L
LunaScript RT SuperMix Kit	New England Biolabs	E3010
TE buffer, Molecular Biology Grade	Sigma-Aldrich	574793
Platinum Taq Polymerase	Invitrogen	10966018
D5000 Reagents, TapeStation	Agilent	5067-5589
D5000 ScreenTapes, Tapestation	Agilent	5067-5588
ExoSAP-IT	Applied Biosystems	78200
1X dsDNA High Sensitivity Kit, Qubit	Invitrogen	Q32851
Native Barcoding Kit 96 V14	Oxford Nanopore	NBD114.96
R10.4.1 Flow Cell	Oxford Nanopore	FLO-MIN114
Blunt/TA Ligase Master Mix	New England Biolabs	M0367
NEBNext Ultra II End repair/dA-tailing Module	New England Biolabs	E7546
NEBNext Quick Ligation Module	New England Biolabs	E6056
Bovine Serum Albumin, UltraPure	Invitrogen	AM2616
dNTP Deoxynucleotide (dNTP) Solution Mix	New England Biolabs	N0447S

BEFORE START INSTRUCTIONS

This protocol assumes that wastewater sample collection, viral concentration and nucleic acid extraction has already been performed.

Inhibitor Removal

1

Note

For best practice, a process control negative and positive sample should be implemented from this point onwards. Norovirus GI and GII positive reference materials can be obtained from the UK Health Security Agency in the form of Virus LENTICULE® Discs. They will need dissolving in 975 µL PBS and then nucleic acid extraction prior to inputting into this process.

Note

The inhibitor removal and reverse transcription procedures in this protocol were adapted from the citation below:

CITATION

Harry T Child, Paul A O'Neill, Karen Moore, Hubert Denise, Matthew Loose, Steve Paterson, Ronny van Aerle, Aaron Jeffries. Wastewater Sequencing using the EasySeq™ RC-PCR SARS CoV-2 (Nimagen) V3.0. protocols.io.

LINK

https://protocols.io/view/wastewater-sequencing-using-the-easyseq-rc-pcr-sar-cih2ub8e

2 Pipette 25 µL of RNA extract into a 96 well PCR plate

3 Add 45 µL (1.8X) of Mag-Bind® Total Pure NGS beads

4 Pipette carefully 10x to mix

5 Leave to stand at room temperature for 00:05:00 5m

6 If there is liquid on the side of the wells, cover the plate with a PCR adhesive seal and briefly spin down plate

7 Place on magnetic rack for 00:03:00 3m

8 Remove 68 µL of supernatant

9 Add 80 µL of [M] 80 % (v/v) ethanol for 00:00:30 (make EtOH fresh every time with nuclease free water) 30s

10 Remove 82 µL of supernatant being careful to avoid the pellet

11 Repeat clean once more go to step #9

12 Set pipette to 20 µL and remove any residual ethanol

13 Leave to dry for 00:02:00 2m

14 Remove from magnetic rack

15 Add 27 µL of molecular grade water and mix by pipetting 10x

16 Allow to stand for 00:05:00 5m

17 Place plate back on magnetic rack for 00:03:00 3m

18 Recover 25 µL of the supernatant

Reverse Transcription

19 Add 2.5 µL of LunaScript® RT SuperMix to a 0.2 mL PCR reaction tube placed on ice or on a cool block

20 Add 10 µL of cleaned nucleic acid extract and mix gently by pipetting 5x

21 Cover or cap the reaction tubes and briefly spin down

22 Incubate the combined reaction mixture in a thermocycler with a heated lid (105°C) for: 48m

25 °C for 00:02:00

55 °C for 00:45:00

95 °C for 00:01:00

Hold at 4 °C

23 Briefly spin down the reaction tubes

24 Add 7.5 µL of molecular biology grade water to each sample

25

Note

The first-strand cDNA can be stored in the fridge at 4 °C overnight or in the freezer at -20 °C if longer term storage is required.

Primer Preparation

26

Note

Cartridge purified primers should be purchased to prevent loss of the degenerate oligonucleotides

27 Reconstitute the lyophilised primers in Table 1 to [M] 100 micromolar (µM) using TE buffer pH 8

Table 1. Norovirus GI and GII Primers

A	B	C	D
Name	Genogroup	F/R	Sequence (5'-3')
NV4478m	GI	F1	AARYTVCCHATHAARGTTGGNATG
NV4562m	GI	F2	GATGCDGAYTAYACRGCHTGGG
GISKRm	GI	R1,2	CCIACCCAICCATTRTACA
NV4611	GII	F1	CWGCAGCMCTDGAAATCATGG
NV4692	GII	F2	GTGTGRTKGATGTGGGTGACTT
GIISKR	GII	R1,2	CCRCCNGCATRHCCRTTRTACAT

1) first round primers and **2**) semi-nested primers

28 Dilute primers to a working concentration of [M] 10 micromolar (µM) in molecular biology grade water and aliquot into suitable volumes to avoid repeated freeze-thaw cycles

First-Round PCR

29

Note

All PCR cycling conditions used a heated lid at 105 °C and a ramping speed of 3°C/s.

Note

Prepare enough mastermix for all of the samples plus the process and PCR positive and negative controls.

Note

GI and GII PCRs are performed independently for each wastewater sample under analysis.

30 Prepare the mastermixes as indicated in **Table 2** and **Table 3** for the first round PCRs

Table 2. Forward and reverse primers for the GI and GII first-round PCRs

A	B	C
Genogroup	Forward Primer	Reverse Primer
GI	NV4478m	GISKRm
GII	NV4611	GIISKR

Table 3. Mastermix Recipe for all PCRs

A	B	C
Reagents	Final concentration	Volume (µL)
H20 (molecular biology grade)	-	15.15

A	B	C
PCR Buffer, without Mg (10X)	1	2.5
dNTP (10 mM)	0.2 mM	0.5
MgCl2 (50 mM)	1.5 mM	0.75
Forward primer (10 µM)	0.2 µM	0.5
Reverse primer (10 µM)	0.2 µM	0.5
Platinum Taq polymerase (10 U/µL)	1 unit/tube	0.1

30.1 Remove all mastermix components from the freezer and place the polymerase on 🌡 On ice

30.2 Defrost all other components at 🌡 Room temperature

30.3 Briefly vortex and spin down all components

30.4 Add the components as outlined in Table 2 and 3 reserving the polymerase until the end

30.5 When adding the polymerase, slowly aspirate and pre-wet pipette tip x3

30.6 After dispensing the polymerase rinse the pipette tip 10x

30.7 Vortex then briefly spin down the prepared mastermix

31 Distribute 20 µL of PCR mastermix 0.2 mL reaction tubes

32 Add 5 µL of cDNA for each sample

33 Add 5 µL of molecular biology grade water or positive control cDNA for your PCR positive and negative controls

34 Seal or cap the reaction tubes then spin down ensuring no bubbles are present

35 Run the reactions in a thermocycler using the conditions in **Table 4** and **Table 5** for GI and GII

Table 4. GI first-round PCR cycling conditions

A	B	C	D
Stage	Temperature (°C)	Time	Cycles
Initial denature	95.0	60 s	1

A	B	C	D
Denature	95.0	30 s	40
Anneal	47.4	30 s	
Elongation	72.0	30 s	
Final elongation	72.0	7 min	1
Hold	4.0		

Table 5. GII first-round PCR cycling conditions

A	B	C	D
Stage	Temperature (°C)	Time	Cycles
Initial denature	95.0	60 s	1
Denature	95.0	30 s	40
Anneal	55.7	30 s	
Elongation	72.0	30 s	
Final elongation	72.0	7 min	1
Hold	4.0		

Semi-Nested PCR

36

Note

Prepare enough mastermix for all of the samples and controls from the first-round PCR plus an additional PCR negative control for the semi-nested PCR.

37 Prepare the mastermixes as indicated in **Table 3** and **Table 6** following

Table 6. GI and GII primer pairs for the semi-nested PCR

A	B	C
Genogroup	Forward Primer	Reverse Primer
GI	NV4562m	GISKRm
GII	NV4692	GIISKR

37.1 Remove all mastermix components from the freezer and place the polymerase on 🌡 On ice

37.2 Defrost all other components at 🌡 Room temperature

37.3 Briefly vortex and spin down all components

37.4 Add the components as outlined in Table 2 and 3 reserving the polymerase until the end

37.5 When adding the polymerase, slowly aspirate and pre-wet pipette tip x3

37.6 After dispensing the polymerase rinse the pipette tip 10x

37.7 Vortex then briefly spin down the prepared mastermix

38 Distribute 20 µL of PCR mastermix into a 96-well plate

39 Vortex and spin down the plate from the first round PCR

40 Add 5 µL of first-round PCR product to each of the sample wells

41 Add 5 µL of molecular biology grade water for your negative control

42 Seal or cap the plate then spin down ensuring no bubbles are present

43 Run using the conditions in **Table 7** and **Table 8** for GI and GII

Table 7. GI semi-nested PCR cycling conditions

A	B	C	D
Stage	Temperature (°C)	Time	Cycles
Initial denature	95.0	60 s	1
Denature	95.0	30 s	40

A	B	C	D
Anneal	57.2	30 s	
Elongation	72.0	30 s	
Final elongation	72.0	7 min	1
Hold	4.0		

Table 8. GII semi-nested PCR cycling conditions

A	B	C	D
Stage	Temperature (°C)	Time	Cycles
Initial denature	95.0	60 s	1
Denature	95.0	30 s	40
Anneal	55.7	30 s	
Elongation	72.0	30 s	
Final elongation	72.0	7 min	1
Hold	4.0		

Analysis of PCR Controls

44

Note

This step uses a TapeStation for PCR product analysis, but gel electrophoresis with 2% agarose tris-borate EDTA gel run at 125 V for 35 min with a 100 bp ladder (Promega, USA), can be used.

45 Remove ScreenTape from packaging, ensure there are no bubbles in any of the electrophoresis lanes, flicking ScreenTape to remove bubbles if neccesary

46 Place ScreenTape into the TapeStation and allow to come to room temperature 00:30:00 30m

47 Load 10 µL of D5000 reagents into the 0.2 mL TapeStation reaction tubes

48 Add 1 µL of ladder into the tube representing position TapeStation position A1

49 Load 1 µL of your positive and negative PCR and process controls into the 0.2 mL TapeStation reaction tubes

50 Cap tubes, vortex and then spin down ensuring no bubbles are present

51 Place tubes into the machine and analyse the samples

52

Note

All controls should give their expected results and appropriate actions should be taken if any of the negative controls show contamination.

PCR Product Clean-Up

53

Note

This is performed for both the GI and GII semi-nested PCR products.

54 Remove ExoSAP-IT from the -20 °C freezer and allow to defrost On ice

55 Briefly vortex and spin down reaction tubes or plates from the semi-nested PCR

56 Briefly vortex and spin down ExoSap-IT

57 Aliquot 10 µL of semi-nested PCR product into a new reaction tube or plate

58 Add 4 µL of ExoSap-IT to each PCR product

59 Seal or cap reaction vessel

60 Incubate with a heated lid (105 °C): 30m

37 °C for 00:15:00

80 °C for 00:15:00

Hold at 4 °C

PCR Product Quantification

61

Note

This is performed for both the GI and GII cleaned semi-nested PCR products.

62 Add 198 µL High-Sensitivity dsDNA Qubit™ reagents to Qubit tubes for each sample being quantified

63 Load 2 µL of cleaned, semi-nested PCR product onto the side of the tube ensuring residual nucleic acids aren't present on the exterior of the pipette tip

64 Add 190 µL High-Sensitivity dsDNA Qubit™ reagents to Qubit tubes for your standards

65 Load 10 µL of each standard

66 Cap tubes, vortex and briefly spin down

67 Check for presence of bubbles, flick tubes and spin down again if necessary

68 Incubate for 00:02:00 and then measure with the Qubit™ 2m

GI and GII PCR Product Pooling

69

Note

At the end of this process, for every sample being analysed, you will have a combined 200 fmol of GI and GII PCR products comprised of 85.3 and 114.7 fmol of GI and GII amplicons, respectively. The final volume will be 11.5 µL ; ready for use directly in end-prep for library preparation.

Note

For samples which do not have sufficient PCR yield (200 fmol of combined GI and GII cleaned, semi-nested PCR products) undiluted, cleaned GI and GII PCR products can pooled and input into end-prep. Adjustment of the quantity of end-prepped DNA input into native barcoding can then be made by varying the sample volume from 0.75-3 µl, reducing the volume of water accordingly, to retain 10 fmol of input DNA.

70 Calculate the volume of water required to dilute 8 µL of the cleaned semi-nested GI and GII PCR products to 14.83 fmol/µL and 19.94 fmol/µL.

70.1 For each sample, convert the GI and GII Qubit derived PCR product concentrations from ng/µL to fmol/µL by inputting into software such as the NEBioCalculator ensuring to enter the amplicon lengths of 1110 bp for GI and 971 bp for GII.

70.2 Calculate the GI dilution factor by dividing the fmol/µL of the PCR product by 14.83

70.3 Calculate the GII dilution factor by dividing the fmol/µL of the PCR product by 19.94

70.4 Calculate the volumes of water required by subtracting 1 from the dilution factor and then multiplying by 8

71 Vortex and then briefly spin down the cleaned, semi-nested PCR products

72 Separately aliquot 8 µL of cleaned, semi-nested GI and GII PCR products into new reaction tubes

73 Add the volumes of water calculated in the previous steps to the GI and GII amplicons

74 Cap the reaction vessels

75 Briefly vortex and then spin down

76 In new reaction tubes, for every sample, combine 5.75 µL of both the diluted GI and GII PCR products for every sample undergoing analysis

77 Briefly vortex and then spin down

Native Barcoding and Sequencing

78 Perform end-prep and barcoding following the manufacturer's instructions for ligation sequencing of amplicons in the Native Barcoding Kit 96 V14 by Oxford Nanopore Technologies

79 Load 45 fmol of library per flow cell, using an amplicon size of 1 kb for molarity calculation

80 Run sequencing at 260 bps using super-accurate basecalling until ≈60,000 reads per barcode have been generated

Set Up Software Environment

81 Install all the required software using mamba:

```
$ mamba create --name=amplion_analysis_norovirus duplex-tools=0.2.14
cutadapt=3.4 seqtk=1.3 minimap2=2.24 yacrd=1.0.0 kma=1.4.9 seqkit=2.3.0
samtools=1.13 bedtools=2.30.0 cd-hit=4.8.1 pyfastx=0.8.4

$ mamba create --name=NGSpeciesID medaka=1.2.2 bcftools=1.11 python=3.6.10
perl=5.32.0 openblas=0.3.3 spoa=4.0.7 racon=1.4.20 minimap2=2.17
tensorflow=2.4.1

$ mamba activate NGSpeciesID

$ pip install NGSpeciesID==0.1.3
```

82 Edit the NGSpeciesID script:

```
$ nano ~/mambaforge/envs/NGSpeciesID/bin/NGSpeciesID
```

83 Change the behaviour of the abundance_cutoff parameter so that it is based on the minimum coverage, rather than the relative abundance of a sequence within the sample:

```
-    abundance_cutoff = int( args.abundance_ratio * len(read_array))
+    abundance_cutoff = args.abundance_ratio
```

84 Edit NGSpeciesID's consensus module:

```
$ nano ~/mambaforge/envs/NGSpeciesID/lib/python3.6/site-
packages/modules/consensus.py
```

85 Change the behaviour of NGSpeciesID so it is possible to output consensus sequences from spoa without polishing the consensus sequences using Medaka:

```
+    polishing = False
     if args.medaka:
         polishing_pattern = os.path.join(args.outfolder, "medaka_cl_id_*")
+        polishing = True
     elif args.racon:
         polishing_pattern = os.path.join(args.outfolder, "racon_cl_id_*")
+        polishing = True

-    for folder in glob.glob(polishing_pattern):
-        shutil.rmtree(folder)
+    if (polishing == True):
+        for folder in glob.glob(polishing_pattern):
+            shutil.rmtree(folder)

     spoa_pattern = os.path.join(args.outfolder, "consensus_reference_*")
     for file in glob.glob(spoa_pattern):
```

Split and Trim Reads

86 Create a folder to store results from the analysis.

```
$ mkdir Analysis_Results

$ cd Analysis_Results
```

87 Copy the fastq_pass or pass folder from the sequencing run to the Analysis_Results folder.

88 Store the name of the barcode and sample to be analysed as a variable:

```
$ barcodeID=barcode09

$ sampleID=Sample-4_PCR-Norm_PoolingType-1
```

89 Activate the correct software environment:

```
$ mamba activate amplion_analysis_norovirus
```

90 Use duplex_tools to split the reads:

```
$ mkdir  Split_Reads

$ duplex_tools split_on_adapter --threads 12 --allow_multiple_splits \
fastq_pass/${barcodeID} Split_Reads/${barcodeID} Native
```

91 Create a text file to store the sequence of different primers being used in the experiment:

```
$ nano primer_sequences.fasta
```

Copy the following into the file and then save it

```
>Long_GI_PCR1
AARYTVCCHATHAARGTTGGNATG...TGTAYAATGGNTGGGTNGG
>Long_GI_PCR2
GATGCDGAYTAYACRGCHTGGG...TGTAYAATGGNTGGGTNGG
>Long_GII_PCR1
CWGCAGCMCTDGAAATCATGG...ATGTAYAAYGGDYATGCNGGYGG
>Long_GII_PCR2
GTGTGRTKGATGTGGGTGACTT...ATGTAYAAYGGDYATGCNGGYGG
```

92 Use cutadapt to trim primer sequences from reads:

```
$ mkdir Trim_Reads

$ cat Split_Reads/${barcodeID}/*_split.fastq.gz >
Split_Reads/${barcodeID}/combined.fastq.gz

$ cutadapt -j12 --action=trim -n 1 -e 0.30 -O 12 --revcomp \
-g file:primer_sequences.fasta --discard-untrimmed \
--output Trim_Reads/trimmed-${sampleID}-{name}.fastq.gz \
Split_Reads/${barcodeID}/combined.fastq.gz \
> Trim_Reads/trimmed-${sampleID}.log 2>&1
```

93 Update the sampleID and barcodeID variables, and repeat steps described in this section for any other samples in the sequencing library.

Detect Chimeras

94 Select reads which include primers from the 2nd round of PCR and are longer than 800bp, then randomly sample 90,000 reads for subsequent filtering and chimera removal:

```
$ cat Trim_Reads/trimmed-${sampleID}-*_G*_PCR2.fastq.gz \
> Trim_Reads/trimmed-${sampleID}-combined-PCR2.fastq.gz

$ seqtk seq -L 800 Trim_Reads/trimmed-${sampleID}-combined-PCR2.fastq.gz |\
seqtk sample - 90000 > Trim_Reads/trimmed-${sampleID}-combined-PCR2-
gt800bp.fastq
```

95 Use minimap2 to carry out an all-vs-all alignment of reads:

```
$ mkdir Detect_Chimeric_Reads

$ minimap2 -k19 -Xw19 -e0 -m100 -r100 -I 30M --cap-kalloc=8000m --cap-sw-
mem=100m -t 12 \
Trim_Reads/trimmed-${sampleID}-combined-PCR2-gt800bp.fastq \
Trim_Reads/trimmed-${sampleID}-combined-PCR2-gt800bp.fastq \
> Detect_Chimeric_Reads/trimmed-${sampleID}-combined-PCR2.overlap.paf
```

96 Use yacrd to identify chimeras and reads with poor support:

```
$ yacrd -t12 -c 10 -n 0.2 \
-i Detect_Chimeric_Reads/trimmed-${sampleID}-combined-PCR2.overlap.paf \
-o Detect_Chimeric_Reads/trimmed-${sampleID}-combined-PCR2.overlap-gt800bp.report.yacrd

$ awk '$1=="NotBad"' \
Detect_Chimeric_Reads/trimmed-${sampleID}-combined-PCR2.overlap-gt800bp.report.yacrd |\
cut -f2 \
> Detect_Chimeric_Reads/trimmed-${sampleID}-combined-PCR2.overlap-gt800bp.report.lst

$ seqtk subseq Trim_Reads/trimmed-${sampleID}-combined-PCR2-gt800bp.fastq \
Detect_Chimeric_Reads/trimmed-${sampleID}-combined-PCR2.overlap-gt800bp.report.lst \
> Trim_Reads/trimmed-${sampleID}-combined-PCR2-gt800bp.filtered.fastq

$ rm  Detect_Chimeric_Reads/trimmed-${sampleID}-combined-PCR2.overlap.paf
```

97 Update the sampleID and barcodeID variables, and repeat the steps described in this section for any other samples in the sequencing library.

Creating Representative Sequences

98 Cluster reads to generate consensus sequences:

```
$ mkdir -p Consensus_Reads/${sampleID}/

$ mamba activate NGSpeciesID

$ NGSpeciesID --t 20 --q 15 --ont --consensus --max_seqs_for_consensus
100000 \
--abundance_ratio 100 --m 1100 --s 200 \
--fastq Trim_Reads/trimmed-${sampleID}-combined-PCR2-gt800bp.filtered.fastq
\
--outfolder Consensus_Reads/${sampleID}/ \
> Consensus_Reads/${sampleID}/NGSpeciesID.log 2>&1 \
|| echo "No Assembly"
```

99 Align reads against consensus sequences, call variants and identify any poorly supported regions:

```
$ mamba activate amplion_analysis_norovirus

$ cat Consensus_Reads/${sampleID}/consensus_reference_*.fasta \
> Consensus_Reads/${sampleID}_consensus_seqs.fasta

$ kma index -NI -i Consensus_Reads/${sampleID}_consensus_seqs.fasta \
-o Consensus_Reads/${sampleID}_consensus_seqs

$ mkdir Align_vs_Consensus_Seqs

$ kma -i Trim_Reads/trimmed-${sampleID}-combined-PCR2.fastq.gz \
-o Align_vs_Consensus_Seqs/${sampleID}_consensus_seqs_kma \
-t_db Consensus_Reads/${sampleID}_consensus_seqs \
-vcf 1 -ConClave 2 -bcNano -bc 0.7 -bcd 100 -t 20 -md 100 -1t1 \
-ont -ml 900 -xl 1100 -ef -mrs 0.92 -mrc 0.90 \
> Align_vs_Consensus_Seqs/${sampleID}_consensus_seqs_kma.log 2>&1
```

100 Update the sampleID and barcodeID variables, and repeat the last two steps for any other samples in the sequencing library.

101 Combine consensus sequences from every sample into a single file and rename them:

```
$ mkdir Cluster_Consensus_Sequences

$ cat Align_vs_Consensus_Seqs/*kma.fsa >
Cluster_Consensus_Sequences/combined_seqs.fasta

$ seqtk rename Cluster_Consensus_Sequences/combined_seqs.fasta OTU_ \
> Cluster_Consensus_Sequences/combined_seqs_renamed.fasta
```

102 Trim consensus sequences to remove any poorly supported regions identified by kma:

```
$ seqkit locate --bed -P -r -p '^[agct]+' -p '[agct]+$' \
Cluster_Consensus_Sequences/combined_seqs_renamed.fasta \
> Cluster_Consensus_Sequences/combined_seqs_renamed.bed

$ samtools faidx Cluster_Consensus_Sequences/combined_seqs_renamed.fasta

$ bedtools sort -i Cluster_Consensus_Sequences/combined_seqs_renamed.bed \
-g Cluster_Consensus_Sequences/combined_seqs_renamed.fasta.fai \
> Cluster_Consensus_Sequences/combined_seqs_renamed_sorted.bed

$ bedtools complement -i
Cluster_Consensus_Sequences/combined_seqs_renamed_sorted.bed \
-g Cluster_Consensus_Sequences/combined_seqs_renamed.fasta.fai \
> Cluster_Consensus_Sequences/combined_seqs_renamed_sorted_inverse.bed

$ bedtools getfasta -fi
Cluster_Consensus_Sequences/combined_seqs_renamed.fasta \
-bed Cluster_Consensus_Sequences/combined_seqs_renamed_sorted_inverse.bed \
-fo Cluster_Consensus_Sequences/combined_seqs_renamed_trimmed.fasta
```

103 Cluster consensus sequences at 95 percent sequence identity:

```
$ cd-hit-est -i
Cluster_Consensus_Sequences/combined_seqs_renamed_trimmed.fasta \
  -o
Cluster_Consensus_Sequences/combined_seqs_renamed_trimmed_clustered.fasta \
  -G 0 -c 0.95 -n 10 -d 0 -M 100000 -T 80 -g 1 -aL 0.9
```

Calculating Coverage

104 Align reads against consensus sequences to calculate coverage:

```
$ mkdir Align_vs_Single_Set_Consensus_Reads

$ kma index -NI \
-i Cluster_Consensus_Sequences/combined_seqs_renamed_trimmed_clustered.fasta
\
-o Cluster_Consensus_Sequences/combined_seqs_renamed_trimmed_clustered

$ kma -i Trim_Reads/trimmed-${sampleID}-combined-PCR2.fastq.gz \
-o Align_vs_Single_Set_Consensus_Reads/${sampleID}_consensus_seqs_kma \
-t_db Cluster_Consensus_Sequences/combined_seqs_renamed_trimmed_clustered \
-nc -vcf 2 -ConClave 2 -bcNano -bc 0.7 -bcd 100 -t 20 -md 100 -1t1 \
-ont -ml 900 -xl 1100 -ef -mrs 0.92 -mrc 0.90 \
> Align_vs_Single_Set_Consensus_Reads/${sampleID}_consensus_seqs_kma.log
2>&1
```

105 Use the following files for downstream statistical analysis:

The coverage of each representative sequence in each sample:

```
Align_vs_Single_Set_Consensus_Reads/*_consensus_seqs_kma.mapstat
```

The sequence of each representative sequence:

```
Cluster_Consensus_Sequences/combined_seqs_renamed_trimmed_clustered.fasta
```

Collating Coverage, Sequencing and Typing Results

106 Use the CDC's calicivirus typing tool to type each representative sequence:

https://calicivirustypingtool.cdc.gov/

107 Upload the entire fasta file

108 Click on 'type it'

109 Once the input sequences have been typed click on 'Download EXCEL'

110 Install R (≥ 4.1.2), RStudio (≥ 1.4.1717) and the openxlsx package (≥ 4.2.5.2)

111 Setup a new project directory using RStudio

112 Download the coverage files to a sub-folder within the project directory called 'Coverage Results'

113 Copy the Excel file downloaded from the CDC's calicivirus typing tool website to the project directory

114 Create a new R script in your RStudio project

115 Copy the following code into the R script:

```
# Load library for opening xlsx files
library(openxlsx)

# List coverage files
coverage_results =
  list.files(path = 'Coverage Results', pattern =
"\\_consensus_seqs_kma.mapstat",
             recursive = TRUE, include.dirs = FALSE)

# List samples
samples = gsub(x = coverage_results,
               pattern = "^(.+)_consensus_seqs_kma.mapstat", replacement =
"\\1")

# Collate coverage results into a single file
combined_coverage_results = NULL

for (sampleID in 1:length(samples)) {
  print(sampleID)
  coverage_filename =
    paste0("Coverage Results/", samples[sampleID],
"_consensus_seqs_kma.mapstat")

  coverage_file = readLines(con = coverage_filename)

  coverage_file[[7]] = gsub(x = coverage_file[[7]], pattern = "# ",
replacement = "")

  noRecords = sum(!grepl(pattern = "##", x = coverage_file,fixed = TRUE)) -
1

  if (noRecords > 0) {
    coverage = read.table(file = textConnection(coverage_file),
                          header = TRUE, stringsAsFactors = FALSE, sep =
"\t", skip = 6)

    coverage$sample = samples[sampleID]

    combined_coverage_results = rbind(combined_coverage_results, coverage)
  }

}

# Import file linking reference sequences to genotype
```

```
taxa_results_field_names =
  c('query','name','sequence','length','Genus','Genotype-B-region-score',
    'Genotype-B-region-plot','Genotype','Genotype-C-region-plot',
    'Genotype-C-region-score','Genotype-plot')

combined_taxa_results =
  read.xlsx(xlsxFile = "calicivirus_typing_output.xlsx",
            sheet = "sheet1", startRow = 2)

colnames(combined_taxa_results) = taxa_results_field_names

# Merge coverage results with genotype
combined_coverage_results2 =
  merge(x = combined_coverage_results,
        y = combined_taxa_results[,c('name','Genotype')],
        by.x = c('refSequence'), by.y = c('name'), all.x = TRUE)

combined_coverage_results2[
  is.na(combined_coverage_results2$Genotype),
  c('Genotype')
  ] = "Unknown"

# Summarise by genotype
coverage_per_genotype =
  aggregate(formula = readCount ~ Genotype + refSequence + sample,
            data = combined_coverage_results2,
            FUN = sum)

coverage_per_genotype_wide =
  reshape(data = coverage_per_genotype,
          v.names = "readCount",
          idvar = c("Genotype","refSequence"),
          timevar = "sample", direction = "wide")

coverage_per_genotype_wide[is.na(coverage_per_genotype_wide)] = 0

colnames(coverage_per_genotype_wide) =
  gsub(x = colnames(coverage_per_genotype_wide),
       pattern = "^readCount\\.(.+)$", replacement = "\\1")

# Save results as a csv file
write.csv(x = coverage_per_genotype_wide,
          file = "coverage_per_genotype_wide.csv", row.names = FALSE)
```

116 Run the R script, check for errors and that the output looks correct.

Read Depth Filtering

117 Using the coverage_per_genotype_wide.csv file generated above, for GI and GII independently, sum the norovirus-aligned reads for each sample

118 Find the median of the total aligned reads per sample across the data set

119 Filter data on a sample-to-sample basis removing reads associated with consensus sequences that have fewer than 0.1% of the median total aligned reads per sample

Reference Database Removal of PCR Chimeras

120

Note

PCR chimeras should be present at relatively low abundances. It is worth taking note of the No. of reads aligned with any putative chimeras to avoid the removal of novel recombinants without further investigation.

121 Collate a list of all known types of norovirus GI and GII from the Centre for Disease Control's (CDC) Human Calicivirus Typing Tool – this should be performed prior the analysis of any new data sets to prevent missing recently observed types

122 Remove all coverage (coverage_per_genotype_wide.csv) and sequence data (combined_seqs_renamed_trimmed_clustered.fasta) relating to consensus sequences that are types not

previously reported by the CDC

Note

This can be done manually for small data sets or using R and tidyverse 2.0.0.

De Novo Removal of PCR Chimeras

123 For each wastewater sample under analysis, annotate the consensus sequences (combined_seqs_renamed_trimmed_clustered.fasta) with the read depths (coverage_per_genotype_wide.csv)

Note

This can be done manually or with R using seqinr 4.2.3 and tidyverse 2.0.0.

124 Perform chimera filtering with USEARCH v11 using the following parameters: -uchime3_denovo -chimeras chimeras.fa

125 Open chimeras.fa to identify potentially chimeric sequences for that sample

126 Remove all coverage (coverage_per_genotype_wide.csv) and sequence data (combined_seqs_renamed_trimmed_clustered.fasta) relating to consensus sequences identified as chimeras in chimeras.fa

Manual Screening for PCR Chimeras

127 For each sample, identify types that may be generated from two other (parent) types in the sample using the data in coverage_per_genotype_wide.csv

Note

E.g. GI.6[P13] may be a chimera of GI.6[P11] and GI.3[P13] if the latter two have a higher number of reads associated with them.

128 Assess the putative chimeras and parent sequences using a multiple sequence alignment tool such as NCBI Multiple Sequence Alignment Viewer v1.25.0 with the putative chimera sequence anchored

129 Check the parent sequences for breakpoints within the terminal or proximal regions of *RdRp* and *VP1* and child-parent sequence similarities ≥95%

130 Remove all coverage (coverage_per_genotype_wide.csv) and sequence data (combined_seqs_renamed_trimmed_clustered.fasta) relating to consensus sequences identified as putative chimeras which match these criteria

131

Note

Your data is now ready for downstream analysis. It should be noted, however, that at present this method is only suitable for qualitative analysis and read-depth/coverage should not be used as a quantitative measurement of the different types detected in a sample.