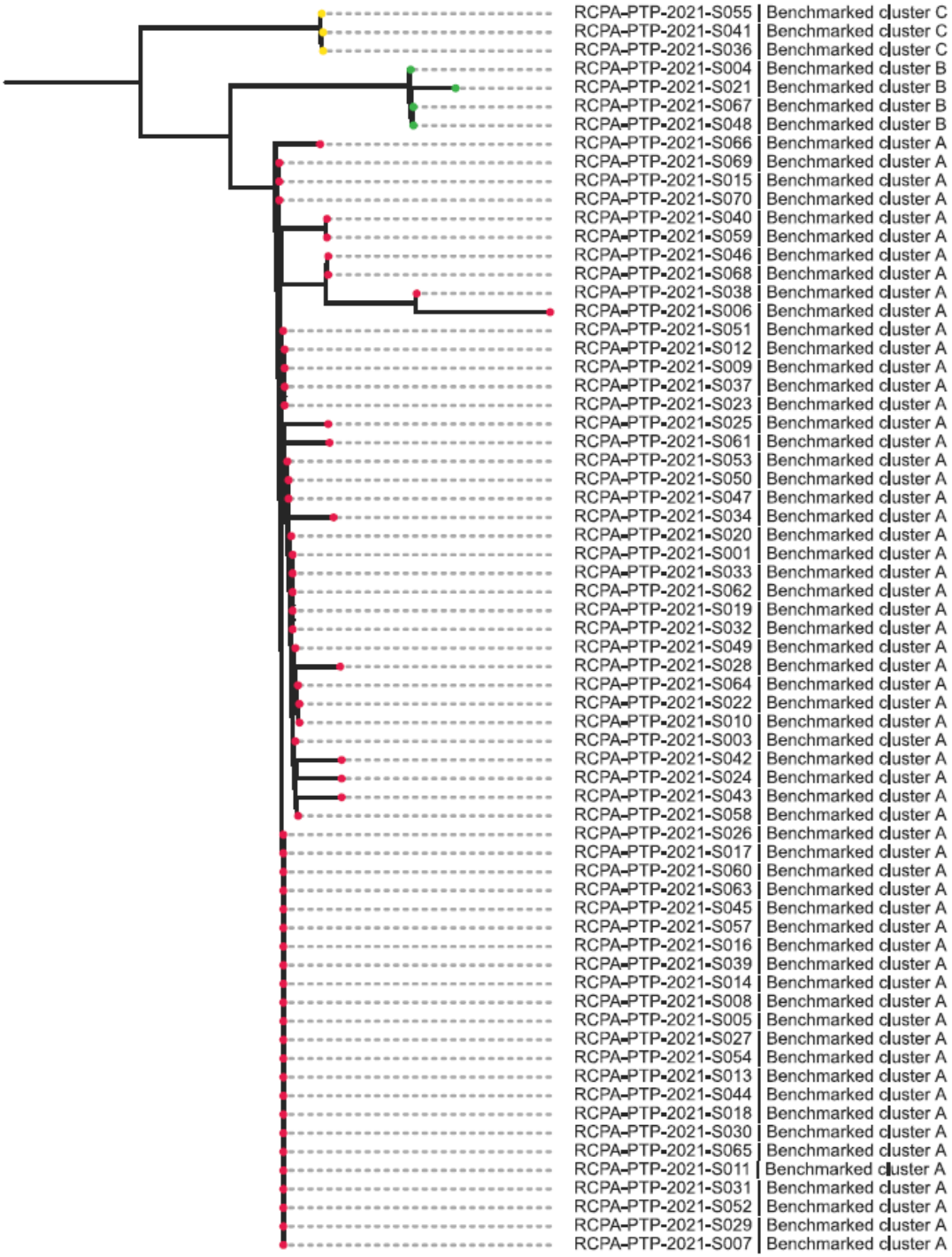
APPENDIX A: SUPPLEMENTARY DATA

Proficiency testing for SARS-CoV-2 whole genome sequencing

Supplementary Fig. 1



Supplementary Table 1

Sample ID	Participant ID & discordant SNP results[1]										
	1	2	3	4	5	6	7	8	9	10	11
BS-SC-21-01	C23615S*					G2644T Missed: G26144T	C23615G	C29750T			C23615S
BS-SC-21-02											
BS-SC-21-03	Missed: G28882A, G28883C					Missed:GGG28881-28883AAC				Missed: All SNPs	U28196C
BS-SC-21-04	C22006M* Missed: G28882A, G28883C	C22006A	C22006A			Missed:GGG28881-28883AAC	C22006G			Missed: All SNPs	C22006M U7540C
BS-SC-21-05	G22205S*				Missed: G22205C	Missed:GGG28881-28883AAC, G22205C		Missed: C14408T		Missed: All SNPs	G22205S U7540C
BS-SC-21-06											
BS-SC-21-07											
BS-SC-21-08	G2012S* A2810R* G7115S* G7469S* T16742Y* T28588Y* C28868Y* G5554C G11114T Missed: C3037T, A23403G		A5874G							Missed: All SNPs	

[1] Concordant results were marked in grey boxes: absence of C241T is considered concordant.
Discordant, with additional SNP results were marked in red boxes: participant 10 identified a list of unmatched SNPs for sample BS-SC-21-03, BS-SC-21-04, BS-SC-21-05 and BS-SC-21-08 that are not listed here.
Discordant, with unreported SNPs (indicated here as missed) were marked in blue boxes.
Ambiguous results were marked in orange boxes: participant submitted SNPs with indication of nomenclature consisted of M, S, R and Y, which did not match with the sequence analysis results performed in-house at the pre-issue testing (S = C or G; M = A or C; R = A or G; Y = C or T).

Supplementary Table 2

Sample ID	Participant ID & discordant amino acid replacement results[1]										
	1	2	3	4	5	6	7	8	9	10	11
BS-SC-21-01	S:R685G*						S:R685G		ORF1ab:L6267P		S:R685G
BS-SC-21-02									Missed: nsp12:P323L		
BS-SC-21-03	ORF14:G50E			Missed: N:G204R		Missed: N:G204R			Missed: nsp12:P323L	Missed: All	
BS-SC-21-04	ORF14:G50E	S: N148K	S: N148K	Missed: N:G204R		Missed: N:G204R	S:N148K		ORF1ab:R5461 ORF1ab:T6097I Missed: nsp12:P323L		S: N148K
BS-SC-21-05	N:N45I ORF9b:I45L ORF14:G50E Missed: N:N48I		ORF9b:I45L Missed: S:D215H ORF10:P10S	Missed: N:G204R S:D215H	Missed: S:D215H	Missed: N:G204R S:D215H	ORF9b:I45L Missed: ORF10:P10S	Missed: S:D215H	ORF1ab:R5461 ORF1ab:T6097I Missed: nsp12:P323L	Missed: S:D215H	
BS-SC-21-06											
BS-SC-21-07									Missed: nsp12:P323L		
BS-SC-21-08	ORF1a:A583P* ORF1a:I849V* ORF1a:K1763N ORF1a:G2284R* ORF1a:D2402H* ORF1b:V1092A* N:P45S* Missed: S:D614G		ORF1a:E1870G						Missed: nsp12:P323L	orf1ab:F871L, orf1ab:C873R, S: N542K, orf1ab:T1653G, orf1ab:K1654Q Missed: All	

[1] Concordant results were marked in grey boxes.
Discordant, with additional amino acid replacement results were marked in red boxes: participant 10 identified a list of unmatched amino acid replacement for sample BS-SC-21-03 and BS-SC-21-08 that are not listed here.
Discordant, with unreported amino acid replacement (indicated here as missed) were marked in blue boxes.
Ambiguous results were marked in orange boxes: participant submitted results which did not match with the sequence analysis results performed in-house at the pre-issue testing.

Supplementary Data

RCPAQAP
The Royal College of Pathologists of Australasia
Quality Assurance Programs

2021 PHLN SARS-CoV-2 and Other Coronaviruses Program

Survey Instructions

Survey Schedule

Survey Number	Open date	Close date
1/1	18 February 2021	17 March 2021

Clinical Notes

Sample ID	Matrix	Clinical Notes
BS-SC-21-01	Simulated Respiratory Sample	N/A
BS-SC-21-02	Simulated Respiratory Sample	N/A
BS-SC-21-03	Simulated Respiratory Sample	N/A
BS-SC-21-04	Simulated Respiratory Sample	N/A
BS-SC-21-05	Simulated Respiratory Sample	N/A
BS-SC-21-06	Simulated Respiratory Sample	N/A
BS-SC-21-07	Simulated Respiratory Sample	N/A
BS-SC-21-08	Simulated Respiratory Sample	N/A

Instructions

Each sample was prepared from total viral RNA material suspended in a 0.5 mL nuclease-free water, simulating a respiratory sample that may contain causative agents of COVID-19 or other coronavirus diseases.

Samples should be treated and tested in the same manner as routine patient samples.

Safe Handling

Before handling please read the RCPAQAP Handling of Proficiency Testing Materials

Consult the relevant Safety Data Sheet (SDS) if you have any queries regarding RCPAQAP samples. These are available for download from myQAP.

Sample storage

All materials should be stored frozen at -80°C at all times.

Sample processing

Part 1. Wet-lab

- Treat the suspension as though it is from a respiratory sample (e.g. nasal swab, bronchoalveolar lavage fluid). Perform RNA extraction and whole genome sequencing according to the laboratory's standard procedure;
  Do not isolate the virus from the specimen using an *in vitro* cell culture.
- Analyse the sequence data using your choice of bioinformatics software. Use the following for your analysis; pangolin v2.1.7 with pangoLEARN version 2021-01-11.
- Report your results in the questionnaire link here (an electronic version of this questionnaire will be available in the email).
- Label FASTQ data for the survey specimen panel with <LABID>_<SAMPLE_ID>_R1.fastq.gz and <LABID>_<SAMPLE_ID>_R2.fastq.gz. The SAMPLE_ID for all specimen are available on Page 1 of this document (Sample ID).
- Label the FASTA file of the consensus sequences for each of the specimens with <LABID>_<SAMPLE_ID>.fasta
- Include your negative control FASTQ data and label it with <LABID>_<NEG>_1_R1.fastq.gz and <LABID>_<NEG>_1_R2.fastq.gz. If submitting multiple NEGATIVE controls (one per run), update the integer suffix with _2_, _3_, ... to identify the distinct set of files.
- For upload of sequence data, a webpage is used. Click here to access with your login details (LABID) and password, which will be available in the email you receive on the day the program opens.
- Submit the raw reads of your sequence data, without pre-processing or trimming, as well as the derived consensus sequence data into the webpage. The detailed instructions to upload this is available on the webpage.

Part 2. Genomic dataset: Dry-lab only

- A genomic dataset for the dry-lab analysis has been created and is available to download. Click here to access with your login details (LABID) and password.
- The file is a multi-FASTA file with 70 consensus genomes of SARS-CoV-2. Using your preferred pipeline, attempt to classify the downloaded sequences into one or more genomic clusters that would be of interest to epidemiological investigations.
- The detailed instructions to upload your clustering results, and if applicable, the NEWICK file (if using a phylogenetic-based method) and the distance matrix file (if using a distance-based method) is available on the webpage.
- Analyse the downloaded sequences using your choice of bioinformatics software and report your results in the questionnaire link here (an electronic version of this questionnaire will be available in the email).

Additional Resources

myQAP portal

Portal help

RCPAQAP Participant Handbook

RCPAQAP Data Analysis and Assessment Criteria Handbook

2021 PHLN: SARS-CoV-2 and other coronaviruses (Part 1)

Welcome to the RCPAQAP 2021 PHLN: SARS-CoV-2 and other coronaviruses Whole Genome Sequencing (WGS) PTP.

This survey opens on 18 Feb and will close on 17 Mar 2021. Please record all your results by answering the following questionnaire.

The results and any information that you submit will be treated confidentially. Please note that the de-identified results and the information obtained will be shared with our collaborators at the Communicable Diseases Genomics Network (CDGN). However, other than this, these details will never be disclosed to a third party without the prior written consent of the participant, unless required by legislation.

The Survey Report will be emailed to you directly when it is ready and should be available in mid-2021 and will not be uploaded into myQAP. While the Report will also be sent to the Australian Department of Health as the funder of RCPAQAP Biosecurity, the results and information will be de-identified.

NOTE:
This Survey Part 1 questionnaire will take approximately 45 minutes to complete. Please only start answering the questionnaire if you can complete it in one sitting as you are NOT able to save your answers and continue later. This will ensure that all your results are recorded and saved.

***** Required

Participant ID

1. Participant name *

2. Name of organization *

3. Participant number *

This information is available in the email sent to you

Snapshot of sequence information

4. % genome recovered *

Separate each specimen with a semi-colon (e.g, BS-SC-21-01 99.5%; BS-SC-21-02 99.7%; etc)

5. Average read depth *

Separate each specimen with a semi-colon (e.g, BS-SC-21-01 read depth; BS-SC-21-02 read depth; etc)

6. Pangolin lineage *

Separate each specimen with a semi-colon (e.g, BS-SC-21-01 pangolin lineage; BS-SC-21-02 pangolin lineage; etc)

7. Single-nucleotide polymorphism (SNP) (if any) *

List each SNP separated by a comma and separate each specimen with a semi-colon (e.g. BS-SC-21-01 T1100C, A1163T; BS-SC-21-02 C9996T, T5740C etc)

8. Amino acid replacement (if any) *

List each mutated protein, followed by a colon and its amino acid replacement, each of the mutated protein is separated by a comma and separate each specimen with a semi-colon (e.g. BS-SC-21-01 nsp4: S481L, N: G204R; BS-SC-21-02 N: N48I, S: S477N; etc)

Report on the protocol used to generate the sequence data

9. Sequencing site *

- [ ] Within own laboratory
- [ ] Externally

10. RNA extraction protocol *

- ◯ Commercial kit
- ◯ Non-commercial assay

11. If using a commercial kit for RNA extraction, specify the name and briefly list relevant deviations from the commercial kit (if applicable)

12. If using a non-commercial assay for RNA extraction, provide a short summary of the protocol

13. RNA extraction system *

Specify the name of the instrument used (if using an automated system)

14. cDNA synthesis method *

15. Sequencing approach *

- ◯ Amplicon-based sequencing
- ◯ Hybrid capture / Target enrichment
- ◯ Shotgun metagenomics
- ◯ Other

16. Other; specify

17. Amplicon size

- ◯ 400 bases
- ◯ 1.2 kilobases
- ◯ 2.5 kilobases
- ◯ Other

18. Other; specify

19. For commercial kit used, please select the following

- ◯ SARS-CoV-2 specific
- ◯ Part of a panel
- ◯ Other

20. Identify the panel

21. Other; specify

22. Amplicon primer scheme

- ◯ ARTIC V2
- ◯ ARTIC V3
- ◯ COVIDSeq
- ◯ Other

23. Other; specify

Please include the number of amplicons and number of PCR pools

24. Any variations to the primer scheme

25. Were there any primer sets that were replaced during the sequencing effort?

- ◯ Yes
- ◯ No

26. How many primer sets were replaced?

27. How were the amplicons quantified?

28. The protocol used to prepare the library for sequencing *

- ◯ Commercial kit
- ◯ Non-commercial assay

29. Specify the name and briefly list relevant deviations from the commercial kit (if applicable)

30. A short summary of the protocol used to prepare the library for sequencing

31. The sequencing platform used *

- ◯ iSeq
- ◯ MiniSeq
- ◯ MiSeq
- ◯ NextSeq
- ◯ Ion Torrent
- ◯ MinION (Oxford Nanopore)
- ◯ Flongle (Oxford Nanopore)
- ◯ Other platform

32. Other platform; specify

33. Was the sequencing run monitored? *

- ◯ Yes
- ◯ No
- ◯ Not applicable

34. How was the run monitored, i.e. which tool was used to show mapping results in real-time, including genome coverage etc?

35. The sequence read type used

- ◯ Paired-end
- ◯ Single-end
- ◯ Not relevant

36. The sequence read length used

- ◯ 75 bp
- ◯ 100 bp
- ◯ 150 bp
- ◯ 250 bp
- ◯ 300 bp
- ◯ Other
- ◯ Not relevant

37. Other; specify

Report the criteria you used to analyse the sequence data

38. The acceptable % of successfully sequenced individual amplicons

- ◯ Above 90%
- ◯ Above 95%
- ◯ Above 98%
- ◯ 100%
- ◯ Other
- ◯ Not applicable

39. Other; provide acceptable %

40. The program used to demultiplex reads (quality-trim); include the version and all parameters used

if applicable

41. The program used to trim the sequence of the primers; include the version and all parameters used

if applicable

42. The version of pangolin currently used in analysing sequence derived from a clinical specimen *

Note: Do not provide the version used in this survey

43. The version of pangoLEARN currently used in analysing sequence derived from a clinical specimen *

Note: Do not provide the version used in this survey

44. Reference-based alignment performed *

- ◯ Yes
- ◯ No

45. The reference sequence ID and indicate whether it is the GISAID sequence or GenBank accession number?

46. Describe the approach used in performing the alignment of the sequence data

47. Reference mapping and consensus calling tool *

- ◯ Custom script
- ◯ Online tool

48. Details of the reference mapping and consensus calling tool

49. Analysis process/workflow engine used *

- ◯ Automated, through graphical user interface
- ◯ Automated, using text-based definition of workflows
- ◯ Manual (in-house pipeline)

50. Which workflow system was used (including the version) and briefly list relevant deviations from the software (if applicable)

51. Describe the approach used in performing the manual analysis (a short summary of the protocol, including details of the QC tools, metrics for trimming if performed and the assembler if used)

52. Consensus sequence calling *

Depth

53. Consensus sequence calling *

Quality

54. Consensus sequence calling *

Minimum frequency threshold

55. The tool used to determine average SNP difference *

56. The approach to remove human reads *

e.g. mapping against the human genome (provide the reference/accession number)

Report the quantitative/qualitative criteria you used for quality control checks

Criteria used to evaluate the quality of the sequence data (quality control checks); if yes, record the threshold/condition used for each criterion. If criteria used is not listed, choose other; specify criteria and record the threshold used

57. % Genome coverage/recovery *

- ◯ Yes
- ◯ No

58. Threshold

59. Sequencing depth *

- ◯ Yes
- ◯ No

60. Threshold

61. SNP distance from the reference genome *

- ◯ Yes
- ◯ No

62. Threshold

63. Ambiguous bases/sites *

- ◯ Yes
- ◯ No

64. Threshold

65. Contamination *

- ◯ Yes
- ◯ No

66. Condition *

67. Other; specify criteria and record the threshold used

Provide a criterion, followed by a colon and the threshold. Separate each criterion with a semi-colon (e.g. Criterion 1: Threshold; Criterion 2: Threshold etc)

Comment & Feedback

68. Comment here, if you experience any difficulties in submitting your response.

    Any feedback that you provide will help us improve our future WGS survey module

2021 PHLN: SARS-CoV-2 and other coronaviruses: Dry-lab ONLY (Part 2)

Welcome to the RCPAQAP 2021 PHLN: SARS-CoV-2 and other coronaviruses Whole Genome Sequencing (WGS) PTP.

This survey opens on 18 Feb and will close on 17 Mar 2021. Please record all your results by answering the following questionnaire.

The results and any information that you submit will be treated confidentially. Please note that the de-identified results and the information obtained will be shared with our collaborators at the Communicable Diseases Genomics Network (CDGN). However, other than this, these details will never be disclosed to a third party without the prior written consent of the participant, unless required by legislation.

The Survey Report will be emailed to you directly when it is ready and should be available in mid-2021 and will not be uploaded into myQAP. While the Report will also be sent to the Australian Department of Health as the funder of RCPAQAP Biosecurity, the results and information will be de-identified.

NOTE:
This Survey Part 2 questionnaire will take approximately 20 minutes to complete. Please only start answering the questionnaire if you can complete it in one sitting as you are NOT able to save your answers and continue later. This will ensure that all your results are recorded and saved.

***** Required

Participant ID

1. Participant name *

2. Name of organization *

3. Participant number *

This information is available in the email sent to you

Analysis of the genomic dataset

4. Select any QC metrics that were used to assess the appropriateness of a sequence for inclusion into an analysis *

- ◯ Genome recovered (%)
- ◯ Distance from a reference
- ◯ Other

5. Other; please specify *

6. Was any filtering of the alignments used? *

- ◯ Yes
- ◯ No

7. If yes, please describe software and/or criteria for filtering *

8. Was Pangolin lineage considered when interpreting genomic relationships? *

- ◯ Yes
- ◯ No

9. If yes, please state version of pangolin and pangoLEARN used *

10. Were the consensus sequences used to generate a phylogenetic tree? *

- ◯ Yes
- ◯ No

11. If yes, please state tree inference software and version *

12. Were distances between sequences calculated? *

- ◯ Yes
- ◯ No

13. If yes, please state the metric (ie SNP distance) and thresholds used *

14. Briefly describe criteria implemented to assess/interpret the degree of the genomic relationships within the dataset *

15. Detail what relationships were identified in the dataset? *

16. Was there an attempt to identify VoC (Variants of concern)? *

- ◯ Yes
- ◯ No

17. If yes, please describe criteria for the assignment of VoC *

18. Was there any VoC identified? *

- ◯ Yes
- ◯ No

19. If yes, please list sequence and VoC identified *

List each VoC separated by a comma and separate each sequence with a semi-colon (e.g. Sequence 1 VoC 1, VoC 2; Sequence 2 VoC 1, VoC 2, VoC 3 etc)

Comment & Feedback

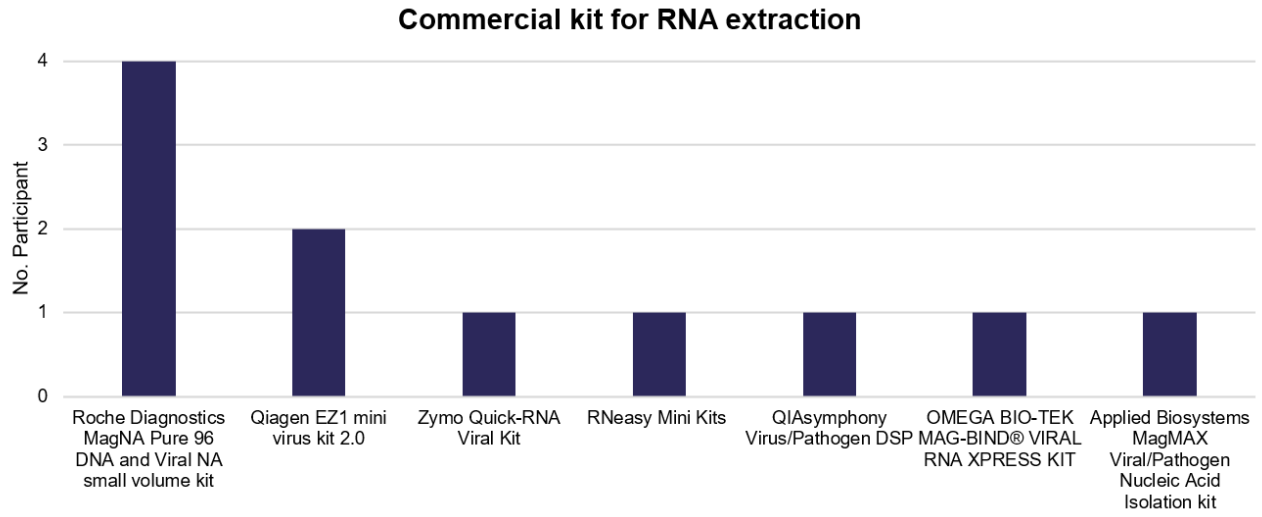
20. Comment here, if you experience any difficulties in submitting your response.

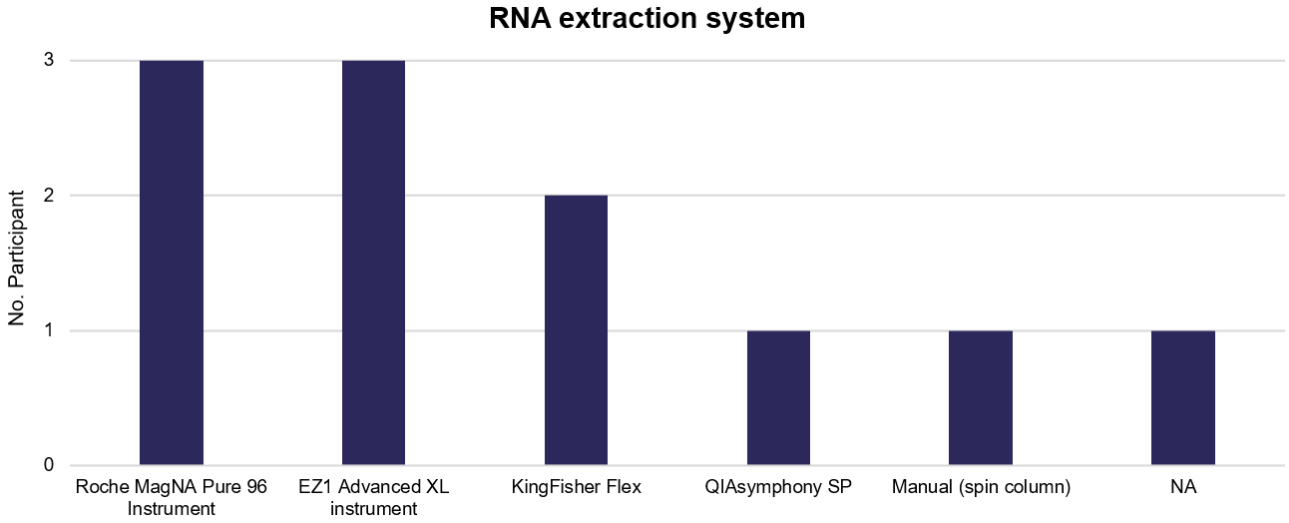
    Any feedback that you provide will help us improve our future WGS survey module

Summary of results submitted (Questionnaire Part 1)

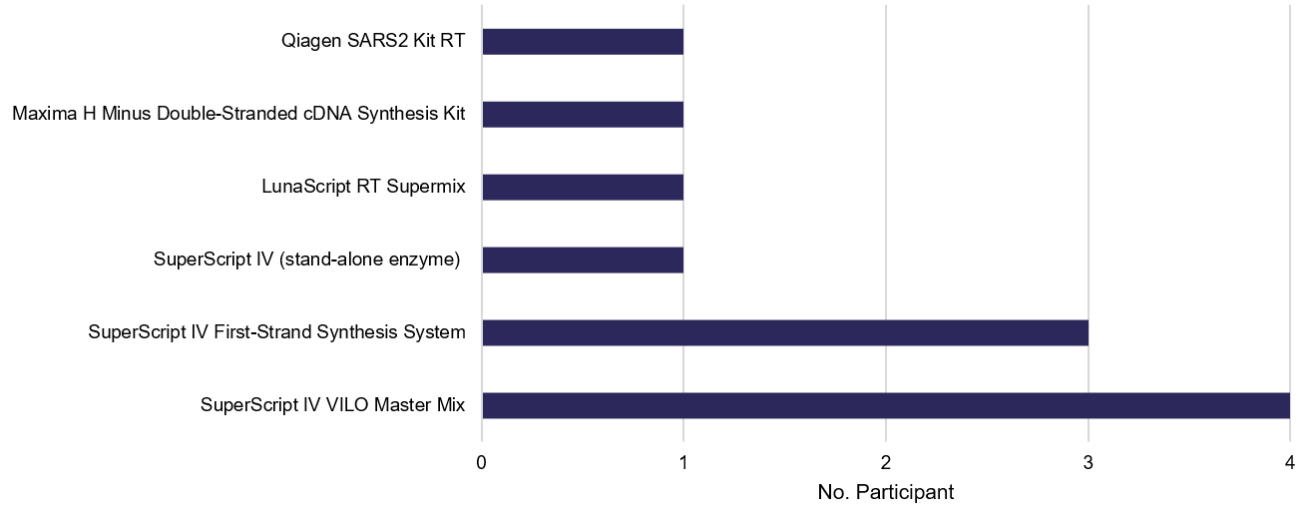
A. RNA extraction

All participants used commercial kits to extract the RNA, with details summarized below.





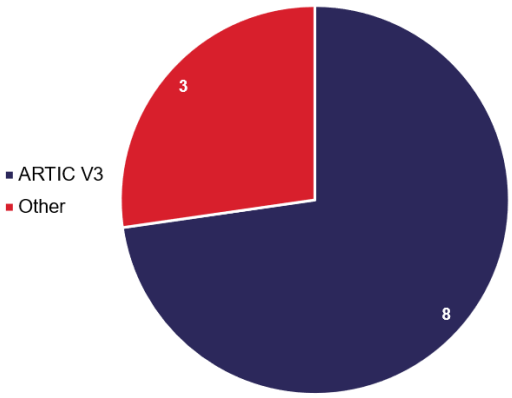
B. cDNA synthesis



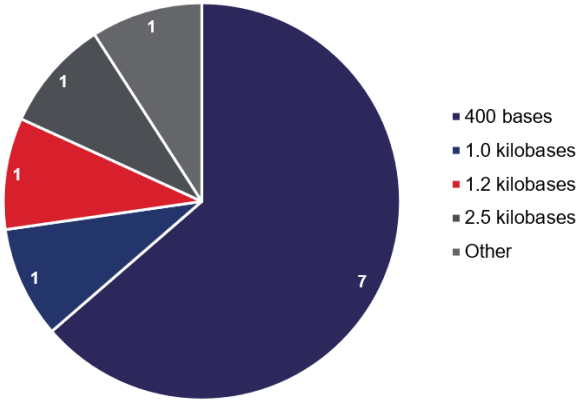
C. Sequencing approach

All participants (n = 11) used the amplicon-based sequencing approach. Details are available below.

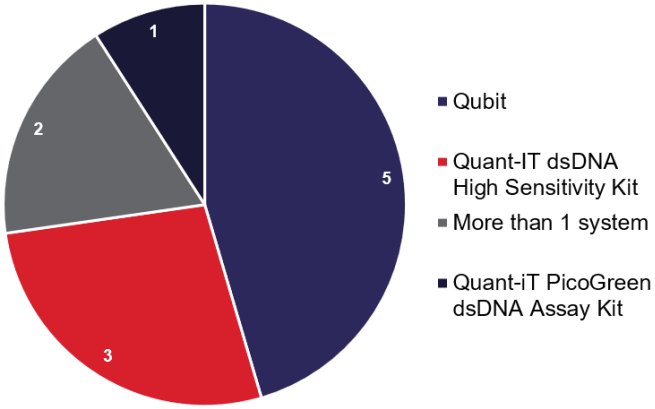
Amplicon primer scheme



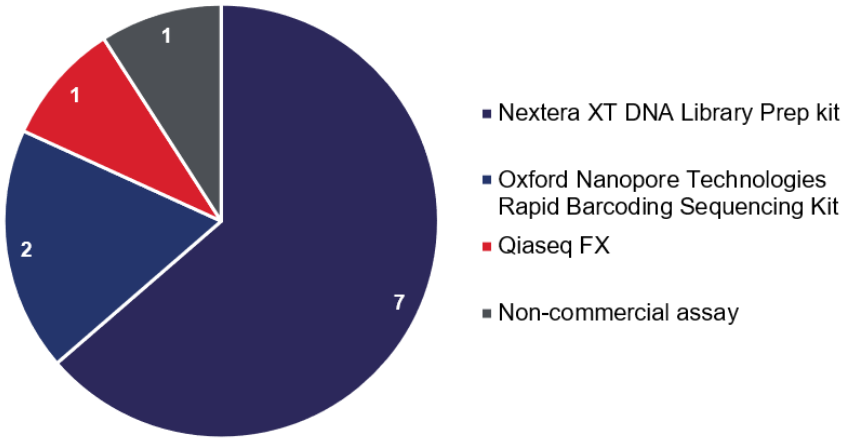
Amplicon size



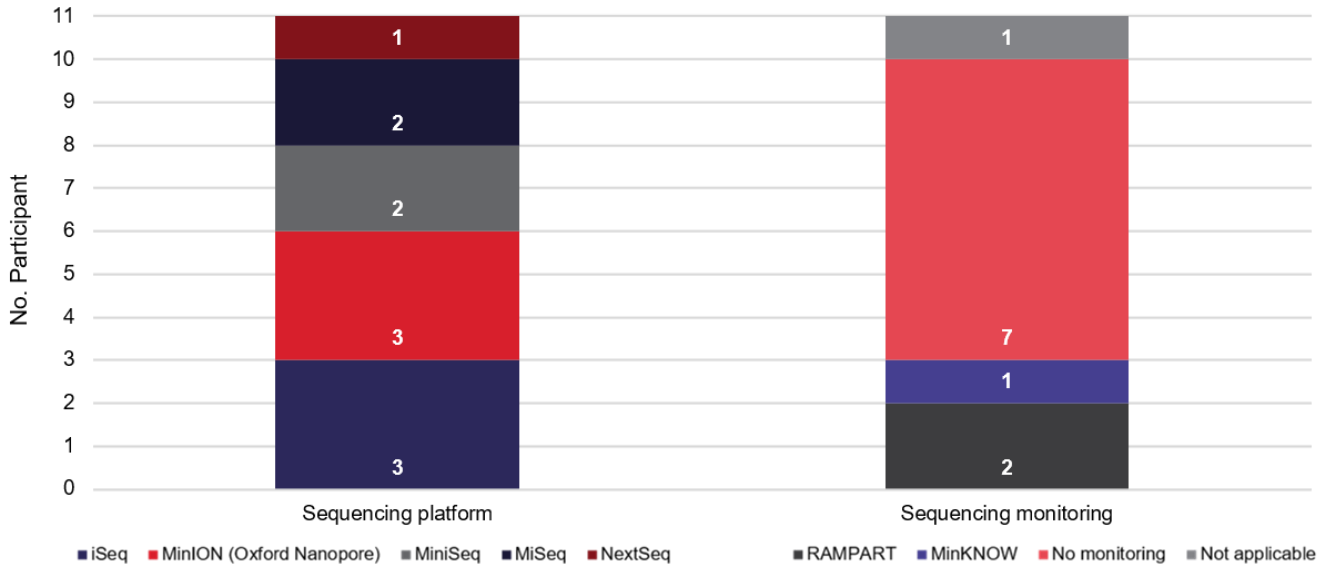
Quantitation of amplicons



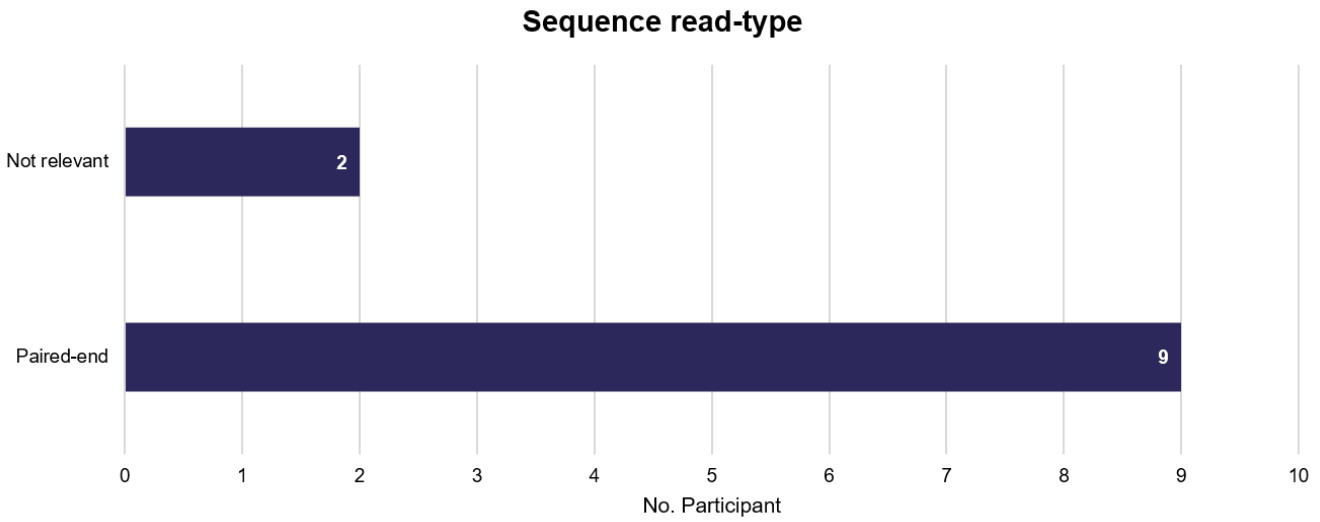
D. Library preparation protocol

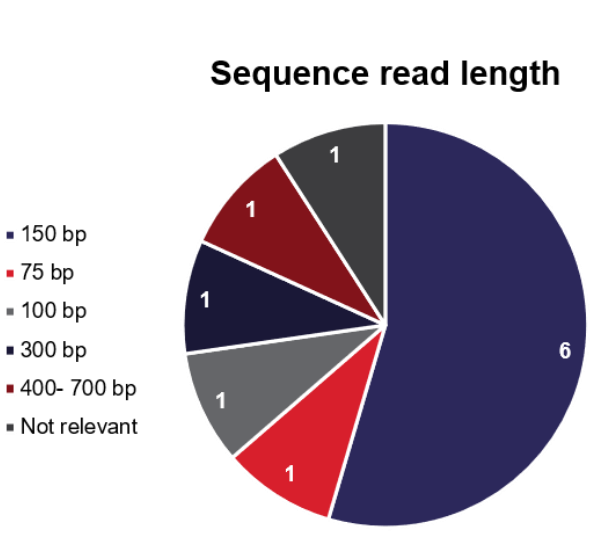


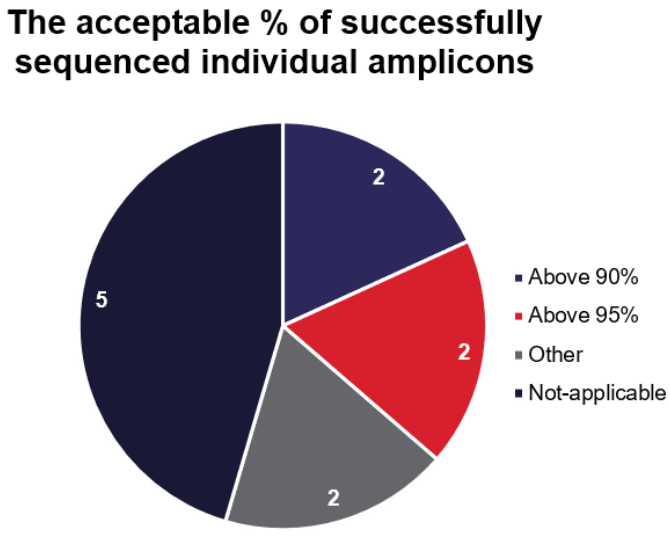
E. Sequencing platform and monitoring system



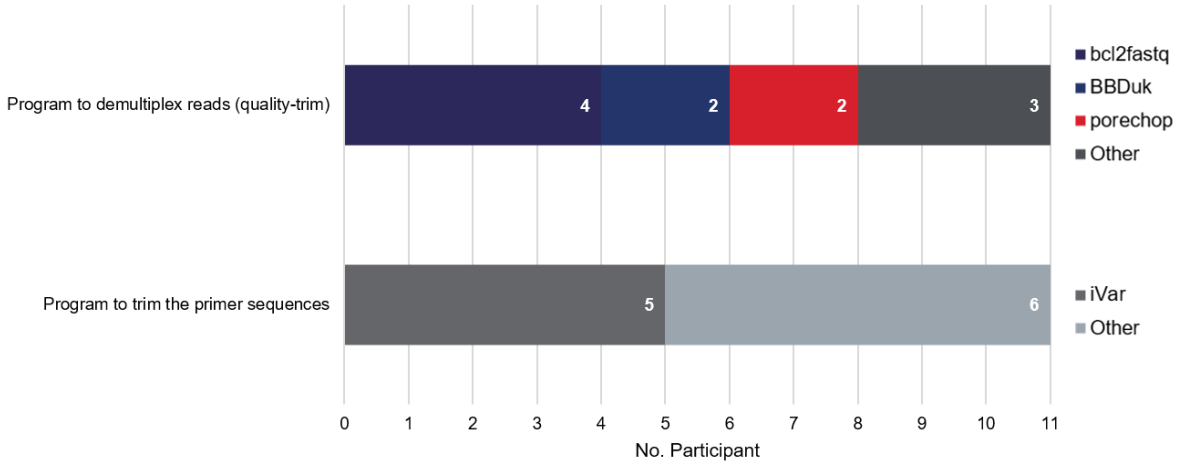
F. Sequencing read



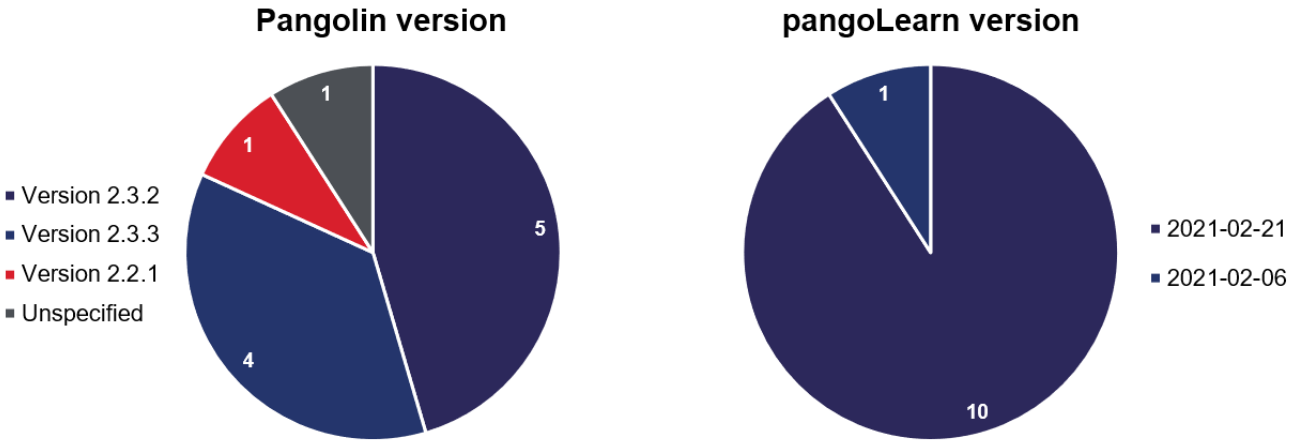




G. Programs used

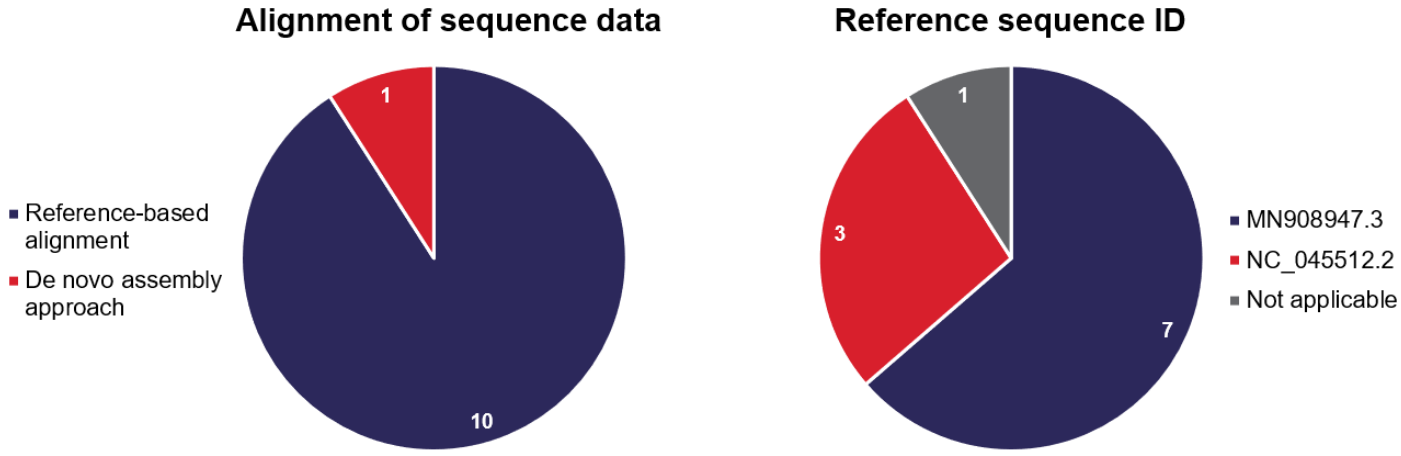


H. Pangolin & pangoLearn

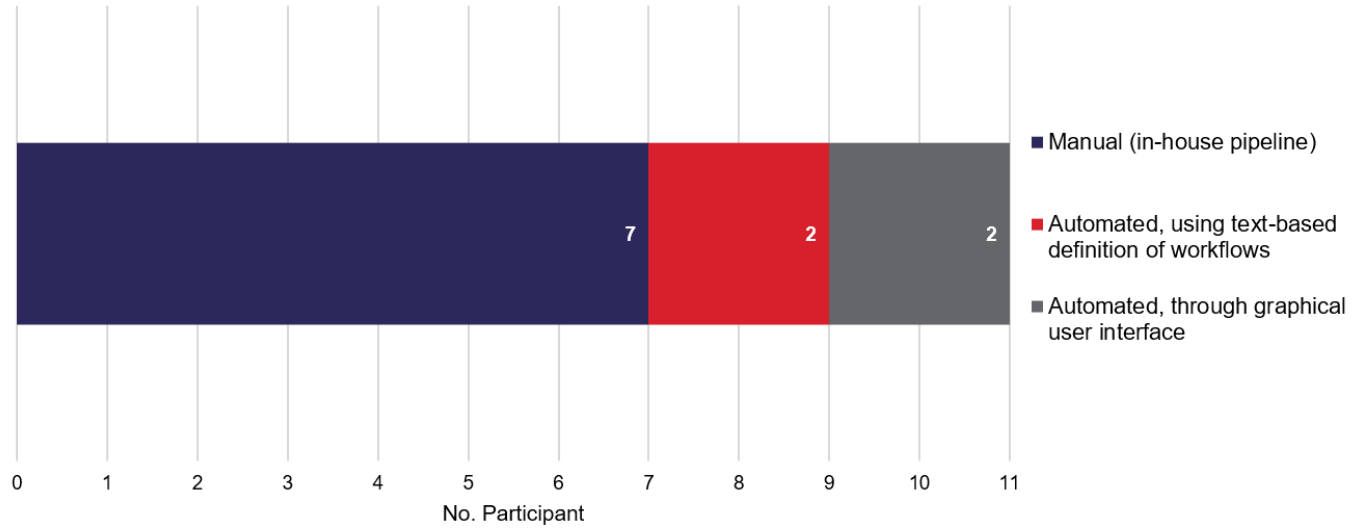


I. Reference mapping

All participants (n = 11) used custom scripts as reference mapping and consensus calling tool. Details of sequence data alignment and reference sequence ID are available, as below.



J. Analysis process (workflow engine used)

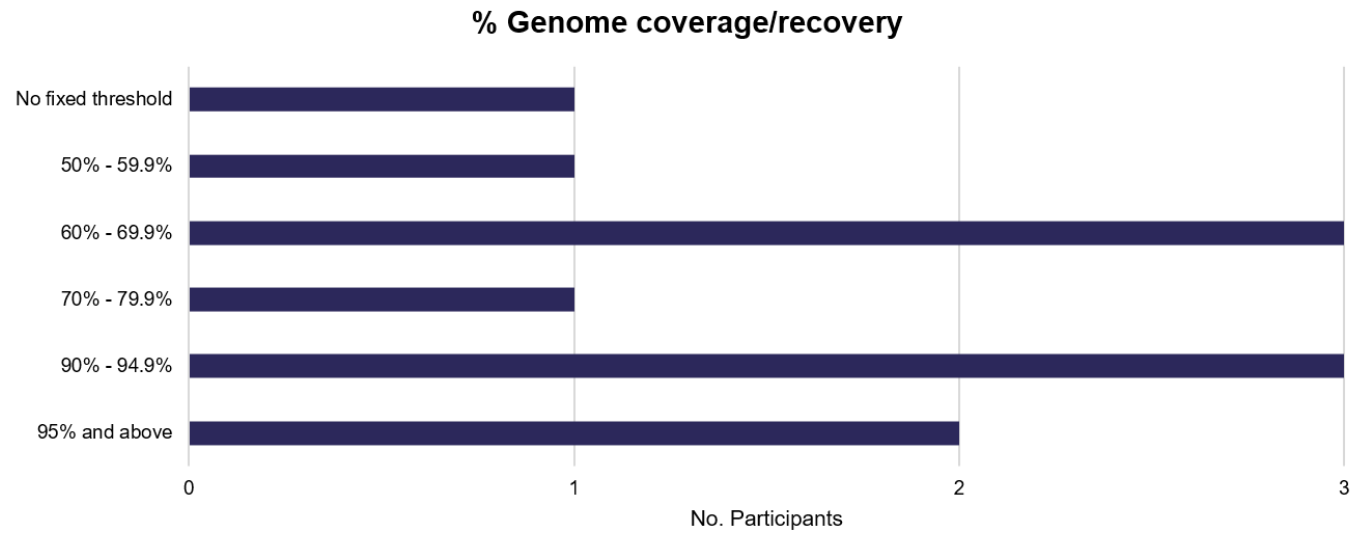


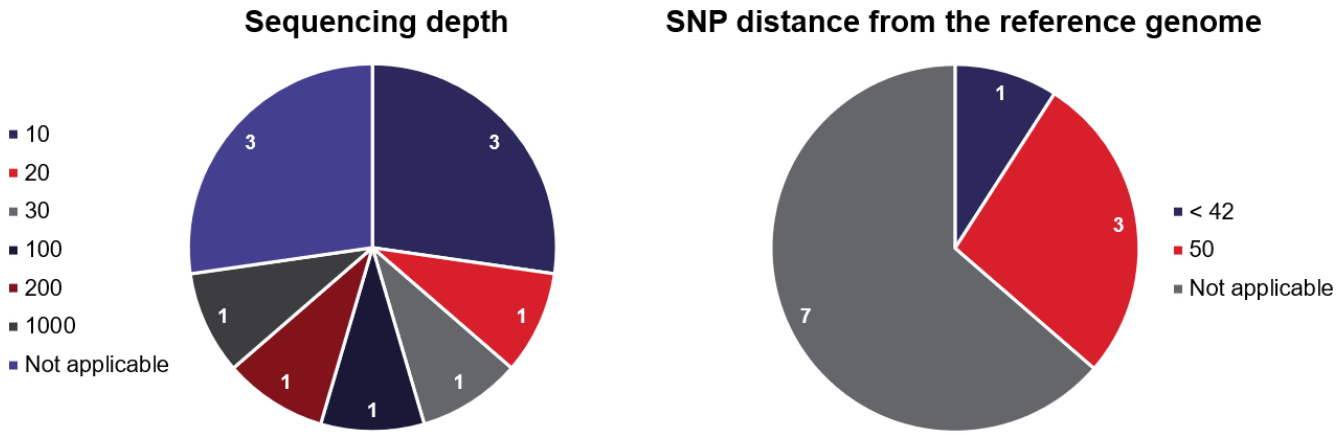
K. Approach to remove human reads

The majority of the participants (10/11) did not remove human reads (not applicable); human reads were excluded by default when the alignment was performed by mapping the sequence to the reference genome of SARS-CoV-2 at the assembly pipeline (amplicon-based approach sequencing). While a step to explicitly remove human reads was not incorporated in the standard pipeline of one of these participants, the participant will use kraken version 2.1.1 to verify the taxonomic group for the potential contaminants if human (or other) contamination is expected in the case when the negative control failed quality check. Similarly, another participant reviewed and analysed the trimmed reads to flag for the presence of human reads, as potential cross-contamination by non-typical organisms may be present if reads were detected in the negative control. Only one participant performed reference mapping to human reference.

L. Quality control (QC)-checks: thresholds

As indicated by participants, the threshold for the ambiguous bases/sites was variable and were not shown here.

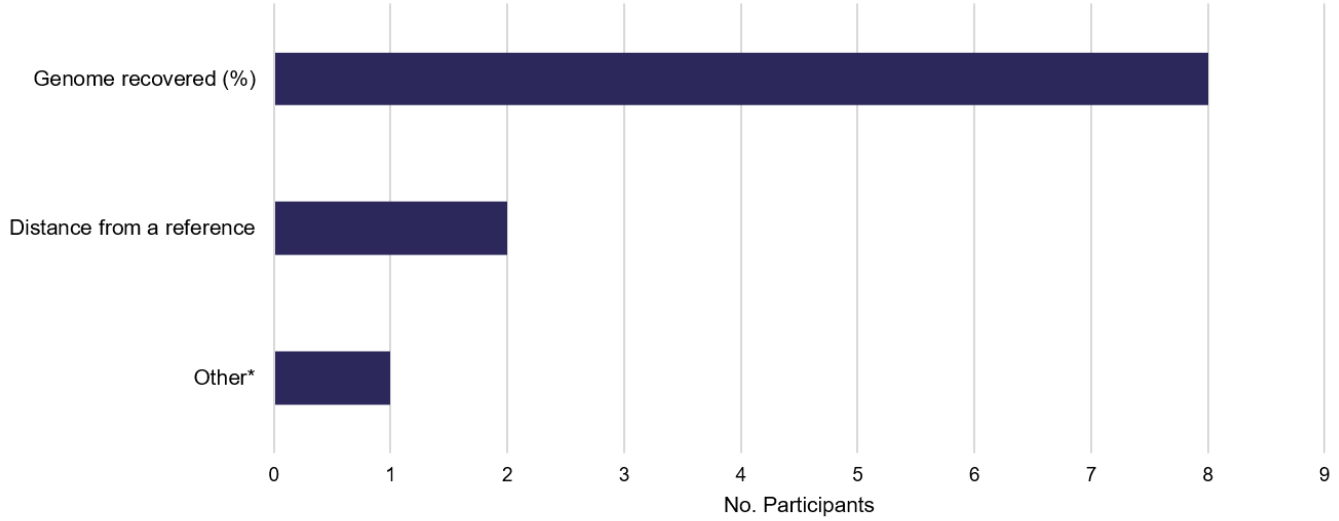




Only one participant indicated that an additional criterion (heterozygous sites) would be considered during the QC-checks process, with a threshold of 40.

Summary of results submitted (Questionnaire Part 2)

A. QC metrics used to assess the appropriateness of a sequence for inclusion into an analysis



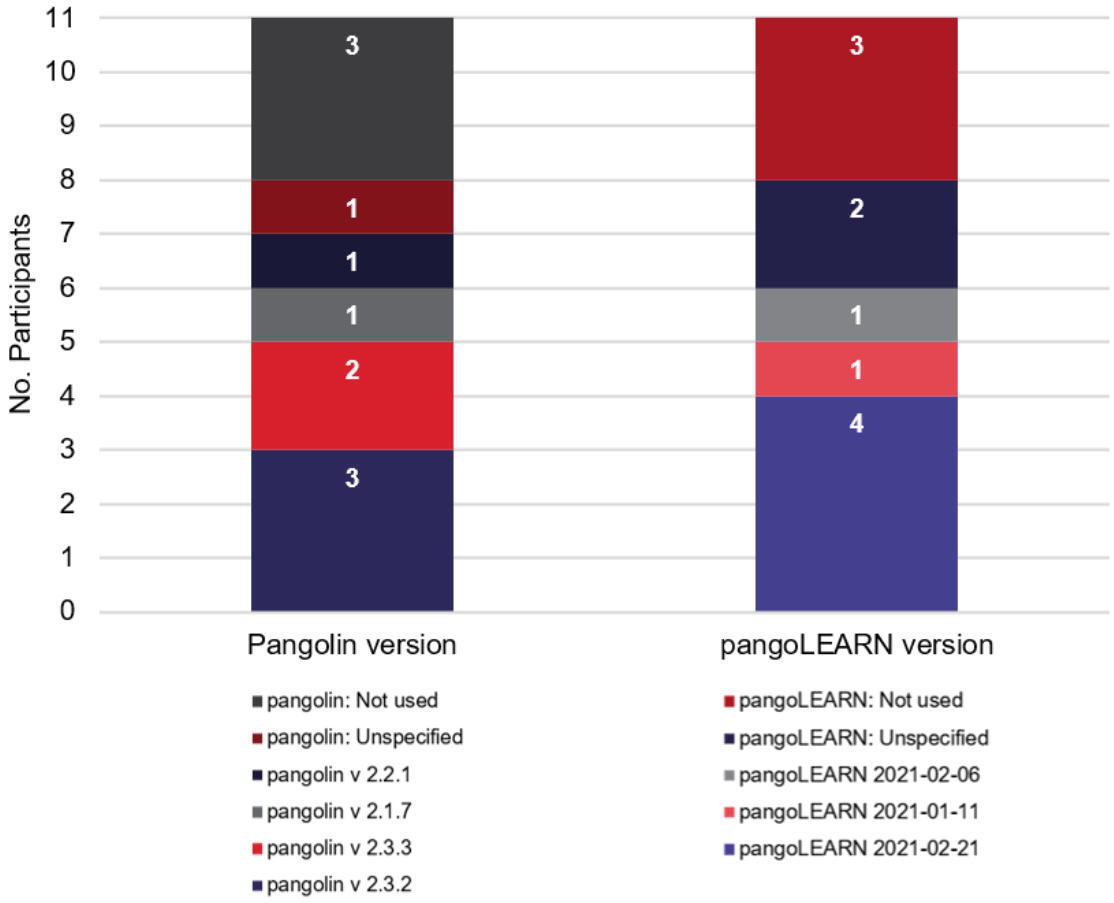
^ Other: All sequences are progressively filtered. Sequences that do not pass several filtering criteria (based on the pipeline adopted by GISAID) are excluded from phylogenetic and clustering analyses. Specifically, sequences are excluded from analyses if they are shorter than 28,000 bp after alignment, and/or contain more than 1000 ambiguities after alignment, and/or are identified as being on a long branch according to 'TreeShrink'.

B. Filtering of the alignment

A total of 3/11 participants did not perform filtering of the alignment, while the remaining participants performed alignment filtering and the software used and/or criteria of filtering are available below.

Participant ID	Software and/or criteria used
1	Manual inspection of alignment, removal of gaps and masking of ambiguous sites
2	3-prime and 5-prime UTRs were filtered (by coordinates from MN908947.3) Problematic sites were masked where they were flagged with 'MASK' in the VCF file here: https://raw.githubusercontent.com/W-L/ProblematicSites_SARS-CoV2/master/problematic_sites_sarsCov2.vcf
4	Alignment and filtering follows the guidelines of https://github.com/roblanf/sarscov2phylo/. The GISAID sample 'hCoV-19/Wuhan/WH04/2020|EPI_ISL_406801' is added to the sequences fasta file prior to alignment. Sequences are progressively profile aligned to the RefSeq SARS-CoV-2 assembly (NC_045512.2) using 'mafft', with file wrangling achieved via 'faSplit', 'faSomeRecords', and UNIX commands ('grep', 'find', 'GNU parallel'). Then, sites in the SARS-CoV-2 genome that are known to be error-prone (sequencing errors, homoplasy issues: http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473) are masked using 'faSplit', 'seqmagick', and UNIX commands (as above). The bash script that carries out the masking of problematic sites also masks the first and last 30 informative base pairs of every sequence, and any 7 bp window that contains at least two gaps or Ns. Any sequences that are shorter than 28,000 bp or have >1000 ambiguities are removed from the alignment using 'esl-alimask'. Prior to cluster estimation, a phylogenetic tree is estimated with 'IQTREE2'. The resulting tree is re-rooted with respect to 'EPI_ISL_406801', then queried with 'TreeShrink' to identify problematic sequences on long branches. If any sequences are recovered as problematic, they are pruned from the alignment and tree. Finally, 'EPI_ISL_406801' is pruned from the alignment and tree. This filtering scheme resulted in 67/70 RCPAQAP samples being retained. Sequences removed were: nCov19/RCPA-PTP-2021-S002/2020, nCov19/RCPA-PTP-2021-S035/2020, nCov19/RCPA-PTP-2021-S056/2020.
5	manual curation
6	Inhouse script
8	Using in-house script to mask problematic sites based on https://github.com/W-L/ProblematicSites_SARS-CoV2
9	Custom in-house script which trims the UTR regions from the consensus sequence.
10	MAFFT 7.221.3

C. Pangolin lineage in interpreting genomic relationships



D. Consensus sequences used to generate a phylogenetic tree

All participants used the consensus sequences to generate a phylogenetic tree; the majority of them used IQ-TREE (6/11) as their software of choice. Details of the tree inference software and its associated version are available below.

Participant ID	Tree inference software
1	iqtree2 v2.0.6
2	Tree topology was calculated using FastTree version 2.1.11. Branch lengths were calculated using RAxML-NG version 1.0.2.
3	IQTREE 2.1.2
4	IQ-TREE multicore version 2.1.2 COVID-edition for Linux 64-bit built October 22 2020
5	Geneious Prime
6	iqtree v2.0.3
7	CLC workbench 20.0.4, Max Likelihood GTR sub
8	IQTREE v2.1.0
9	Iqtree version 1.6.7
10	FASTTree v2.1.10+galaxy1 for approximately-maximum-likelihood phylogenetic trees from alignments
11	PhyML v3.3.20180621

E. Distances calculated between sequences

The majority of participants calculated distances between sequences (9/11) using a variable metric such as SNP distance. Details are summarised below.

Participant ID	Metric (SNP distance) and threshold used in distances calculation
1	likelihood (patristic) distance matrix
2	Both evolutionary & SNP distances were used, but no hard thresholds were applied
4	During analysis with ClusterPicker v1.2.5, the maximum within-cluster genetic distance is calculated using pairwise distances (p-distances) between sequences, taking into account all sites except gaps (-, N, ~). In combination with tree support metrics (see below), a threshold of 0.035% within-cluster genetic distance is used to define clusters. This threshold corresponds to allowing ~10 bases (out of 29784 in the alignment I used) to differ between sequences within a cluster. The threshold might vary for an alignment of a different length.
5	Number of SNPs; no specific threshold
6	snp-dists
8	0-1 SNP
9	snp-dists version 0.6. Samples would be considered to be clustered is the snp-distance between them was ≤ 2
10	Jukes-Cantor distance -0.75*log(1 - 4/3 d), where d is the proportion of positions that differ, and Generalised time-reversible model
11	Phylogenetic clusters that differed by no more than 2 SNPs from the index case or core sequence