Bioinformatic investigation of discordant sequence data for SARS-CoV-2: insights for robust genomic analysis during pandemic surveillance

Authors

Sara E. Zufan[1], Katherine A. Lau[2], Angela Donald[1,3], Tuyet Hoang[1,3], Charles S.P. Foster[4,5], Chisha Sikazwe[6,7], Torsten Theis[2], William D. Rawlinson[2,4,5,8,9], Susan A. Ballard[3], Timothy P. Stinear[1], Benjamin P. Howden[1,3], Amy V. Jennison[10], Torsten Seemann[1,3]

Affiliation(s)

[1]Department of Microbiology and Immunology, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

[2]RCPAQAP Biosecurity, St. Leonards, NSW, Australia

[3]Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia

[4]Serology and Virology Division (SAViD) SEALS Microbiology, NSW Health Pathology, Sydney, NSW, Australia

[5]School of Biomedical Sciences, Faculty of Medicine and Health, University of New South Wales, Sydney, NSW, Australia

[6]Department of Microbiology, PathWest Laboratory Medicine Western Australia, Nedlands, WA, Australia

[7]School of Biomedical Sciences, The University of Western Australia, Nedlands, WA, Australia

[8]School of Women's and Children's Health, Faculty of Medicine and Health, University of New South Wales, Sydney, NSW, Australia

[9]School of Biotechnology and Biomolecular Sciences, Faculty of Science, University of New South Wales, Sydney, NSW, Australia

[10]Public Health Microbiology, Forensic and Scientific Services, Queensland Department of Health, Brisbane, Australia

Corresponding author and email address

Torsten Seemann, t.seemann@unimelb.edu.au

Keywords

SARS-CoV-2, public health genomics, bioinformatics, quality assurance

Abstract

The capacity to undertake whole genome sequencing (WGS) in public health laboratories (PHLs) has grown rapidly in response to COVID-19, and SARS-CoV-2 genomic data has been invaluable for managing the pandemic. The public health response has been further supported by the rapid upgrade and implementation of laboratory and bioinformatic resources. However, there remains a high degree of variability in methods and capabilities between laboratories. In addition to evolving methodology and improved understanding of SARS-CoV-2, public health laboratories have become strained during surges in case numbers, adding to the difficulty of ensuring the highest data accuracy. Here, we formed a national working group comprised of laboratory scientists and bioinformaticians from Australia and New Zealand to improve data concordance across PHLs. Through investigating discordant sequence data from Australia's first external SARS-CoV-2 WGS proficiency testing program (PTP), we show that most discrepancies in genome assessment arose from intrahost variation. While others could be remedied using reasonable, parsimonious bioinformatic quality control. Furthermore, we demonstrate how multidisciplinary national working groups can inform guidelines in real time for bioinformatic quality acceptance criteria. Provision of technical feedback allows laboratory improvement during a pandemic in real time, enhancing public health responses.

Data summary

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

Impact statement

The COVID-19 pandemic has brought the utility of genomics to the forefront of microbial surveillance in public health. Our findings provide recommendations for monitoring bioinformatic quality controls in a pandemic context and how multidisciplinary national working groups can provide technical feedback for actionable improvement towards sequence data concordance.

Introduction

The global scientific community responded to the COVID-19 pandemic with the rapid development of laboratory and bioinformatic methods for whole genome sequencing (WGS) of SARS-CoV-2. These methods have been widely adopted, with more than 45 countries conducting routine genomic surveillance (1). Genomic data have been used to inform the local and national pandemic response. Applications range from monitoring the distribution and emergence of variants of concern (VOC) to supplementing contact tracing efforts (2). These efforts have been supported by an open-access framework for the wide-scale sharing of data and resources, including GISAID, Nextstrain, and protocols.io, among others. By May 2020, problematic sequence data attributed to contamination or sequencing and bioinformatic errors had been detected in the public domain, highlighting the need for quality control and remediation strategies to improve consensus genome accuracy (3).

Amplicon sequencing has been the main method of SARS-CoV-2 WGS, with bioinformatic methods dependent on the platform (4,5). Previous interlaboratory and interprotocol comparative studies of SARS-CoV-2 WGS have found differences in single nucleotide variants (SNV) during the assessment of wet and dry methodologies (6,7). The appropriate selection of bioinformatic workflows and thresholds is key to the accuracy of genomic data and varies depending on the sequencing approach and technology. While selection varies, some commonly accepted thresholds include: read depth (>=10 for Illumina, >=20 for Nanopore), phred score (>25), and genome coverage >=90% (8). Discrepancies at SARS-CoV-2 mutation sites can affect the interpretation of genomic metrics of clinical and public health importance, such as Pango lineage classification, phylogenetic placement, or genomic epidemiological clustering (9–11). As such, bioinformatic quality control processes ensure that methodological variability yields consistent results.

Australia and New Zealand have been at the forefront in the use of fine-scale genomic data to manage their COVID-19 responses. Genomic data are shared via AusTrakka to provide context for local and interstate transmission. AusTrakka is a national real-time pathogen genomics surveillance platform for public health laboratories that facilitates analytical harmonisation and national reporting as well as provides equitable access to computational resources and expertise (12). Due to rapid transmission and Australia's high case sequencing rate, interpretations can be made by as little as one SNV. As is common in many countries, each jurisdiction has a unique laboratory and bioinformatic workflow to suit its facility. In response, a national working group facilitated by the Communicable Diseases Genomics Network (CDGN), consisting of public health laboratorians and bioinformaticians across Australia and New Zealand, formed to develop a framework for laboratory and bioinformatic standards and quality control that allows for procedural flexibility while producing high-quality, consistent SARS-CoV-2 genomes.

Between February and March 2021, 11 Australian laboratories participated in the first external SARS-CoV-2 WGS proficiency testing program (PTP) conducted by the Royal College of Pathologists of Australasia Quality Assurance Programs (RCPAQAP). The results of this PTP revealed varied capacities and difficulty in achieving consistent SARS-CoV-2 consensus genomes in Australia (13). The working group obtained this data with the aim of characterising discordant SNVs and investigate bioinformatic parameters for remediation.

Methods

Data Source: 2021 RCPAQAP SARS-CoV-2 WGS PTP

Data were obtained from the 2021 RCPAQAP SARS-CoV-2 WGS PTP, sample characteristics and methods have previously been described (13). Briefly, viral RNA extracted from five unique lineages of SARS-CoV-2 isolates were each diluted in 0.5 ml of nuclease-free water to assess the participant's ability to sequence different variants and viral titers. Contrived samples were sent to 11 participating laboratories for WGS using the laboratory's standard procedure. The raw reads of sequence data without pre-processing or trimming and the derived consensus sequence data, as submitted by the participants, were assessed based on three quality metrics, as previously described by Lau et al (2022). The participants also completed a questionnaire documenting laboratory and bioinformatic methods performed and self-reported genomic characteristics. Responses to the questionnaire were excluded from the PTP assessment. A summary of the comparison of discordant results was provided to participants in the PTP report for self-assessment purposes only. We obtained consensus sequences from participants in FASTA format, primary sequencing data in FASTQ format, and the method questionnaire for retrospective analysis. One participant belonging to the working group inadvertently omitted consensus genomes from submission to the PTP and provided them independently for this study. Here, we refer to the participants by their laboratory number (LB01 – LB11) and the samples by their biological sample number (BS01 – BS08) (Table 1).

Investigation of submitted consensus genomes

The consensus genomes submitted in FASTA format were used to obtain genome coverage and PANGO lineages. Genome coverage was calculated as the percentage of non-N alleles in a consensus genome relative to the SARS-CoV-2 reference genome (Wuhan-Hu-1; GenBank MN908947.3). Lineages were assigned to the samples using Pangolin v.2.3.3 with pangoLEARN 2021-21-02, versions available at the time of the PTP. Discordant, missed, and ambiguous SNVs relative to expected SNVs defined by the PTP were identified using Nextclade v.2.8.0 relative to expected SNVs defined by the PTP (Figure 1). Consensus genome positions where observed alleles matched expected alleles were characterised as 'concordant', 'discordant' if the consensus allele did not match the expected allele, 'missing' if there was an N or gap at an expected SNV position of the SNV, or 'ambiguous' if any position contained an IUPAC ambiguity code. A phylogenetic tree was constructed using RAxML-NG v.1.1 with a GTR+G4 model, starting trees from FastTree v.2.1.11 with 1,000 bootstraps, and alignment column weight using gotree goalign compress v.0.4.2 (14). BS03 and BS08 from LB10 were omitted from the final input alignment due to disproportionately long branches that negatively affected tree readability.

Investigation of submitted reads

The submitted FASTQ reads were mapped to the SARS-CoV-2 reference using minimap2 v. 2.24-r1122 with options -ax map-ont for Nanopore or -ax sr for Illumina reads. The corresponding amplicon primers (Supplementary Table 1) were trimmed from the mapped reads using Samtools ampliconclip v.1.16.1 (options -- both ends -- strand -- soft-clip -u). Basepair characteristics of total mapped reads were calculated using bam-readcount v.1.0.1 with default options, and parsed with the parse_brc.py script (15). Trimmed reads were processed similarly with option -b 20 to discard low-quality reads. Per amplicon read depth was calculated with mosdepth v.0.3.3 using the corresponding amplicon insert regions as the input bam file.

171 # Results
172
173 ## Methodology Characteristics

174 Participant workflows varied for laboratory and bioinformatic methodologies (Supplementary
175 Tables 1-2). Most participants (N=9/11) performed multiplex amplicon sequencing, with
176 seven using the ARTIC V3 primer scheme (16,17), another using a modified ARTIC V3
177 primer scheme to create 1 kbp amplicons, and one using the 'Midnight' scheme (4,18). Two
178 participants used a long-pooled amplicon approach, 'JSE', with primers that produce 2500bp-
179 length amplicons (19,20). Most of the participants used an Illumina sequencing platform
180 (N=8/11), with 6/8 using ARTIC V3 primers. All participants using Nanopore technology use
181 the artic-ncov2019 pipeline (https://github.com/artic-network/artic-ncov2019). Amongst
182 Illumina users, four used an in-house analysis pipeline, two used a mix of command line and
183 commercial software, and two used solely commercial software.
184
185 ## Characteristics of Consensus Genomes and Reads Submitted

186 Genome coverage for consensus sequences submitted of SARS-CoV-2 positive samples was
187 generally >98%, though mean coverage of BS08 was 93.75% due to a high cycle threshold
188 value (Ct=27.48) (6). LB05 and LB09 had consistent coverage <99%, suggesting dropout of
189 the amplicon(s). Upon further inspection, LB05 had two or more amplicon dropouts across
190 all samples, generally in amplicons 64 and 70. Similarly, LB09 failed to amplify amplicons 1
191 and 98 in all samples. While most participants detected few to no SARS-CoV-2 reads in the
192 negative control BS06, LB03 and LB05 recovered 11.0% and 4.5% of the genome,
193 respectively, and LB10 recovered 79.2% (Supplementary Figure 1).
194
195 ## Investigation of Discordant SNVs

196 Four out of 11 participants had 100% concordant genomes when excluding ambiguous results
197 (Figure 1). However, when inspecting read pileups at discordant positions associated with
198 ambiguity codes, we found that all participants had sequenced isolates with intrahost
199 variation (Figure 2a). Therefore, the allele called at the position depended upon a) the allele
200 mixture sequenced; b) the variant calling threshold; and c) whether ambiguity codes were
201 used in variant calling. Taking this into account, six participants had 100% concordant
202 genomes, with an additional two having appropriate ambiguity codes.
203
204 LB01 and LB03 yielded discordant or missed SNVs in BS08, the high-Ct sample. At
205 nucleotide (nt) positions 25,588 and 5,574 in LB01 and LB03, respectively, we observe an
206 allele mixture where the discordant allele is the major allele according to the variant calling
207 thresholds of the participant. We investigated the allele frequency at identical sites in the
208 remaining samples and found them homogeneous. Finally, at nt position 5,455 in the sample
209 of LB01, the discordant allele was present at a frequency of 1 (Figure 2b). Given that this
210 false SNV was present in 100% of the reads and was not detected in the other matched
211 specimen, it may be the product of chimeric PCR amplification (7).
212
213 LB10 had the highest number of discordant SNVs, with 51 and 39 in BS03 and BS08,
214 respectively. Given the high recovery of the SARS-CoV-2 genome in the negative control of
215 LB10 and the high number of total reads, laboratory contamination was suspected. However,
216 an inspection of the read pileup found a disproportionately high read depth among several
217 regions and found that the concordant allele was the major allele at discordant positions in
218 many instances. In BS03, there was overamplification of amplicons 2, 3 and 4 with a mean
219 read depth of 43866, 302975 and 47687, respectively. The mean coverage across all other

amplicons was 33.6. In BS08, the mean amplicon coverage was 18.2. However, the read depth in the 3'UTR was 5910.7. Inspection of the read pileup in this region revealed an abundance of poly-A reads. All LB10 samples had similar read mapping characteristics in the 3' UTR, possibly suggesting poly-A carryover from cDNA synthesis. When investigating the distribution of alleles at discordant positions, in many cases, the discordant allele was present at a depth of 1 or 0, with the concordant allele present at a greater depth or frequency (Supplementary Figure 2). In fewer cases, the discordant allele was present in greater frequency than the concordant allele, which may be attributed to contamination (Supplementary Figure 2). We could not determine this participant's specific bioinformatic methods or parameters from their questionnaire responses to further investigate how variants were being called (Supplementary Table 2). However, we can infer that the minimum depth threshold for variant calling was below 10, a commonly accepted minimum criterion for Illumina amplicon sequence data (19).

Impact on Key Surveillance Metrics

PANGO lineage assignments were largely tolerant of missing data and false SNVs, with 95% of consensus genomes submitted correctly designated at the highest taxonomic level (Figure 3a). Because PANGOLIN classification is hierarchical, genomes with missing data can be expected to have parent lineage assignments, such as BS03 from LB10 (21). BS08 samples from LB01 and LB10 were probably designated as a higher, incorrect lineage due to a combination of low genome coverage and false SNVs.

In general, the consensus genomes submitted were placed in their expected phylogenetic clades (Figure 3b). BS03 and BS08 from LB10 were omitted from the input alignment as the relatively high number of false mutations resulted in long branches, making it difficult to interpret the tree. Among the remaining 75 genomes, 84% were placed within their expected clade. Subclades formed with low bootstrap support in BS04 and BS05 which included specimens with ambiguity codes, those that had a majority allele different from the isolate reference, and one specimen that was concordant with the basal clade. BS08 samples from LB01 and LB03 formed longer branches within the expected clade, with the former having the longest branch due to lower genome coverage (63.0% vs. 96.7%) and more discordant SNVs (N=4 vs. N=2).

Discussion

The assessment of PTP data for all SARS-CoV-2 positive samples was based on three quality metrics (*viz.* genome recovered, PANGO lineage and sequencing accuracy of consensus sequence submitted) decided with input from CDGN Genomics Implementation and Bioinformatics Working Groups, as reported by Lau et al. (2022). Here, we report findings from a further retrospective analysis performed to characterise discordant SNVs. This was performed using data submitted by participants for the PTP, including their consensus genomes, raw reads and the method questionnaire. Overall, the participants data largely produced concordant genomes despite using a variety of laboratory and bioinformatic methods for SARS-CoV-2 WGS. Most discrepancies could be resolved with reasonable bioinformatic quality thresholds. These differences had little to no impact on key surveillance metrics, lineage assignment, and phylogenetic clustering.

Similar to Foster et al. (2022), we primarily observed interlaboratory differences between matched samples when iSNVs were present (20,21). As in the assessments performed in

Foster et al. (2022), where it was concluded IUPAC ambiguity codes were beneficial for quality control, we conclude these may signal laboratory contamination. They may also indicate epidemiological importance where true iSNVs arise from host-to-host transmission or co-infection (22,23). Based on our observations, no universal minimum frequency threshold would have prevented the calling of a false SNV. Instead, a good practice could be to call two consensus genomes using: 1) a strict minimum variant frequency threshold with ambiguity codes for the primary purpose of quality monitoring (i.e. a consensus sequence control); and 2) a major allele threshold without ambiguity codes for downstream analysis.

Otherwise, we find that consensus genomes generated with bioinformatic parameters at or above the commonly accepted minimum thresholds were highly accurate. It should be noted that two participants (LB03 and LB10) generated consensus genomes using read depth <10, the commonly accepted minimum threshold, with mixed results. LB03 used a depth of 3, resulting in 100% concordant genomes when including ambiguity. While LB10 used a depth of 1, resulting in 6/8 consensus genomes with 100% concordance and two (BS03, BS08) with >40 discordant SNVs. One concern regarding variant calling with low read depth is false SNVs due to contamination, as observed here for a small number of false SNVs in LB10-BS03 and LB10-BS08. This can be ameliorated by monitoring reads in a negative sequencing control or a consensus sequence control to determine if a read set should use a higher depth threshold or fail quality control. The latter approach is more bioinformatically challenging to extract and interpret allele frequencies.

In characterising discordant SNVs, we found that technical analysis of this dataset could provide actionable feedback for participants. For PHLs already engaged in WGS quality control, they have typically had more experience with methods designed for prokaryotic pathogens in which shotgun sequencing of isolates is performed (24–26). The characteristics of amplicon sequencing differ in several ways in which technical bioinformatic analysis can support laboratory optimisation. First, PCR efficiency often results in non-uniform coverage across the SARS-CoV-2 genome. Although genome coverage is an important quality indicator for unbiased sequencing approaches, it can be misleading when a small number of amplicons are overamplified or poly-A is carried through the library preparation and sequenced, as observed here with LB10. Instead, monitoring the mean read depth per amplicon would help participants troubleshoot amplicon drop outs, a common issue with this method that requires optimisation (27–29). Second, the proximity of discordant SNVs to primers can indicate bioinformatic trimming errors. It has previously been observed that CLC Genomics Workbench (Qiagen, https://www.qiagenbioinformatics.com/products/clc-genomics-workbench, accessed 03 Jan 2023) has a lower sensitivity to trimming partial primers and calling SNVs in primer regions, compared to iVar (7). Although we were unable to replicate the results of LB10 using CLC 22 (Identify ARTIC V3 SARS-CoV-2 Low-frequency and Shared Variants (Illumina) Workflow) we did observe that false SNVs occurred near primer sites, which could signal the need for bioinformatic troubleshooting by the participant.

While the working group found this analysis informative for improving sequence data concordance across laboratories, there were limitations given the retrospective design of this investigation. For instance, in attempting to replicate methods to troubleshoot SNVs, we found data entry errors, including typos and null responses, leading to discrepancies in the PTP questionnaire responses. Although the questionnaire was designed to capture the participants' capacity to correctly report, interpret, and communicate these results to relevant stakeholders, our analysis provides a different perspective on the capacity of PHLs, especially

the difficulties they may have faced amid a pandemic surge. It may also reflect the varied capacity of PHLs to identify key bioinformatic characteristics. In the future, multidisciplinary national working groups could develop an internal, easy-to-use bioinformatic analysis dashboard for participants to upload reads to receive technical feedback, like the metrics described here. These insights would provide the working group with data-driven strategies to inform living guidelines to ensure sequence data concordance as knowledge and methods evolve.

Since the PTP analysed here concluded in 2021, several laboratory and bioinformatic developments have occurred, reflecting the evolving nature of the emergent pandemic response. The ARTIC primer scheme was redesigned in terms of laboratory methods and commercial schemes were offered to account for increasing variation (30,31). From a bioinformatic perspective, AusTrakka increased its minimum genome coverage criteria from >50% to >=90% (ACGT bases), which was the basis for the internal quality criterium of several participants. Regardless of AusTrakka participation, we agree with the conclusions of Lau et al. (2022) and recommend this criterion as a minimum given the shared characteristics of the discordant metrics observed here (LB01-BS08, LB10-BS03, and LB10-BS08) being genome coverage <80% in addition to previous benchmarking studies (8,21). These developments further demonstrate the need for living guidelines to continually review and update the quality standards for SARS-CoV-2 WGS.

Conclusions

In summary, despite various laboratory and bioinformatic approaches for SARS-CoV-2 WGS, participants generally produced concordant results. We find that consensus genomes generated with parameters at or above the commonly accepted minimum thresholds were highly accurate. We suggest caution is needed when calling variants at positions with minimum read depths of below 10. We recommend PHLs generate two consensus genomes for: 1) quality control using ambiguity codes and a strict minimum variant frequency threshold with ambiguity codes; and 2) downstream analysis using a major allele threshold without ambiguity codes. Given the results presented here, we recommend these additional PTP assessment criteria, if not already implemented: 1) per-amplicon read depth from sequence data submitted as a measure of primer efficiency; 2) SNVs derived from sequence data to account for intrahost variation within the received specimen; and 3) minimum genome coverage of 90%.

In a genomics-forward era of public health pathogen surveillance, we show how multidisciplinary national working groups can employ technical analysis to provide feedback for sequence data concordance across PHLs. We also identified additional assessment criteria for PTPs to further assist laboratories in identifying areas for improvement.

Author statements

Author contributions

Contribution	Authors
Conceptualization	S.Z., A.J., T.S.
Supervision	T.P.S., T.S., B.H.
Data Curation	K.L., T.T.
Data Analysis	S.Z.
Writing – original draft	S.Z.
Writing – review and editing	All authors contributed to review and/or editing of the manuscript
Project management	A.D., T.H.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

This study was funded by the National Health and Medical Research Council (NHMRC) through the Medical Research Future Fund (MRFF) – Coronavirus Research Response: 2020 Tracking COVID-19 in Australia using Genomics Grant Opportunity (MRF9200006).

Acknowledgements

We acknowledge members of the MRFF COVID-19 Quality Control Working Group for their feedback and contribution to this paper. Members include: Jaisy Arikkatt, Susan Ballard, Angela Donald, Charles Foster, Rikki Graham, Tuyet Hoang, Amy Jennison, Jessica Johnson-Mackinnon, Jurissa Lang, Terrence Lee, Lex Leong, Jaimie McMahon, Paran Rayan Samuel, Torsten Seemann, Chisha Sikazwe, and Chenwei Wang. We also owe appreciation to the Royal College of Pathologists of Australasia Quality Assurance Programs (RCPAQAP), and The Communicable Diseases Genomics Network, supported by the Office of Health Protection and Response, within the Australian Government Department of Health and Aged Care, which has led the national coordination and integration of SARS-CoV-2 genomic data in Australia.

382 ## References
383
384 1. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of SARS-
385 CoV-2 genomic surveillance and data sharing. Nat Genet [Internet]. 2022
386 Apr;54(4):499–507. Available from: http://dx.doi.org/10.1038/s41588-022-01033-y

387 2. Lane CR, Sherry NL, Porter AF, Duchene S, Horan K, Andersson P, et al. Genomics-
388 informed responses in the elimination of COVID-19 in Victoria, Australia: an
389 observational, genomic epidemiological study. Lancet Public Health [Internet]. 2021
390 Aug;6(8):e547–56. Available from: http://dx.doi.org/10.1016/S2468-2667(21)00133-X

391 3. Issues with SARS-CoV-2 sequencing data [Internet]. Virological. 2020 [cited 2022 Nov
392 15]. Available from: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473

393 4. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome
394 sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid
395 Barcoding. Biol Methods Protoc [Internet]. 2020 Jul 18;5(1):bpaa014. Available from:
396 http://dx.doi.org/10.1093/biomethods/bpaa014

397 5. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al.
398 Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus
399 genomes directly from clinical samples. Nat Protoc [Internet]. 2017 Jun;12(6):1261–76.
400 Available from: http://dx.doi.org/10.1038/nprot.2017.066

401 6. Foster CSP, Stelzer-Braid S, Deveson IW, Bull RA, Yeang M, Au J-P, et al. Assessment
402 of Inter-Laboratory Differences in SARS-CoV-2 Consensus Genome Assemblies
403 between Public Health Laboratories in Australia. Viruses [Internet]. 2022 Jan 19;14(2).
404 Available from: http://dx.doi.org/10.3390/v14020185

405 7. Liu T, Chen Z, Chen W, Chen X, Hosseini M, Yang Z, et al. A benchmarking study of
406 SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples.
407 iScience [Internet]. 2021 Aug 20;24(8):102892. Available from:
408 http://dx.doi.org/10.1016/j.isci.2021.102892

409 8. Xiaoli L, Hagey JV, Park DJ, Gulvik CA, Young EL, Alikhan N-F, et al. Benchmark
410 datasets for SARS-CoV-2 surveillance bioinformatics. PeerJ [Internet]. 2022 Sep
411 5;10:e13821. Available from: http://dx.doi.org/10.7717/peerj.13821

412 9. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic
413 nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat
414 Microbiol [Internet]. 2020 Nov;5(11):1403–7. Available from:
415 http://dx.doi.org/10.1038/s41564-020-0770-5

416 10. Frampton D, Rampling T, Cross A, Bailey H, Heaney J, Byott M, et al. Genomic
417 characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in
418 London, UK: a whole-genome sequencing and hospital-based cohort study. Lancet
419 Infect Dis [Internet]. 2021 Sep;21(9):1246–56. Available from:
420 http://dx.doi.org/10.1016/S1473-3099(21)00170-5

421 11. Geoghegan JL, Ren X, Storey M, Hadfield J, Jelley L, Jefferies S, et al. Genomic
422 epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa

423 New Zealand. Nat Commun [Internet]. 2020 Dec 11;11(1):6351. Available from:
424 http://dx.doi.org/10.1038/s41467-020-20235-8

425 12. Hoang T, da Silva AG, Jennison AV, Williamson DA, Howden BP, Seemann T.
426 AusTrakka: Fast-tracking nationalized genomics surveillance in response to the COVID-
427 19 pandemic. Nat Commun [Internet]. 2022 Feb 14;13(1):865. Available from:
428 http://dx.doi.org/10.1038/s41467-022-28529-9

429 13. Lau KA, Horan K, Gonçalves da Silva A, Kaufer A, Theis T, Ballard SA, et al.
430 Proficiency testing for SARS-CoV-2 whole genome sequencing. Pathology [Internet].
431 2022 Aug;54(5):615–22. Available from: http://dx.doi.org/10.1016/j.pathol.2022.04.002

432 14. Lemoine F, Gascuel O. Gotree/Goalign: toolkit and Go API to facilitate the development
433 of phylogenetic workflows. NAR Genom Bioinform [Internet]. 2021 Sep;3(3):lqab075.
434 Available from: http://dx.doi.org/10.1093/nargab/lqab075

435 15. Khanna A, Larson DE, Srivatsan SN, Mosior M, Abbott TE, Kiwala S, et al. Bam-
436 readcount -- rapid generation of basepair-resolution sequence metrics. ArXiv [Internet].
437 2021 Jul 27; Available from: https://www.ncbi.nlm.nih.gov/pubmed/34341766

438 16. Quick J. nCoV-2019 sequencing protocol v3 (LoCost). protocols io. 08 2020;

439 17. Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, et al. Improvements
440 to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using
441 nanopore. bioRxiv [Internet]. 2020 Sep 4; Available from:
442 http://dx.doi.org/10.1101/2020.09.04.283077

443 18. Freed N, Silander O. SARS-CoV2 genome sequencing protocol (1200bp amplicon
444 "midnight" primer set, using Nanopore Rapid kit). protocols io. 07 2021;

445 19. Eden J-S, Rockett R, Carter I, Rahman H, de Ligt J, Hadfield J, et al. An emergent clade
446 of SARS-CoV-2 linked to returned travellers from Iran. Virus Evol [Internet]. 2020
447 Jan;6(1):veaa027. Available from: http://dx.doi.org/10.1093/ve/veaa027

448 20. Eden J-S, Sim E. SARS-CoV-2 Genome Sequencing Using Long Pooled Amplicons on
449 Illumina Platforms. protocols io. 04 2020;

450 21. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment
451 of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus
452 Evol [Internet]. 2021 Jul 30;7(2):veab064. Available from:
453 http://dx.doi.org/10.1093/ve/veab064

454 22. Armero A, Berthet N, Avarre J-C. Intra-Host Diversity of SARS-Cov-2 Should Not Be
455 Neglected: Case of the State of Victoria, Australia. Viruses [Internet]. 2021 Jan
456 19;13(1). Available from: http://dx.doi.org/10.3390/v13010133

457 23. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al.
458 SARS-CoV-2 within-host diversity and transmission. Science [Internet]. 2021 Apr
459 16;372(6539). Available from: http://dx.doi.org/10.1126/science.abg0821

460 24. Lau KA, Gonçalves da Silva A, Theis T, Gray J, Ballard SA, Rawlinson WD.
461 Proficiency testing for bacterial whole genome sequencing in assuring the quality of

462 microbiology diagnostics in clinical and public health laboratories. Pathology [Internet].
463 2021 Dec;53(7):902–11. Available from: http://dx.doi.org/10.1016/j.pathol.2021.03.012

464 25. Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E, et al.
465 Proficiency testing for bacterial whole genome sequencing: an end-user survey of
466 current capabilities, requirements and priorities. BMC Infect Dis [Internet]. 2015 Apr
467 3;15:174. Available from: http://dx.doi.org/10.1186/s12879-015-0902-3

468 26. Gargis AS, Kalman L, Lubin IM. Assuring the Quality of Next-Generation Sequencing
469 in Clinical Microbiology and Public Health Laboratories. J Clin Microbiol [Internet].
470 2016 Dec;54(12):2857–65. Available from: http://dx.doi.org/10.1128/JCM.00949-16

471 27. Borcard L, Gempeler S, Terrazos Miani MA, Baumann C, Grädel C, Dijkman R, et al.
472 Investigating the Extent of Primer Dropout in SARS-CoV-2 Genome Sequences During
473 the Early Circulation of Delta Variants. Frontiers in Virology [Internet]. 2022;2.
474 Available from: https://www.frontiersin.org/articles/10.3389/fviro.2022.840952

475 28. Lambisia AW, Mohammed KS, Makori TO, Ndwiga L, Mburu MW, Morobe JM, et al.
476 Optimization of the SARS-CoV-2 ARTIC Network V4 Primers and Whole Genome
477 Sequencing Protocol. Front Med [Internet]. 2022 Feb 17;9:836728. Available from:
478 http://dx.doi.org/10.3389/fmed.2022.836728

479 29. Cotten M, Lule Bugembe D, Kaleebu P, T Phan M V. Alternate primers for whole-
480 genome SARS-CoV-2 sequencing. Virus Evol [Internet]. 2021 Jan;7(1):veab006.
481 Available from: http://dx.doi.org/10.1093/ve/veab006

482 30. Davis JJ, Long SW, Christensen PA, Olsen RJ, Olson R, Shukla M, et al. Analysis of the
483 ARTIC Version 3 and Version 4 SARS-CoV-2 Primers and Their Impact on the
484 Detection of the G142D Amino Acid Substitution in the Spike Protein. Microbiol Spectr
485 [Internet]. 2021 Dec 22;9(3):e0180321. Available from:
486 http://dx.doi.org/10.1128/Spectrum.01803-21

487 31. Bei Y, Pinet K, Vrtis KB, Borgaro JG, Sun L, Campbell M, et al. Overcoming variant
488 mutation-related impacts on viral sequencing and detection methodologies. Front Med
489 [Internet]. 2022 Oct 28;9:989913. Available from:
490 http://dx.doi.org/10.3389/fmed.2022.989913

491 # Figures and tables

492

493 Table 1. Characteristics of contrived study samples.

Sample ID	Isolate (GISAID ID)	Average Ct (E-gene NAT)	Lineage
BS01	EPI_ISL_406844	16.77	B
BS02	EPI_ISL_419750	16.95	B.1
BS07		24.14	
BS08		27.48	
BS03	EPI_ISL_480701	17.28	B.1.1.136
BS04	EPI_ISL_519314	17.31	D.2
BS05	EPI_ISL_563416	17.29	D.2
BS06	Negative Control	Not detected	NA

494

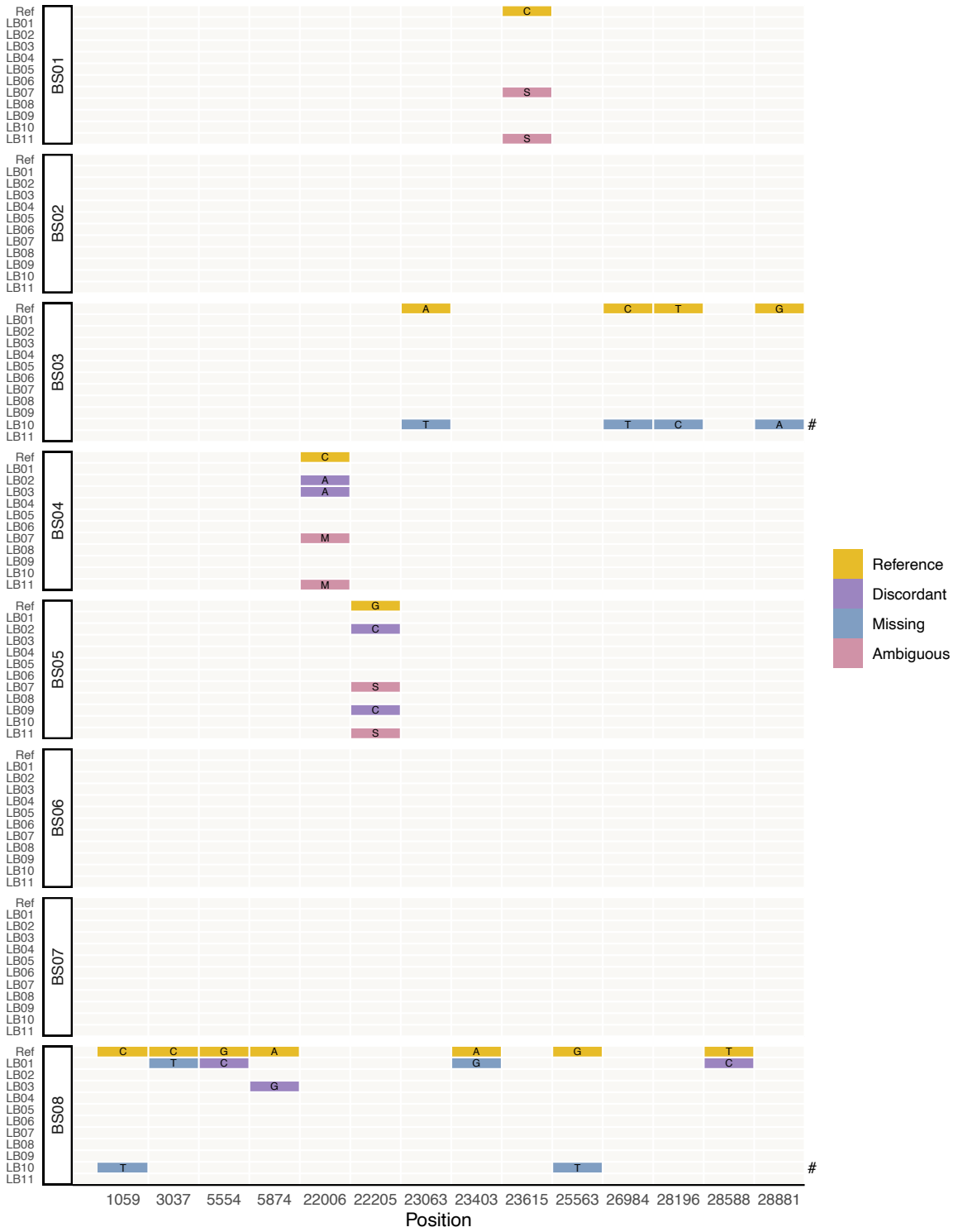


Figure 1.

Discordant SNVs observed in participant consensus genomes submitted. Reference indicates the expected allele of the sample isolate reference sequence. Discordant SNVs are those where the observed participant allele differs from the expected reference allele. #LB10 had additional discordant SNVs for samples BS03 (N=51) and BS08 (N=39) that are not listed here. Missed SNVs are a subset of discordant SNVs in which a participant called an allele different from the expected SNV. Ambiguous SNVs are another subset of discordant SNVs in which the participant called an IUPAC ambiguity code where a standard base (ATGC) was expected (S = C or G; M = A or C). According to the criteria used in the PTP, the absence of

505 C241T was considered concordant, and thus omitted from this figure if present. No
506 annotation implies concordance.

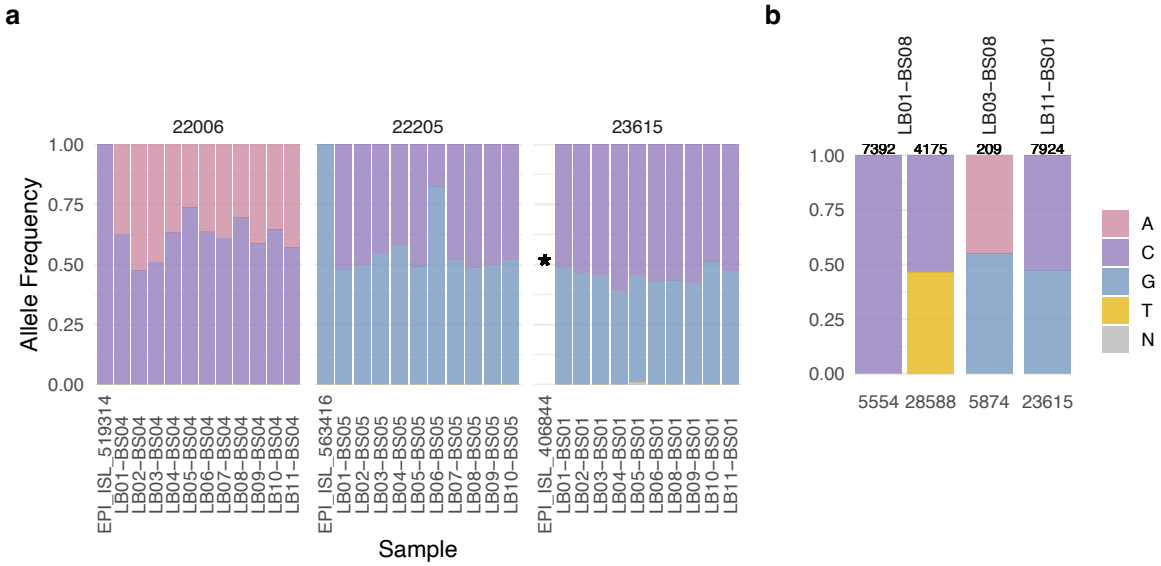
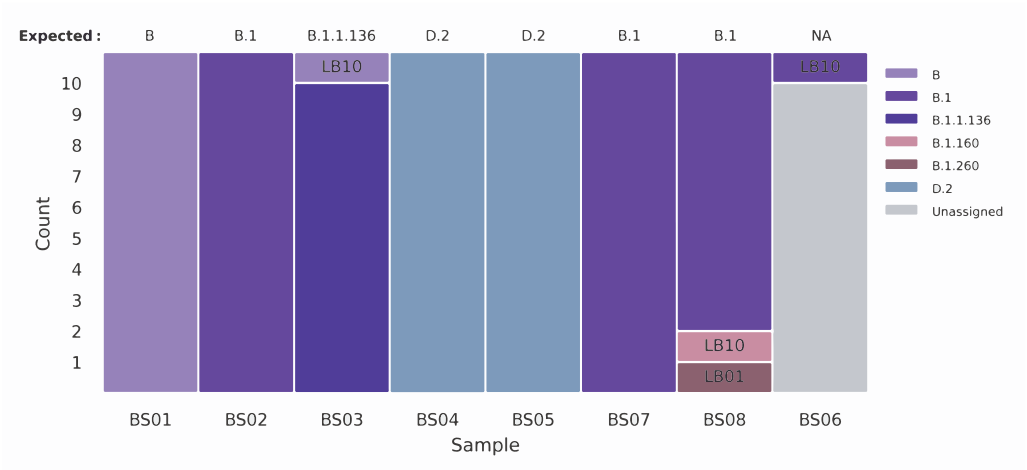
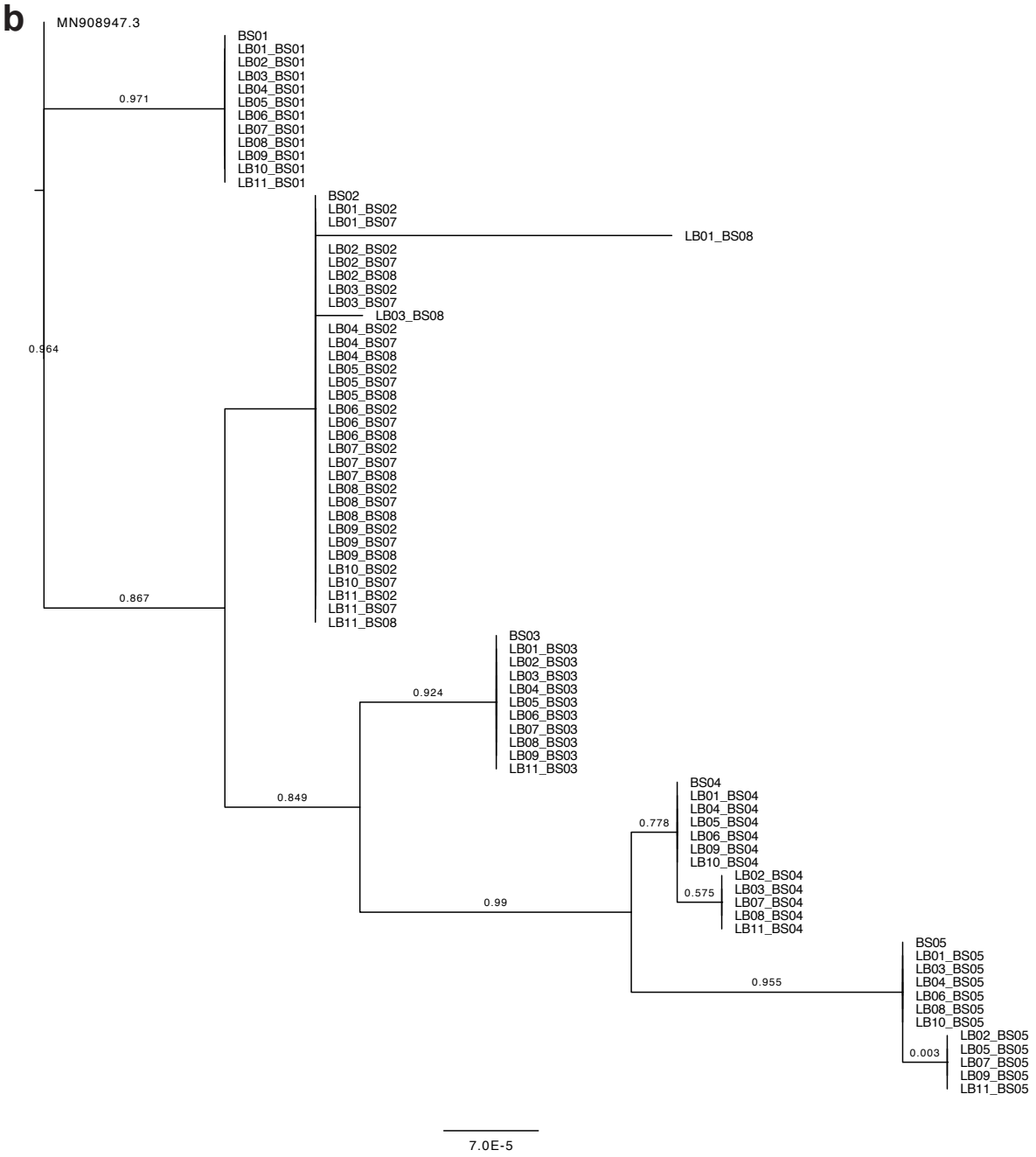


Figure 2.
Allele distribution at discordant positions. a) Allele frequencies at positions containing ambiguity codes where the further left sample is the primary sample from which the isolate was derived. * Read data for EPI_ISL_406844 was not available for analysis. b) Allele frequencies of false SNVs by sample and position. Total read depth is annotated at the top of each bar.

a


Expected :
B
B.1
B.1.1.136
D.2
D.2
B.1
B.1
NA
LB10
LB10
LB10
LB01
Count
10
9
8
7
6
5
4
3
2
1
BS01
BS02
BS03
BS04
BS05
BS07
BS08
BS06
Sample
B
B.1
B.1.1.136
B.1.160
B.1.260
D.2
Unassigned

b


MN908947.3
0.971
BS01
LB01_BS01
LB02_BS01
LB03_BS01
LB04_BS01
LB05_BS01
LB06_BS01
LB07_BS01
LB08_BS01
LB09_BS01
LB10_BS01
LB11_BS01
BS02
LB01_BS02
LB01_BS07
LB01_BS08
LB02_BS02
LB02_BS07
LB02_BS08
LB03_BS02
LB03_BS07
LB03_BS08
LB04_BS02
LB04_BS07
LB04_BS08
LB05_BS02
LB05_BS07
LB05_BS08
LB06_BS02
LB06_BS07
LB06_BS08
LB07_BS02
LB07_BS07
LB07_BS08
LB08_BS02
LB08_BS07
LB08_BS08
LB09_BS02
LB09_BS07
LB09_BS08
LB10_BS02
LB10_BS07
LB11_BS02
LB11_BS07
LB11_BS08
0.964
0.867
BS03
LB01_BS03
LB02_BS03
LB03_BS03
LB04_BS03
LB05_BS03
LB06_BS03
LB07_BS03
LB08_BS03
LB09_BS03
LB11_BS03
0.924
0.849
BS04
LB01_BS04
LB04_BS04
LB05_BS04
LB06_BS04
LB09_BS04
LB10_BS04
0.778
LB02_BS04
LB03_BS04
LB07_BS04
LB08_BS04
LB11_BS04
0.575
0.99
BS05
LB01_BS05
LB03_BS05
LB04_BS05
LB06_BS05
LB08_BS05
LB10_BS05
0.955
LB02_BS05
LB05_BS05
LB07_BS05
LB09_BS05
LB11_BS05
0.003
7.0E-5

514

515 **Figure 3.**
516 Key surveillance metrics derived from consensus genomes submitted. A) Pangolin lineage
517 assignments from consensus genomes submitted. Participant ID annotated where discordant.
518 b) Maximum likelihood tree of consensus genomes submitted, excludes genomes <50%
519 coverage; and BS03 and BS08 for LB10. Branch labels indicate bootstrap support (N=1000).
520 Phylogenetic tree visualised with FigTree v1.4.4 (https://github.com/rambaut/figtree/).