IN THE MATTER OF THE *INQUIRIES ACT 2014* **NTP-027**

AND

IN THE MATTER OF A BOARD OF INQUIRY INTO
THE COVID-19 HOTEL QUARANTINE PROGRAM

WITNESS STATEMENT OF PROFESSOR BEN HOWDEN

I, PROFESSOR BENJAMIN PETER HOWDEN, medical microbiologist, of the University of Melbourne, 792 Elizabeth Street, Melbourne say as follows:

A. Preliminary questions

A.1 Please describe your professional background and qualifications

1. I am the Director of the Microbiological Diagnostic Unit Public Health Laboratory (**MDU PHL**) at the University of Melbourne. I have been the Director since 2014.

2. I obtained a Bachelor of Medicine, Bachelor of Surgery from Monash University in 1993. I am a Fellow of the Royal Australasian College of Physicians (Infectious Diseases, 2001) and a Fellow of the Royal College of Pathologists of Australasia (Microbiology, 2004). I was awarded a PhD in Molecular Biology from Monash University in 2009.

3. Before my appointment to MDU PHL, I was Head of Microbiology and an Infectious Diseases Physician at Austin Health.

4. A copy of my CV is attached.

A.2 Please describe your expertise in genomic sequencing

5. I started research into genomics in 2005. I was one of the first people in Australia to sequence bacterial genomes and have conducted about 15 years of research into bacterial genomes. Since 2014, I have implemented genomic sequencing as a public health diagnostic tool at MDU PHL.

6. I have co-authored 104 peer reviewed manuscripts that include aspects of pathogen genomics. I am accredited by the Royal College of Pathologists of Australasia and the National Pathology Accreditation Advisory Council as a microbiologist for supervision of genetic/genomic testing in a medical pathology laboratory.

A.3 What is your role in the current genomic sequencing program in Victoria?

7. As Director, I lead a team of scientists, computer scientists and epidemiologists who conduct genomic sequencing and analyse and report on genomic sequencing data. The team includes scientists who do the sequencing, bioinformaticians who take the raw data and undertake quality control checks and analysis and epidemiologists who report on the data. There is also a quality team in the laboratory who ensure all the work that is done is within the laboratory's quality framework and meets the laboratory's guidelines.

8. There is no single person in the team who does all the genomic sequencing work. As I have said, it involves a team of people who have different but complementary fields of specialised knowledge.

9. The opinions I express in this statement are based wholly on my specialised knowledge arising from my training, study and experience. In expressing those opinions, I am necessarily relying on the work and expertise of all the members of the team who have contributed to the genomic sequencing work.

10. The unit aims to sequence samples from all SARS-CoV-2 cases in Victoria. As a result of my work at MDU PHL, I know that comprehensive sequencing of a pathogen assists with understanding the relevance of sequence data. Sub-sampling can lead to biases in interpretation and these biases are reduced with comprehensive sequencing. Sequencing more samples improves the accuracy and robustness of the bioinformatic process (that I describe below). When we commenced SARS-CoV-2 sequencing, there were few cases in Victoria, and it was around that time the MDU PHL decided to sequence all samples we received.

11. I am a co-author of an article entitled "Tracking the COVID-19 pandemic in Australia using genomics". The article describes the genomic sequencing work done by the MDU PHL in the period 25 January 2020 to 14 April 2020. The MDU PHL continues to do the same genomic sequencing work that is described in the paper. The paper is available at: https://doi.org/10.1101/2020.05.12.20099929. The article is undergoing peer review for acceptance in a medical journal.

A.4 At a high-level, can you please describe the Doherty Institute, including the nature of its current corporate structure and the sources of its funding?

12. It is the MDU PHL that does the genomic sequencing work.

13. MDU PHL is part of the Department of Microbiology and Immunology at The University of Melbourne. The unit is located within the Peter Doherty Institute for Infection and Immunity.

14. The Doherty Institute is a joint venture between the University of Melbourne and the Royal Melbourne Hospital. It is governed by the Doherty Council. The strategic plan is led by its Executive Team. As MDU's Director, I sit within the Operation Management Committee which is responsible for the day-to-day operational activities of the Doherty Institute.

15. The MDU PHL provides services to the Department of Health and Human Services (the **Department**) and is funded by the Department to provide those services. The unit provides microbiology services for the investigation and control of communicable disease and food and waterborne outbreaks and serves as a microbiological reference laboratory delivering specialist public health microbiological services and surveillance activities for the State of Victoria. The predominant activity of the MDU PHL is providing services to the Department.

16. MDU PHL's program areas include food and water borne disease, hospital acquired infections and antimicrobial resistance, community acquired infections, sexually transmitted infections, vaccine preventable diseases and gastroenteric diseases. The MDU PHL is also the designated World Health Organization Regional Reference Laboratory for Invasive Bacterial-Vaccine Preventable Diseases.

17. The MDU PHL is the primary laboratory in Victoria doing sequencing of pathogens for public health purposes. It uses high throughput DNA sequencing technology to detect, identify and characterise public health pathogens. A pathogen is a microorganism that can cause disease.

18. With respect to COVID-19, the MDU PHL does genomic sequencing of SARS-CoV-2 cases and it provides some diagnostic testing services.

19. The MDU PHL is accredited under ISO 15189 for Human Pathology (Medical testing) and ISO/IEC 17025 for Environmental, Food and Beverage, Healthcare, Pharmaceutical and Media products testing (Biological) and Animal Health (Veterinary) testing and holds accreditation for Forensic Operations across all fields.

20. The MDU PHL operates under a Quality Management System, meeting the requirements of ISO 15189 and 17025 specifications, and associated regulatory

documents, which impact on MDU PHL's microbiological testing under the National Association of Testing Authorities (NATA), National Pathology Accreditation Advisory Council , Department of Agriculture, Water and the Environment, Department of Health and Human Services in relation to Security Sensitive Biological Agents, Office of the Gene Technology Regulator and the Therapeutic Goods Administration.

21. A public health laboratory carries out disease surveillance work that most other laboratories cannot do. It carries out enhanced testing and pathogen characterisation and reports the results to government to protect the public against disease. A public health laboratory does not usually undertake work for diagnostic purposes on a fee-for-service model.

B. Genomic sequencing: general

B.1 What is genomic sequencing?

22. A genome is an organism's complete set of genes or genetic material. The genetic material may comprise DNA or RNA. The human genome, bacterial genomes and some viral genomes are made up of DNA. However, SARS-CoV-2, which is a viral genome, is made up of RNA.

23. A genome sequence is the complete list of nucleotides (A (adenine), C (cytosine), G (guanine), and either T (thymine) for DNA genomes or uracil (U) for RNA genomes) that make up the genetic material of the organism.

24. Whole genome sequencing is the process to determine the complete sequence (DNA or RNA) of an organism's genome.

B.2 How does it work?

25. Whole genome sequencing may be broken down into two processes.

    (a) First, there is an analytical process undertaken in a specialized genome sequencing laboratory, using sophisticated laboratory hardware, to determine the complete genome of an organism in a single reaction.

    (b) Then, this genome sequence is investigated and compared with other genome sequences using bioinformatic software.

26. The first process is an analytic wet bench process. It encompasses the handling of samples, extraction of nucleic acids, library preparation (fragmentation, barcoding

(indexing) of DNA and amplification), flow-cell preparation and generation of sequence reads (FASTQ files). A published example of a whole genome sequence for SARS-CoV-2 that was sequenced at MDU PHL is at https://www.ncbi.nlm.nih.gov/nuccore/1803016604.

27. The second, bioinformatic process, uses software to interrogate the quality of the sequence data and to characterize the genome. Characterising the genome involves detection of genes and mutations. It also involves comparative genomic studies to compare the sequences from multiple different samples.

28. As a result of these processes, findings can be made about the quality of the sequence data. Comparisons between genomes enable conclusions to be drawn about the presence of mutations and allow inferences to be drawn about genomic clusters.

29. In the context of pathogens, the first sequence of a bacterial genome was reported in 1995. Since then, there has been a rapid escalation in technology and, in particular, in the last decade. All well-known human pathogens are now sequenced. Whole genome sequencing has been used for surveillance of other viruses including Zika, Ebola and influenza. This has been done for a range of reasons including to track the movement of a virus around the world and to inform vaccine development.

30. For SARS-CoV-2, there is no alternative to sequencing to identify, and discriminate between, clusters. Further, whole genome sequencing (as opposed to partial sequencing) is the highest resolution technology available to identify, and discriminate between, clusters.

B.2.1 What is a mutation in the viral genome? What do mutations tell us?

31. A genetic mutation is a permanent alteration in the genetic makeup of an organism.

32. Mutations can result from errors during normal replication of the viral genome or due to damage to the RNA genome (for example, exposure to radiation or chemicals). Mutations may or may not produce a change to the characteristics of the organism. Mutations in genes can either have no effect, alter the product of a gene or prevent the gene from functioning properly. Mutations can also occur in regions between genes (although, these are relatively small in viruses).

33. Mutations are the source of genetic variation of an organism and thus play a role in the evolution of the organism. Once a mutation occurs in the genome of a virus, it is

copied to and it is shared by all its descendant copies, this creates groups of viruses that share a mutation because of their shared ancestry to the exclusion of others. This shared ancestry forms the basis of the phylogenetic analyses discussed below. RNA viruses generally have high mutation rates compared to DNA viruses (100-1000X greater) because they lack the ability to repair errors that occur during viral replication.

34. Mutations help identify differences between viral genomes and make it possible to compare the genomes of multiple different sources to see how similar or different they are. As mentioned in the previous paragraph, viruses that share a group of mutations over all others are assumed to share a closer ancestor to each other than to any other viruses.

35. SARS-CoV-2, the virus that causes COVID-19, has a genome that is approximately 30,000 RNA nucleotides (represented by A, C, G, U) in length. The genome includes 10 genes which encode four structural proteins known as S (spike), E (envelope), M (membrane) and N (nucleocapsid) proteins and 16 non-structural proteins. The sequencing process can recover and reconstruct up to 99.8% of the SARS-CoV-2 genome, but this percentage varies based on several biological and testing factors.

36. Mutations in the SARS-CoV-2 genome have been occurring slowly. Mutations are often found in the gene that encodes the spike protein. Because SARS-CoV-2 is a newly evolved virus and because mutations occur slowly, there has not been time for many mutations to develop. These mutations can act as a "passport stamp" for the virus such that bioinformatic analyses can determine where a virus sample may have originated.

37. This bioinformatic analysis compares all the available sequences to each other and defines the differences at each point in the genome (single nucleotide polymorphisms, or SNPs).

38. The information about the location and type of mutation or SNP at each point in the genome is used to infer the likely relationships between different samples (and the cases they originated from). This type of analysis is called a phylogenetic analysis, and the data are visualised on a phylogenetic tree (discussed further below). A phylogenetic analysis is a statistical process used to infer the most likely ancestral relationships between samples given the observed sequence data and differences between them.

39. Sequences that share the same patterns of mutations are said to be highly genomically related, and the sequences occur close to each other on the phylogenetic tree (i.e., share a common ancestral/internal node to the exclusion of all others). If sequences are highly genomically related, then an epidemiological link is likely. This is seen, for example, in cases where the virus has been acquired from the same transmission network, e.g. transmission between members of a household.

40. In contrast, sequences with very different patterns of mutations are not closely related by genomics, and the sequences are more distant on the phylogenetic tree. This occurs, for example, with returned travellers who acquired COVID-19 in different countries.

B.2.2 What are epidemiological and genomic clusters? What do they tell us?

41. A cluster, whether epidemiological or genomic, is a group of people or samples with a condition or disease who have some similarity which suggests they may have acquired the condition from each other, from a common source or due to a common cause.

42. *Epidemiological clusters* are based on similarity in the epidemiological characteristics of person (e.g. demographics), place (e.g. attending the same location) and time, or a combination of these.

43. *Genomic clusters* are based on the degree of genomic similarity between the pathogens (such as a virus or a bacteria). Genomic clusters indicate the sequences contained within the cluster are more related to each other than they are to any other sequences in the dataset.

44. Identifying clusters is important in public health as it allows targeted investigation of cases within the cluster to identify and remove the source of infection and/or disrupt transmission chains.

45. The criteria used to define these genomic and epidemiological clusters are dynamic and are based on an understanding of the condition or disease, the pathogen and its biology, and how it may be transmitted and/or acquired. In both cases, cluster definitions and the results of any clustering analysis are dependent on the completeness of the included dataset.

46. Genomic sequences can be used to generate a phylogenetic tree. Figure 1 is an example of a phylogenetic tree. A phylogenetic tree is a visual representation of the likely evolutionary relationships between samples or sequences.

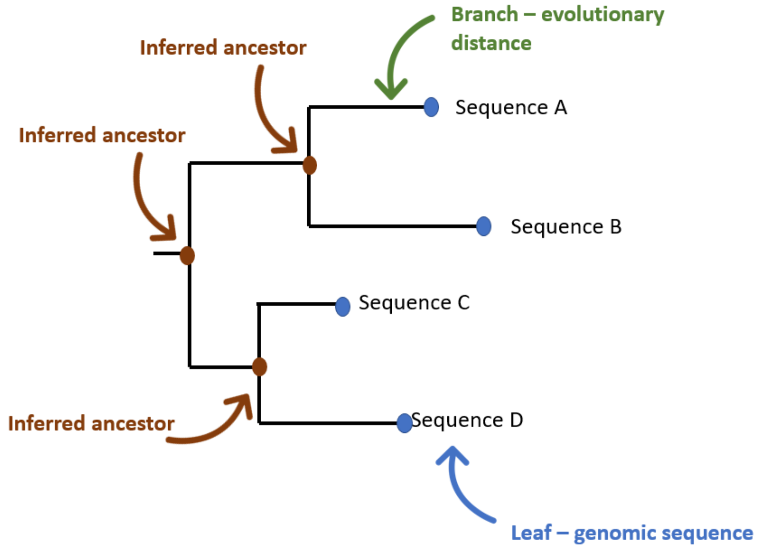


Figure 1: Annotated phylogenetic tree describing the evolutionary relationships between sequences

47. Each sample or sequence is represented as a *leaf* of the tree. In figure 1, the leaves are blue dots. The *branches* represent the genetic distance between the sequence and its inferred ancestral or parent sequences. The sum of the lengths of the horizontal lines that form the path between two *leaves* is a measure of the evolutionary distance between the two sequences, based on the pattern of mutations. So, for example, Sequence A in figure 1 is more genomically related to Sequence B than to Sequence C (longer total horizontal branch distance). The vertical lines are added to aid in visualisation and have no interpretative value.

48. The inferred parent sequences are *forks* in the tree and each one has a support value (called a "branch support value"). The higher the value, the higher the degree of confidence the "fork" gave rise to the child sequences. In the specific case of the analyses performed at MDU PHL, the support values can be interpreted as the probability of existence of the branch given the observed data. The degree of confidence is based on reconstructing the evolutionary history of the virus and the pattern of mutations. It is also based on running multiple iterations of the tree to show that each iteration infers the same relationship between the sequences. Put another way, the observed sequences stemming from that node are more closely related to each other than to other sequences in the analysis. This type of analysis is called a phylogenetic analysis.

49. For analyses of SARS-CoV-2 genomes currently conducted at MDU PHL, *genomic clusters* are defined using these support values and genetic distances.

50. Multiple clusters (or samples) may be merged only where they share a most recent common genomic ancestor and available epidemiological data supports the clusters as forming part of the same immediate transmission network; or split where the converse is true.

B.3 What is genomic sequencing used for?

51. In a public health laboratory, genomic sequencing is used for pathogen surveillance and outbreak detection/investigation. Amongst other things, genomic sequencing allows findings to be made about the resistance of a pathogen to antibiotics, how a pathogen is evolving, whether a pathogen is bringing in new genes and what is the disease-causing potential is for a given pathogen. As described above genomic sequencing also permits analysis of the relationship between pathogen genomes from different samples to detect disease outbreaks and identify the potential source of disease and transmission networks.

52. Genomic sequencing of SARS-CoV-2 is used at MDU PHL to identify genomic clusters that are likely to be epidemiologically linked.

53. In other contexts, genomic sequencing may have other applications.

B.3.1 How does sequencing allow us to determine the source of a particular case of coronavirus?

54. Genomic sequencing can identify possible transmission networks and help identify the probable origin of cases. Epidemiological investigations are then needed to support the hypotheses generated by genomic sequencing. Genomic sequencing data is not fully informative without epidemiological data.

55. The relationships identified from bioinformatic analyses of genomic sequences may provide information about whether a group of samples/cases are likely to have acquired the pathogen (here, SARS-CoV-2) from a common source or chain of transmission. Cases with sequences that are highly genomically related (in the sense described above, so that an epidemiological link is likely) are more likely to come from the same transmission network, compared to cases with sequences that are not closely related. In that way, genomic relationships can be used to identify cases likely to be part of the same transmission network where epidemiological data or case

identification is incomplete, or multiple epidemiological hypotheses exist regarding how a person acquired the infection.

56. Genomic sequencing can also be used to rule out a transmission network as a possible source of COVID-19 for a given person. For example, in Victoria, epidemiological data suggested a possible transmission network across four health services involving some 54 possible cases. However, genomic sequencing showed the cases in fact appeared in four separate genomic clusters. Most notably, samples from 3 cases from 2 health services were most closely related to others who attended the same social event and not those at the health services, excluding transmission within the facilities as a source of COVID-19 for these cases.

57. It is important to identify what genomic sequencing does not do. It might suggest two cases are part of transmission network, but does not and cannot prove transmission between the cases. Nor does genomic sequencing in isolation suggest the direction of transmission.

58. On the other hand, without genomic sequencing data, it would be difficult to draw conclusions about the success of public health measures taken in March-May 2020 to control SARS-CoV-2 in Victoria. And, without genomic sequencing data, epidemiological data alone could not reveal how the new cases since May 2020 emerged from expansion of a small number of new introductions.

B.3.2 How does sequencing allow us to determine strains of the virus? How does sequencing allow us to determine whether two cases of the virus are related?

59. Methods have been developed internationally to determine and describe different lineages (or "strains") of the SARS-CoV-2 virus. While description of these lineages is useful at a global level, they are not useful for cluster detection and transmission analysis as they lack sufficient resolution.

60. Sequencing and subsequent analysis can determine whether two or more samples have likely come from the same transmission cluster. Phylogenetic analysis of sequencing data can establish that the sequences are more closely related to each other than to other sequences in the analysis, indicating a common source or transmission network. Using these clusters and temporal data, it is possible to propose the emergence of new transmission chains, which may be visible as new genomic clusters or groups emerging from existing clusters, correlating with temporal data.

B.3.3 How does sequencing allow us to determine how closely two cases are related – for example, whether two cases are an instance of direct transmission, or whether two cases are separated by intermediate cases in the chain of transmission?

61. The genome sequencing data for SARS-CoV-2 alone cannot determine if two cases are caused by direct transmission or through an intermediate.

62. Analysis of the genome sequencing data will demonstrate clustering of these cases, but attribution of direct transmission events require support of epidemiological data. Additional data from epidemiological investigations, such as geographical relationships between cases, dates of travel or contact, can be combined with the genome sequencing analysis to establish direct transmission or if transmission was possibly via an intermediate.

B.4 Can you please explain, in simple terms:

B.4.1 the process for conducting genomic sequencing tests for COVID-19 (including what information you collect about individuals);

63. The *first* step is a patient with suspected COVID-19 disease (caused by the virus SARS-CoV-2) has a sample taken by a doctor or nurse, such as a nose and throat swab. This sample is tested for the presence of the virus at one of the diagnostic microbiology labs in the State. This is a diagnostic test. This initial test only looks for a small portion of the SARS-CoV-2 virus as an indicator that the patient has the infection.

64. If the sample is positive, the *second* step is to send it to MDU PHL for genome sequencing. The sample is registered on receipt at MDU PHL and checked against the paperwork received from the referring lab.

65. MDU PHL operates as a NATA accredited public health laboratory. Samples are referred under National Pathology Accreditation Advisory Council regulations, which require a minimum of two identifiers on the sample and three identifiers on the request form. Minimum identifiers for the sample include:

    (a) patient first name;

    (b) patient surname;

    (c) patient date of birth;

    (d) sample collection date;

(e) submitter's sample identification number.

66. Additional details collected from the referral form include:

(a) patient UR number (if associated with a hospital);

(b) submitting organization;

(c) patient address;

(d) patient phone number;

(e) patient Medicare number;

(f) symptoms;

(g) association with a known case – where available; and

(h) health care worker status.

67. Further data is provided to MDU PHL by the Department:

(a) date of diagnosis;

(b) date of symptom onset;

(c) contact with a known COVID-19 case during relevant infectious and incubation periods, and case identification numbers for such contacts;

(d) linkage to an exposure site and/or epidemiological cluster;

(e) overseas or interstate travel during the 14 days prior to onset/diagnosis, and location of travel;

(f) health care worker status;

(g) aged care residence;

(h) suspected source of acquisition (overseas travel, contact with a known case, unknown source, under investigation).

68. As set out below, sometimes the Department asks for genomic sequencing data for specific cases. When such a request is made, the Department usually provides additional data to MDU PHL.

69. The *third* step is the sequencing process. It involves:

(a) extracting the RNA genome from the sample;

(b) converting the RNA of the SARS-CoV-2 genome into a more stable form (complementary DNA, cDNA);

(c) amplifying the cDNA ("copying" the cDNA to increase the total amount for analysis) in small overlapping segments to cover the whole genome;

(d) sequencing the segments (library preparation, including fragmentation, barcoding of DNA fragments and amplification, flow-cell preparation and generation of sequence reads);

(e) analysing the sequence data to reconstruct the genome.

70. The sequencing process is performed using protocols from the international ARTIC network.

71. Each part of the sequencing process includes quality control checks to ensure the procedures are working as expected. Additionally, each sequencing result (genome) is also subjected to quality control checks to ensure the sequence quality is adequate for further analysis.

72. All samples that are positive for SARS-CoV-2 that arrive in MDU PHL undergo attempted whole genome sequencing. For the majority of samples, sequencing is successful. For some samples, however, especially those with low levels of virus in the sample, whole genome sequencing is not possible or is unable to yield genome data that passes quality control metrics.

73. The bioinformatics process is the *fourth* step and it involves the following steps.

(a) demultiplexing – multiple samples are pooled into a single sequencing run to ensure high throughput, thus in this first step, the data for each of the samples are identified and separated into separate files by using unique indexes that are added to each sample during the preparation of the sequencing libraries;

(b) creating a consensus sequence – the process of sequencing of DNA can be error prone, thus each portion of the genome is sequenced many times over to ensure accuracy (the international recommendation for SARS-CoV-2 is 1000X, and MDU PHL aims at 3000X), thus the process of recovering the genomic sequence of a virus in a sample involves finding the consensus nucleotide at each position of the genome over the 1000s of reads that are collected at each position through the sequencing process;

(c) multi-sequence alignment – once consensus sequences are identified for each sample, they must be aligned so that sequences from different samples are compared at the same genomic positions in the genomes in order to identify the differences needed to perform the phylogenetic analysis;

(d) tree inference – using a statistical model of how evolution occurs different tree topologies are generated, tested and discarded until an optimal tree is found that maximises the probability of the relationships found in the tree given the observed genomic data – because of the very large number of possible trees, the generation of new tree topologies for testing is directed towards new topologies that increase the probability of the new tree relative to the current tree; and

(e) identifying clusters.

74. From receipt of the sample, the whole process takes about 3 to 5 days, depending on the workload. The timeframes are currently longer than that because of the large number of samples.

B.4.2 the differences between the tests to establish whether a person is positive for COVID-19 and the genomic sequencing tests conducted by Doherty Institute;

75. The test to establish whether a person is positive for COVID-19 is a *diagnostic test,* which means the sample is tested by a method approved by the TGA (meeting certain criteria for accuracy and precision). Its purpose is to provide a simple yes/no answer to the question of whether a person is currently infected with the virus.

76. A diagnostic test involves testing a sample for the presence of SARS-CoV-2 viral RNA using multiplex polymerase chain reaction (PCR test). The process involves extraction of the viral RNA genome, conversion of the RNA to cDNA and then amplification targeting one or more regions of the genome. Usually, two regions are amplified to confirm COVID-19 infection in a patient. The regions that are amplified are small (<300 nucleotides) and are in conserved regions of genes, including E, S, N, RdRP and Orf1ab. The regions chosen are based on their ability to differentiate SARS-CoV-2 from other Coronaviridae, including MERS-CoV, SARS-CoV-1 and common coronavirus (causing a mild respiratory illness, similar to a cold). They are also regions that are less likely to mutate due to their essential role in viral function. These regions are informative for diagnosis, but not informative for

identifying diversity of the virus between patients. These regions represent less than 1% of the genome. The exact regions amplified for diagnosis vary depending on the diagnostic assay used. Note that commercial companies do not identify the exact nucleotide region that their assay detects, instead they only indicate the gene the assay targets, such as S, E, Orf1ab, N.

77. Samples that have been identified as positive for SARS-CoV-2 by the PCR test are then used to perform genome sequencing of the virus present in those samples. The sequencing process can start from the original sample or the RNA extracted from the original sample.

B.4.3 the phenomenon of, and potential for, false positives in the Victorian context?

78. The answer to this question depends on what is meant by false positive.

79. False positives can be an issue in diagnostic testing.

80. False positive tests occur when a positive diagnostic test result is issued from a laboratory in the absence of disease in the patient. Some potential causes of false positive test results include mislabelling of specimens, data entry errors, contamination of the primary specimen, misinterpretation of a test result, or off-target test reactivity caused by an unsuitable testing platform set up or use or limitations of the test. The TGA provides guides on what levels of false positives are acceptable for diagnostic tests.

81. Off-target reactivity is relatively uncommon, but is an inherent characteristic of a PCR assay, the relative robustness or otherwise of assay design and of clinical implementation. Off-target reactivity may include (a) cross reaction with non-target genetic material, and (b) self-priming phenomena in the absence of target. Examples of non-target genetic material can include related coronaviruses.

82. In addition, highly sensitive assays may detect non-viable virus from past infections. In these cases, the test result is truly positive, however, the patient does not have an active infection.

83. The Australian Public Health Laboratory Network emphasises that the likelihood of false positive diagnostic results occurring in Australia is very low. Australian laboratories performing SARS-CoV-2 diagnostics by PCR testing are required to implement the National Pathology Accreditation Advisory Council's Requirements

for Medical Testing of Microbial Nucleic Acids quality framework, which includes procedures to minimise the risk of false positives tests.

84. False positive are a theoretical possibility for genomic sequencing, as the process might identify a variant (mutation) in the sequence where no variant exists. However, as I describe further below, the MDU PHL has validated its processes specifically looking for false positives (amongst other things) and did not find any.

C. Genomic sequencing: application to Victorian cases

C.1 Can you please explain the process by which the Doherty Institute and DHHS collaborate with regards to genomic sequencing for COVID-19 and how this has evolved over time?

85. The Department funds the genomic sequencing of SARS-CoV-2 samples. MDU PHL carries out the genomic sequencing and bioinformatics analysis.

86. MDU PHL provides the results to the department at least weekly during online meetings, primarily in the form of phylogenetic trees and identifying genomic clusters. In these meetings, the Department provides some epidemiological data to MDU PHL to assist with confirming or understanding transmission networks. Sometimes, the Department asks for genomic analysis of specific cases of interest and the Department provides MDU PHL with the patient ID and known epidemiological exposure details. An example of this is the putative transmission network involving four health care services that I referred to above.

87. There is a strong collaborative effort on the part of MDU PHL and the Department on developing pathogenic genomics as a public health tool.

C.1.1 Who owns the sample and the intellectual property?

88. I believe that MDU PHL own the samples and all intellectual property that is developed or created. I understand the Department has a licence to use any intellectual property the Department does not own and the Department can use it, including by publishing it for public health purposes.

C.1.2 Who owns the data?

89. I understand all genomic sequencing data, including the interpretation of the sequence data, is owned by MDU PHL. Again, I understand the data, and related information and reports, are provided to the Department and the Department is able to use that information, including by publishing it for public health purposes.

90. MDU PHL publishes the genomic sequences of SARS-CoV-2 samples on public databases. I referred to an example above.

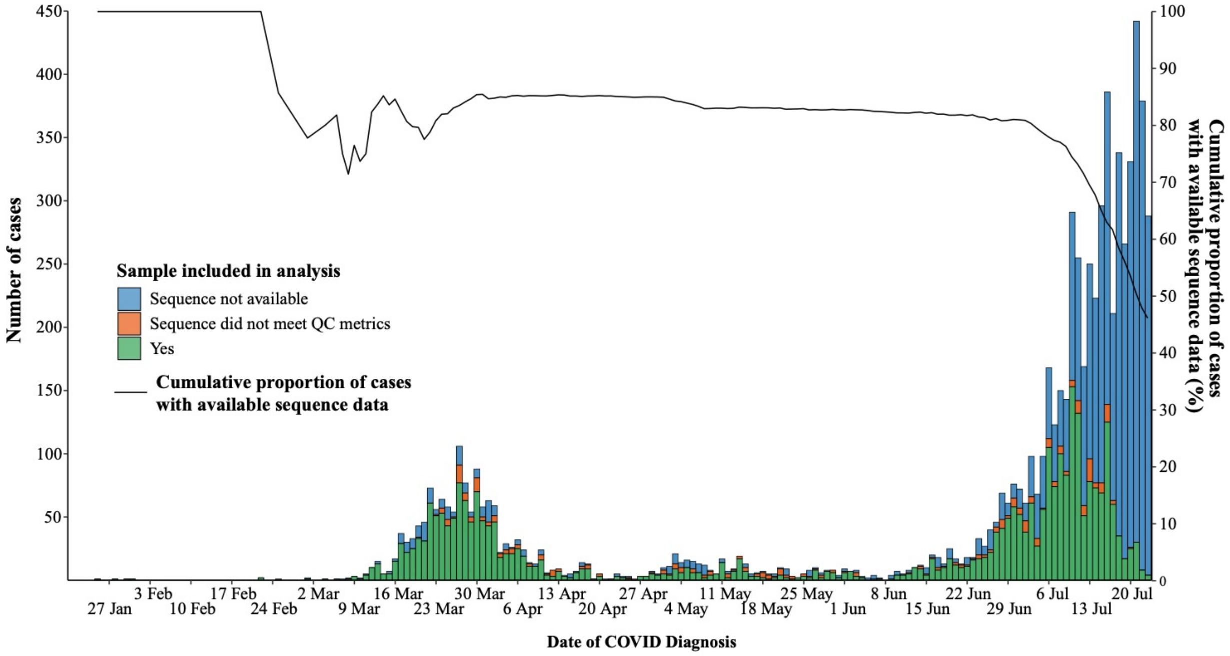
C.2 What percentage of Victorian cases were you able to sequence, over what date range?

91. As of 29 July 2020, sequence data was available for 3392/7347 (46%) of Victorian COVID-19 cases diagnosed up to and including 23 July 2020.

92. Sequencing is ongoing or yet to be performed for samples from many recent cases, resulting in recent weeks in a reduced proportion of cases with available sequence data overall. As figure 2 shows, the cumulative proportion of cases with available sequence data was previously around 80% and has fallen as case numbers have increased.

Figure 2: Epidemic curve of Victorian COVID-19 cases diagnosed between 20th January and 23rd July 2020, by sequence data available as of 29th July 2020.

Total number of cases diagnosed each day, as provided by DHHS, shown on the x-axis, with each day stratified by the availability of sequence data, as indicated in the primary figure legend. The epidemic curve is annotated with the cumulative proportion of cases sequenced, indicated on the secondary Y-axis. The drop in proportion of cases sequenced in July may include cases currently undergoing and/or pending sequencing.



93. Figure 2 displays the total number of cases diagnosed each day on the x-axis, as provided by the Department. The number of cases each day is stratified by the availability of sequence data (as indicated in the primary figure legend). The epidemic curve is annotated with the cumulative proportion of cases sequenced, indicated on the secondary Y-axis. This demonstrates the fall in July of the proportion of cases with available sequence data. As a result, none of the identified clusters include these recent cases marked as "not available" (blue). A sequence is not available where the

results of the sequencing are pending, the sample was unable to be sequenced, a sample was lost or not provided by the diagnosing laboratory.

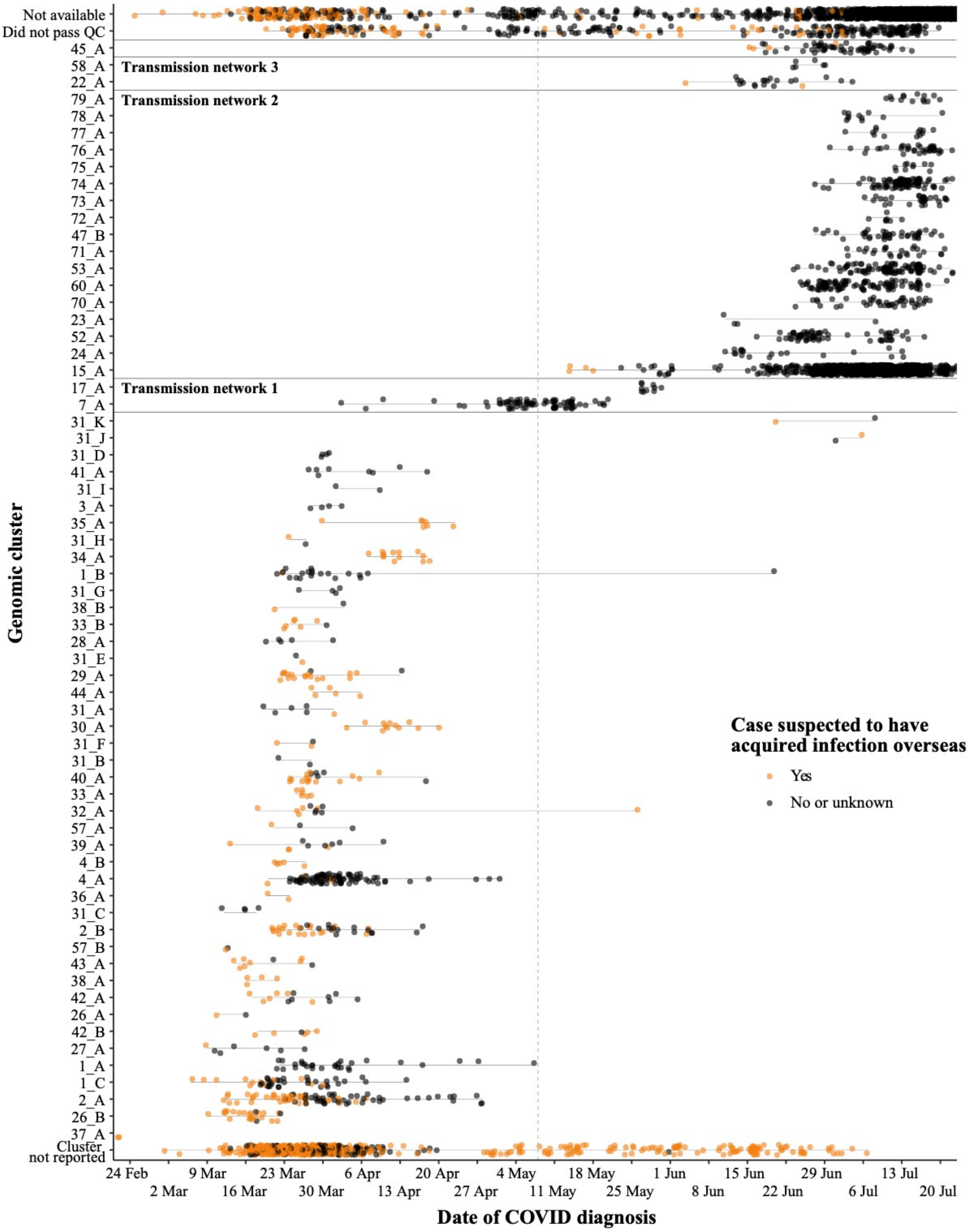
94. Sequences for 42% (3080) of cases during the same time period met all quality control metrics and have been included in the results below.

C.3 What were your findings?

95. As of 29 July 2020, 65 *genomic clusters* were identified. These genomic clusters ranged in size from two to 1071 cases (median 10 cases). A timeline of cases associated with each genomic cluster is seen in Figure 3. The figure presents the relationships from a phylogenetic tree in clusters and presents those clusters over time.

Figure 3: Genomic clustering of Victorian COVID-19 cases diagnosed between 21st February and 23rd July 2020, with sequence data available as of 29th July 2020.

Each graph point represents a single case, with date of diagnosis represented on the X-axis and reported genomic cluster on the Y-axis. Cases are coloured orange if the case is suspected to have acquired COVID-19 through overseas travel (data provided by DHHS) and as indicated in the Figure legend.



Note: A transmission network, in this context, represents a group of very closely related genomic clusters with a most common recent ancestor. Each transmission network is thought to represent a separate but single importation of the virus into Victoria, with genomic diversity, which is well supported but small in magnitude, arising in Victoria as the virus has circulated within the community and resulting in mutliple closely related genomic clusters.

96. Each graph point in Figure 3 represents a single case. The date of diagnosis of the case is represented on the X-axis. The reported genomic cluster is on the Y-axis. Cases are coloured orange if the case is suspected to have acquired COVID-19 through overseas travel, based on data provided by the Department. The names given to clusters are arbitrary and do not indicate relatedness between genomic clusters in any way.

97. The "not available" line has the same meaning as for Figure 2. To reiterate, each dot represents a case, rather than a sequence and the horizontal line is present for visual representation only. Once sequences become available for the cases on this line, they will appear in the main image.

98. The "did not pass QC" line identifies the cases where the sample did not pass quality control and hence no sequence is reported for that case.

99. The "cluster not reported" line identifies cases where a sequence was obtained, but the sequence was not within a reported cluster with other sequence(s). Each of these cases (except one) since 8 May 2020 is from a case where the infection is believed to have been acquired overseas.

100. Each horizontal line (except the heavy lines dividing transmissions networks 1, 2 and 3, and the lines for "not available", "Cluster not reported" and "did not pass QC") represents a genomic cluster.

101. The reported genomic clustering is broadly categorised into two periods (marked by the vertical dotted line):

(a) Period 1 (~1st March – 7th May 2020); and

(b) Period 2 (~ 8th May onwards)

102. Period 1 is characterised by the presence of many diverse genomic clusters each containing a small number of cases. 44 clusters were identified in this period. From 7 May 2020, based on the sequence data included, transmission appears to have ceased in most of these genomic clusters, with cases identified in only three of these clusters following this date.

103. Period 2 is characterised by the expansion of three transmission networks and an additional genomic cluster (45_A), each with a high branch support value. Each transmission network is a group of closely related genomic clusters with a most common recent ancestor. Each transmission network is believed to represent a

single importation of the virus into Victoria, supported by epidemiological clustering and travel history data.

104. The transmission networks are marked on Figure 3 and can be described as follows:

(a) Network 1 (91 sequenced cases) was first identified in March and expanded rapidly throughout May. No further cases have been identified within this transmission network since 30 May 2020.

(b) Network 2 (1705 sequenced cases) was first identified in mid-May in a group of returned travellers. Additional cases were identified within this transmission network throughout June and continuing into July. This network includes 17 clusters which appear to have originated from the earliest cluster (15_A) based on the data available to date.

(c) Network 3 (27 cases) and genomic cluster 45_A (65 cases) were both first identified in returned travellers during June, with additional cases identified throughout June and into July.

105. I referred above to the branch support value and said the higher the value, the higher the degree of confidence the grouped sequences share a more recent common ancestor to the exclusion of all other samples. For each of transmission networks 1, 2 and 3, the branch support values were 88%, 100% and 98% for each network respectively, and 100% for cluster 45_A. These would all be considered a high degree of confidence that these clusters are truly a separate group from the surrounding sequences.

106. Of the 1837 cases diagnosed since 8 May 2020 where overseas acquisition was not suspected and with available sequence data, 1833 (99.8%) were identified within one of the three local transmission networks, or genomic cluster 45_A. There is no evidence of ongoing transmission of any other known genomic clusters within Victoria since this date.

107. Transmission networks 1, 2, and 3 each contain some genomic diversity which is strongly supported statistically, but small in magnitude. This diversity has been used to classify cases within these transmission networks into multiple genomic clusters to assist public health investigations. However, I note that for each of these networks this diversity has likely arisen within Victoria as the virus has circulated within the community. I say this because of the phylogenetic clustering, and the strong branch

support for the clusters from which the transmission networks are derived. For that reason, each transmission network is believed to represent a single, independent importation of the virus into the state.

C.3.1 How many clusters and what proportion of cases were linked to quarantine travellers in the Hotel Quarantine program both (a) prior to 14 April 2020; and (b) since 14 April 2020?

C.3.2 Were these cases confined to particular hotel sites? Which ones?

C.3.3 What proportion of cases were linked to private security staff at those sites?

C.3.4 What proportion of cases were linked to other staff in the Hotel Quarantine program?

C.3.5 What proportion of cases were linked to cases or other sources not related to the Hotel Quarantine program?

C.3.6 What were your findings in relation to onward transmission of infection from staff in the Hotel Quarantine program to persons in the community? Where did these cases occur?

C.3.7 Based on these findings, to what extent was the increase in the spread of the COVID-19 virus in Victoria attributable to quarantined travellers or staff in the Hotel Quarantine program spreading the virus to the broader Victorian community? Over what date range?

C.3.8 If that spread of the COVID-19 virus was attributable to quarantined travellers or staff in the Hotel Quarantine program, what do your findings suggest about whether the sources of that spread included all staff in all roles, at all sites – or whether the sources of spread were specific to: (a) certain individual quarantined travellers; (b) certain individual staff; (c) staff in specific roles (eg security staff); (d) staff at particular sites; or € staff employed by particular sub-contractors?

C.3.9 Is there any data that would allow us to compare the rates of spread attributable to the Hotel Quarantine program in Victoria, compared to other equivalent programs in Australia (for example, in New South Wales)?

108. To answer each of the above questions C.3.1 to C.3.9 in the terms asked requires the addition of epidemiological data. This epidemiological data is not held by the MDU PHL. I believe it is held by the Department. MDU PHL does not and cannot make findings of the kind referred to in these questions.

C.4 What are the limitations on the reliability of these findings?

109. Identification and interpretation of genomic relationships between SARS-CoV-2 samples is dependent upon:

(a) the accuracy of sequencing and the subsequent generation of consensus genomes;

(b) the identification of genomic clusters from trees that are derived from the consensus genomes; and

(c) interpretation of genomic clusters for epidemiological purposes.

110. Each of these was considered in detail and analysed in a validation report prepared by MDU PHL about sequencing SARS-CoV-2.

111. To demonstrate the accuracy of sequencing and the generation of consensus genomes, one sample was resequenced 20 times. The validation process involved looking for genetic variations between all the repeat sequences and a control. The 20 repeat sequences were completely identical, indicating 100% accuracy of the sequencing process. The process also involved checking the repeat sequences against a published, reference version of the sequence of the same genome, to ensure concordance between the validation sequences and the reference sequence. In this process, the accuracy was 98.9%. The 1.1% is accounted for by instances where no consensus base could be identified in the resequences (i.e. incomplete sequencing data, consistent with the known limitations of SARS-CoV_2 sequencing). In no case was an incorrect base identified in the repeat sequences (i.e. no false variants introduced by sequencing process).

112. To validate identification of genomic clusters, we sought to ensure the process would identify a genomic relationship between multiple sequences of the same sample. A total of 63 samples for which multiple sequences were available were used, with an additional 50 publicly available sequences used for context. A relationship was identified between multiple sequences of the same sample for 93.5% of the samples. This result means there is some possibility of under-inclusion in the identified clusters, in that a sequence might not be included within a cluster when it should have been. The remaining 6.5% was accounted by samples where there were deletions in the consensus sequence (missing data, as described above), resulting in their false exclusion from the cluster.

113. The completeness of genome recovery for SARS-CoV-2 is variable due to factors such as sample degradation prior to sequencing, and low viral load in the sample. This may result in missing genomic data, where the base at a certain position may not be able to be determined due to low coverage (insufficient data) or ambiguity in

sequence (unclear which base is present at that position). This can impact on the calculations of genetic distances and the support values used to identify genomic clusters, as demonstrated above.

114. The process of identifying genomic clusters uses an algorithm that takes as input the phylogenetic tree and threshold values of branch support and genetic distance. As a first step, it breaks the tree up into smaller sub-trees based on the branch support values. It then calculates the genetic distance among the *leaves* of the sub-trees. If the maximum genetic distances are larger than the threshold, then it breaks the sub-tree into smaller trees based on the branch support values and repeats the genetic distance calculation. The algorithm iterates over this process until it has exhausted all the possible sub-trees. Critical to this approach are the branch support values. When the tree is being built, if there are deletions in the consensus sequence, the algorithm sums over the four possibilities of nucleotides at each deleted site (e.g., assuming the site had an "A" what would be the probability of the tree, then assuming the site had a "C" what would be the probability of the tree, etc). This is the international best practice to account for the uncertainty in the data. However, this process may reduce the branch support values, thus potentially affecting the clustering algorithm.

115. I referred above to the branch support values for transmission networks 1, 2, 3 and genomic cluster 45 being 88%, 100%, 98% and 100% respectively. For samples within these networks, the algorithms have been run 1000 times. For a given sample within transmission network 2, every time they were run, the sample was placed within transmission network 2.

116. To validate the interpretation of genomic sequences, three genomic epidemiologists independently analysed a given phylogenetic tree, and allocated each sequence to a genomic cluster. The validation report demonstrated that this process of genomic cluster allocation reliably identified genomic relationships for sequences with known epidemiologic links (e.g. household contacts), and never identified a genomic relationship where the epidemiological data showed no relationship.

117. Because SARS-CoV-2 is a new pathogen with, at present, low diversity, this can make it harder to differentiate clusters and this may result in the lower certainty about the identification of clusters. However, for transmission networks 1, 2 and 3, and genomic cluster 45_A we have a high degree of certainty that these transmission networks truly exist, due to the high branch support values, as outlined above.

C.4.1 Do they depend on any assumptions?

118. I have answered this question in my previous answer, particularly by reference to the algorithms used.

C.4.2 In your field of expertise, could you please outline any dissenting opinions on the efficacy of genomic sequencing in relation to COVID-19?

119. Within the field of microbiology, I am not aware of any dissenting opinions on the efficacy of genomic sequencing in relation to COVID-19 for the purpose of identifying and discriminating between clusters. Put another way, I am not aware of any microbiologists who would say that genomic sequencing should not be used to identify or discriminate between clusters.

Ben Howden

4 August 2020

IN THE MATTER OF THE *INQUIRIES ACT 2014* **NTP-027**

AND

IN THE MATTER OF A BOARD OF INQUIRY INTO THE COVID-19 HOTEL QUARANTINE PROGRAM

INDEX OF DOCUMENTS ANNEXED TO WITNESS STATEMENT OF PROFESSOR BEN HOWDEN

Number	Document Title	Document Date	Paragraph of statement
1.	CV of Professor Howden	4th August, 2020	4

None of the documents listed above is subject to a claim for "reasonable excuse".