Rapid-Review: The role COVID-19 genomics can play in contact tracing, cluster analysis and viral evolution.

Prof Michael Bunce (8th-15th September 2020).

Context & Scope:

"Systematic pathogen surveillance is within our grasp, but is still undervalued and underfunded relative to the magnitude of the threat" **– Prof Michael Worobey, 2017 - *Nature* 547**

In the COVID-19 global pandemic over 100,000 viral genomes have now been characterised, and made accessible to the global community[1,2]. In the context of Aotearoa New Zealand's pandemic response, the genomic sequencing of positive COVID-19 samples will become an increasingly important tool that must be seamlessly integrated into our holistic health response. This is the first time, globally, that genomic data has been truly integrated with epidemiological data in real time. The fact that this has occurred 'at pace' has highlighted a number of areas for improvement and capacity building.

The aim of this rapid review is to describe and explore the role genomic sequencing and (phylogenetic) analysis has played, and may play, in managing the pandemic across Aotearoa New Zealand. Furthermore, I explore the importance of integrating genomic data into our analysis of COVID-19 clusters and contact tracing, as well as how the data are presented to decision-makers and the public. Finally, this paper will provide some genomic-focused recommendations that might better prepare Aotearoa New Zealand to manage not just the current phase of this pandemic, but also our wider capacity to tackle other disease(s) that impact our people, industries and biodiversity. Below I present a set of 'key recommendations' (within executive summary) and recommendations for consideration and action.

Executive Summary: Observations and Key Recommendations:

Observation 1: Aotearoa New Zealand's genome sequencing effort to date has been world-leading. Over 50% of all cases that tested PCR positive now have a genome sequence recorded. There has been adoption of rapid (and semi-automated) workflows and well executed genomic analyses. Decision-makers have, rapidly appreciated the strategic value genomic analyses can bring to pandemic management in both the short, and long-term. Aotearoa New Zealand is one of the few countries to generate and publish a nationwide genome study[3]. The challenge ahead is to build capacity, redundancy, precision, dynamic reporting and speed (of sample delivery, processing and reporting).

> ***Key Recommendation 1.1*** Aotearoa New Zealand should continue its innovation in the COVID-19 genomics space. The speed at which samples are shipped and genomes can be sequenced remains a priority if genomic data is going to be useful in real-time contact tracing and cluster

[1] https://nextstrain.org/ncov/global
[2] https://www.gisaid.org
[3] Geoghegan et al. 2020: doi: https://doi.org/10.1101/2020.08.05.20168930

analysis. There are a number of 'pipette ready' projects that could rapidly build capacity in this area and make a tangible difference to Aotearoa New Zealand's pandemic readiness, both now and in the future. Importantly, Aotearoa New Zealand needs to build redundancy (both laboratory and expertise) into the system to ensure there is surge capacity and no single point of failure that would disrupt the flow of genomic data when it might be needed under urgency.

Observation 2. While our health system is familiar with reporting positive/negative test results from genetic studies, the rapid incorporation of genomic data is a new area that requires considered approaches - especially in how the data is graphically presented and communicated. The reporting of genomic data can, and should, be improved. There are short- and long-term solutions to this.

Key Recommendation 2.1 A priority system for genomic analysis needs to be embedded in the pandemic response system that can rapidly report viral lineages (and mutations therein) from 'urgent samples' to contact tracers. The flow of sample metadata (e.g. dates, addresses) that should be pinned to the genomic data needs to be improved between stakeholders. A rapid genome reporting pathway incorporating all relevant metadata is essential to real-time responsiveness.

Key Recommendation 2.2 A common language and consistency of data presentation is needed when exploring genomic data that acknowledges the strengths, weakness and precision of different data types. This will involve genomic experts generating text and graphics that health professionals and decision-makers can easily and accurately comprehend (often in a time-sensitive setting). Critically, reports should also convey 'time & space' as a cluster expands or is contained. It is vital that these same professionals and decision-makers attain a level of genetic literacy so they can ask the 'right questions' of the data - case studies and educational materials could help expedite this process and build capacity.

Key Recommendation 2.3 A closer-working relationship between epidemiologists, geneticists and local health responders is needed to extract the most from genomics. ESR and MoH have started this collaborative journey - there is now a need to interweave public health units across the country into these new workflows. Investment is likely needed to train and develop individuals who can work proficiently at the interface of these fields and be effective communicators.

Key Recommendation 2.4 Investment and collaboration is needed to develop bioinformatics and data science solutions (software and databases) for real-time reporting. While manual curation of data will be required in the short term, this cannot be conducted at pace or at scale, nor does it position Aotearoa New Zealand well for future outbreaks/pandemics. There are options available within Aotearoa New Zealand to 'prime' this area.

Observation 3. Missing genomic data from Aotearoa New Zealand's COVID-positive samples has, and will continue to impact our ability to track the source of any outbreaks that might escape border quarantine. There is room to improve both processes and priorities in this area.

Key Recommendation 3.1 Protocols and workflows should be put in place to sequence ALL samples that test positive. This may require multiple samples be taken for every MIQ person, alternatively, rapid follow-up testing as soon as a positive PCR result is recorded. Secondary or follow-up swabs would be to obtain a variety of sample types/quality specifically with genomics in mind. There should not be a reliance on the single positive sample as storage, shipping and sample degradation will introduce gaps/weaknesses in the genomic surveillance network that severely impact our ability to infer where or when breeches might have occurred.

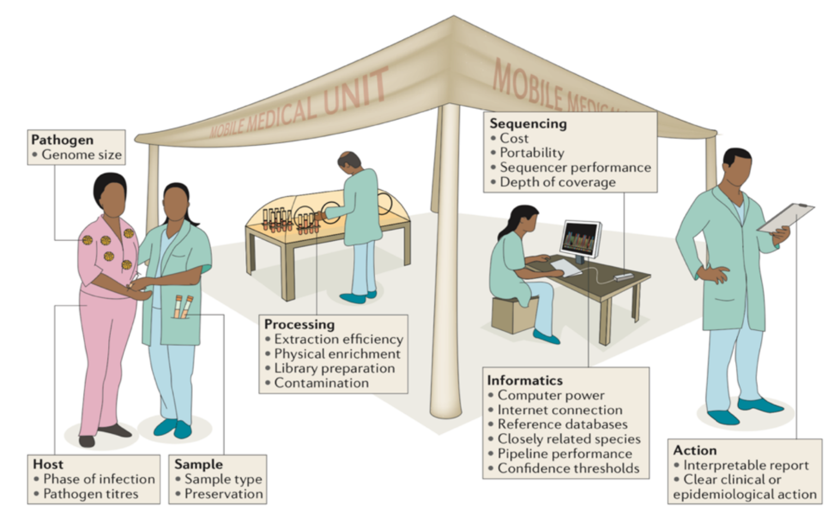
Key Recommendation 3.2 Samples that fail to yield full viral genomes can, and should, be analysed using shorter fragments of the genome that are centred around the variable sites (i.e. mutations) in the virus. This approach will ensure that, in the absence of a full genome, there is still the possibility of determining the viral strain from a small (diagnostic) section of the virus (this research capability is underway).

Observation 4. The novelty of the SARS-CoV-2 virus in human hosts means the virus has not experienced significant genomic selection – the ~2000 or so variations observed have been largely ‘random’ and are scatted across the genome. This will change when vaccines and medicines start to select for viral lineages.

Key Recommendation 4.1 Aotearoa New Zealand needs to embrace genomic tools and analyses for long-term monitoring of viral evolution. This is not simply an academic exercise, rather there is a pressing need to monitor the viral lineages that are circulating (akin to seasonal influenza tracking). This is vitally important once a vaccine is deployed in Aotearoa New Zealand and border restrictions are relaxed. Genetic characterisation of circulating lineages needs to become routine practice. Capacity that is built in this area can easily be adapted to other disease outbreaks in people as well as animals (e.g. *mycobacterium bovis*).

Observation 5. The high public and political interest in COVID-19 means there is a need to develop tools for; education, communication and engagement.

Key Recommendation 5.1 Resources such as video, graphics, animations and webinars should be developed to explain the science behind COVID-19 genomics and how it is being implemented at our front lines. This should build on the foundational communication work by ESR and University of Otago. Visual tools must accurately represent the data and be cognisant of cultural and privacy concerns. Given the public interest in this science there is an opportunity to explain and educate about the wider benefits of genomic testing/technologies including how vaccines are made and tested.



2018 illustration that envisages a genomics-informed heath response that has fast become a reality under COVID-19 pandemic scenarios. “*When informed by a One Health approach, in which human, animal and environmental health are considered together, such a genomics-based system has profound potential to improve public health*” Source: https://www.nature.com/articles/nrg.2017.88.pdf

Rapid Review – 8th -15th September 2020.

Background to Genomic Analysis of COVID-19:

1. COVID-19 is a disease caused by the SARS-CoV-2 coronavirus. The virus has a genome composed of RNA. The first entire genome of SARS-CoV-2 was released on a database on Jan 12th January 2020 and published soon after[4]. Once researchers determined the sequence of a virus (approximately 30,000 RNA bases in length) PCR tests were designed to quickly detect it – these tests have been at the core of global testing.

2. The testing for COVID-19 (using a technique called PCR) detects the presence of only a small piece of the virus – this is not the same as sequencing the entire genome. There are a number of approaches to sequence viral genomes, the most common being a derivative of the ARTIC protocol[5] that is in used here in Aotearoa New Zealand by ESR. The term WGS (whole genome sequencing) is used interchangeably with genomics to describe this analysis.

3. Genomic sequencing enables the changes (i.e. viral mutations) to be analysed. Since the first genome was sequenced there have only been a handful of 'core' viral lineages that have arisen. For example, one of the 30,000 letters in the genome of SARS-CoV-2 changed from an A to a G at position 23,403. This is the predominant lineage (termed **'B'**) that has spread around the globe[6]. If any two viral genomes are compared across the globe they will differ on average by ~10 mutations[7] (however up to 25 mutations may be observed).

4. Within these major viral lineages 'spot' mutations are also observed - these background mutations leave a trail of 'genetic breadcrumbs' that are increasingly being utilised in contact tracing and cluster analysis.

How are Genomes Sequenced?

5. When a patient is swabbed for a possible COVID-19 infection the first step is to collect and isolate the genetic material – namely RNA – from the sample. The PCR test for COVID-19 involves converting part of this RNA to DNA and then 'photocopying' (amplifying) a small diagnostic part of the virus. A sample is deemed positive if viral RNA 'photocopies' are detected. Conversely, a sample is deemed negative if the viral RNA is not detected. As well as returning a positive/negative test PCR will also usually provide a measure of how much viral RNA is there (a metric referred to as a C_T value). Despite the analytical sensitivity of PCR there is a still a false negative rate. This is important to consider, as false negative samples are never flagged for genomic analysis.

6. Samples from positive COVID-19 tests can vary in quality and quantity. A real possibility is that (i) there is very little virus present, (ii) sometimes swabbing may not have been conducted well, (iii) other times the viral RNA can degrade (for example if the sample is stored poorly). The quality

[4] https://www.nature.com/articles/s41586-020-2008-3

[5] doi: https://doi.org/10.1101/2020.09.04.283077

[6] https://www.sciencemag.org/news/2020/07/pandemic-virus-slowly-mutating-it-getting-more-dangerous#

[7] https://www.nature.com/articles/d41586-020-02544-6

and quantity of the viral RNA is a key determinant in the success or otherwise of efforts to obtain a genomic sequence of SARS-CoV-2.

7. If a PCR positive test is returned, there is a decision to make – should the sample be sent for genomic sequencing? Aotearoa New Zealand's viral burden is so low that the answer to this should be 'yes'. The sequencing of viral genomes has, to date, been primarily conducted by ESR, who have developed in-house expertise and robotics to sequence and analyse up to 200 viral genomes per week. There is also developing capacity at the Massey University Albany Campus who have sequenced genomes and are collaborating with ESR.

> ***Recommendation:*** That Aotearoa New Zealand make it an immediate priority to sequence the genomes of all positive cases that are detected within the borders (this action is already underway). There should be the expectation that testing centres around the country rapidly ship swab samples and/or RNA to a nominated genomic testing centre together with key information (metadata) on the PCR test result (for example, C_T values where appropriate).

8. At ESR viral genome sequencing is based around a protocol called 'ARTIC'[8] which, breaks the virus into lots of bite-sized overlapping pieces (of about 1200 or 400 bp each). Each of these pieces are then sequenced and then reassembled to determine the viral genome. Some viral genomes are technically incomplete because some pieces fail to sequence. Raw data from sequencing is filtered and checked to ensure the quality of each base-call (i.e. A, T, C or G) across the genome. Despite these checks, there is still some residual possibility for errors, especially in samples with low viral numbers.

9. ESR has been using a combination of Nanopore sequencing and Illumina sequencing technologies. Both approaches are fit for purpose, though the Nanopore protocol sequences longer fragments of virus and can generate the data quicker (typically 24hrs earlier). For this reason, it has been favoured by ESR as they look to rapidly process samples for contact tracing.

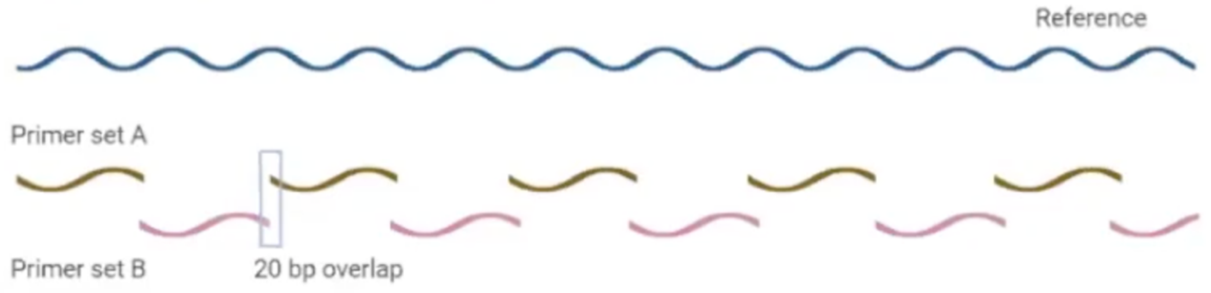


Figure 1. Schematic of how the COVID-19 genome is sequenced in overlapping 'pieces' that are assembled into a viral genome that is hopefully both complete and accurate.

How are genomes analysed?

10. Viral genome data can be analysed in a number of different ways. In many respects, the way in which data is presented depends on the questions asked of the genetic analyses. One common

[8] doi: https://doi.org/10.1101/2020.09.04.283077

way of representing viruses, including COVID-19, is in the form of trees (also called a phylogenies). These can show everything from the total global viral lineages (see Figure 2a) or just a tree containing cases from a recent cluster (e.g. Auckland outbreak, Figure 2b).

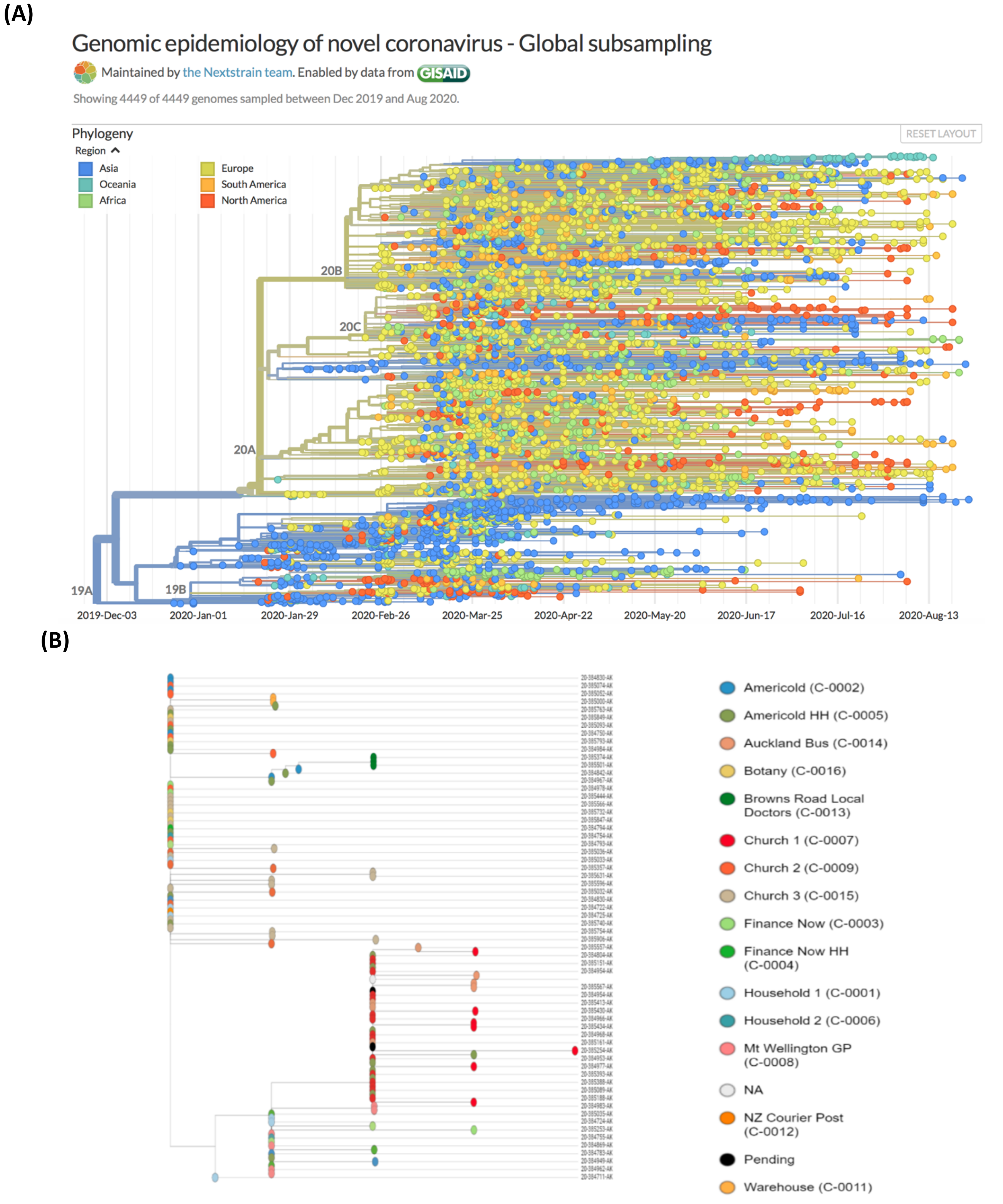


Figure 2. Examples of COVID-19 genome data represented as trees (phylogenies). Such trees are a hypothesis of how viral genomes are related to each other. One key point when interpreting trees is that they represent 'time' – samples at the root of the tree are back in time relative to the tips. **(A)** global tree of viral genomes from *NextStrain* and **(B)** a more localised tree that focuses on a single cluster imaged in *Microreact* (Source; ESR report dated 8th September)

11. One of the aims of building trees is to accurately assign viral genomes to a lineage (also commonly referred to as viral 'strains'). In mid-July 2020 there were more than 80 distinct viral lineages. These lineages assigned to the major group designations; **A** (~6 lineages) **B** (~16 lineages) and **B.1.** (subdivided into ~70 sublineages)[9]. The recent lineage assigned to the July/August outbreak in Auckland (over 160 cases) was exclusively B.1.1.1. While all these viruses have the same lineage designation, it does not mean they are all 100% identical at the genome-level. Figure 2B shows the point mutations that have accumulated as the virus has spread (all are still designated B.1.1.1). These mutations hold key information relevant to contact tracing and cluster analysis.

12. Data science has the ability to clarify and inform on the underlying patterns and drivers in the context of this pandemic. However, it is clear that better explanation, clarity and transparency are necessary to help end-users make sense of what can and should be done with this data.

Why are some genomes missing from the NZ dataset?

13. During the first wave of the pandemic, genomic sequencing was not seen as a priority – this perspective has now changed. Indeed, some PCR positive samples were discarded by testing laboratories, making (retrospective) genomic analysis impossible.

    Recommendation: The genome data should be viewed as a core component of the Aotearoa New Zealand's health response (this transition is happening). There should be a directive for every positive PCR sample to undergo genome sequencing. Aotearoa New Zealand's case numbers are low enough that this should be easily to achieve both logistically and financially. MoH should look to mandate genome testing into its COVID testing strategy immediately (this is currently being actioned).

14. There are a number of underlying reasons why we might be missing genomic data from some COVID-19 positive samples. The most common scenarios are; (i) the sample had very low viral templates (i.e. high C_T values in PCR tests) meaning that ESR protocols cannot work effectively. (ii) the swab was poorly taken resulting in low number of viral template on the swab (iii) samples (swab or RNA) we not stored optimally, which results in degradation of viral genetic material.

15. While it may never be formally established, the most likely explanation for the recent Auckland outbreak is likely that it 'escaped' from managed quarantine. As only ~50% of 'managed in quarantine' (MIQ) samples were genome-sequenced, it is not possible to rule this possibility either in nor out.

    Key Recommendation: Protocols and workflows should be put in place to sequence ALL samples that test positive. This may require multiple samples be taken for every MIQ person, alternatively, rapid follow-up testing as soon as a positive PCR result is recorded. Secondary or follow-up swabs would be to obtain a variety of sample types/quality specifically with genomics in mind. There should not be a reliance on the single positive sample as storage, shipping and sample degradation will introduce gaps/weaknesses in the genomic surveillance network that severely impact our ability to infer where or when breeches might have occurred.

[9] https://www.nature.com/articles/s41564-020-0770-5

Key Recommendation: Samples that fail to yield full viral genomes can, and should, be analysed using shorter fragments of the genome that are centred around the variable sites (i.e. mutations) in the virus. This approach will ensure that, in the absence of a full genome, there is still the possibility of determining the viral strain from a small (diagnostic) section of the virus (this research capability is being developed).

Recommendation: Given that any location across Aotearoa New Zealand where a COVID-19 cluster has emerged will likely be in some form of lockdown, the transport of samples to ESR (Porirua) may become a factor that dictates the rapid turnaround of samples (for contact tracing intelligence). Alternative transport may have to be considered (road, air force or private charter) if commercial flights are not operating. Alternatively, genome sequencing protocols should be put in place at multiple centres across Aotearoa New Zealand (this is being actively explored by ESR). While there are benefits to decentralising genome sequencing, the informatics/analytics could (and I would strongly argue, should) remain centralised for consistency in quality control, analytics and reporting. A hub-and-spoke model of operation is the next logical step to build surge capacity.

What questions can be addressed by way of viral genomic analysis?

16. There are a number of possible scenarios where genomic data can help us understand the COVID-19 pandemic. However, viral genomes are not a 'magic bullet' and interpretations need to be interwoven into a wider information landscape that surrounds pandemics. In the context of the Aotearoa New Zealand outbreak the core questions genomics can address include:

(i) ***Where on the globe is the likely origin of a circulating viral lineage?*** Given the global effort to sequence and restrictions to international travel it is likely there will be a probabilistic framework around possible transmission routes. These data may inform a risk matrix when managing (or risk stratifying) quarantine placements.

(ii) ***Can a specific person be traced as the origin of a given circulating viral lineage?*** Debates into the origin of the Auckland outbreak underpin the importance of including or excluding various infection route scenarios. While not foolproof (due to false negatives when PCR testing), genomic data from all MIQ samples should offer the best chance in identifying routes of transmission. Connections that can be drawn to MIQ may help to alleviate public concern about other transmission routes, such as shipped goods from overseas.

(iii) ***Can genomes predict if a positive case has been circulating in Aotearoa New Zealand as a 'burning ember'?*** The residual risk that COVID-19 can simmer in the background, but remain undetected, remains a possibility. The lineage of virus and the patterns of mutations (when compared to genomes known to have 'landed' in Aotearoa New Zealand) provide a window to test if this scenario is likely or not.

(iv) ***Are we dealing with one incursion or more?*** The likelihood of importing two viruses with exactly identical genomes (that both escape containment) is low. Given this, there is value in quickly understanding if there is more than one incursion. The example of the 'maintenance worker' assigning to lineage B.1 and not the Auckland community strain (B.1.1.1) is an example of excluding some infection pathways.

(v) ***What information can genomes tell us in the context of contact tracing?*** By tracing viral lineages and the mutations (sub-clusters) that inevitably develop, there exists the possibility to help inform contact tracing and cluster analysis. Starting with an identical source linage, mutations will occur (see figure 2b). Once they do, it is possible to infer transmission history (albeit with resolution that is dependent on mutations that are largely random). There is the possibility that the genome might also inform when a patient might have become exposed (based on when those mutations first appeared). Because this form of mutational analysis is dependent on single changes to the genetic code care must be taken that data at these key sites is real and not from PCR amplification or sequencing. In situations where patients cannot be contact-traced to a source, the genome of the virus may include or exclude infection pathways. This intelligence can prioritise and inform leads for contact tracing teams. Ultimately, the speed by which the genomic information can be provided will underpin how useful it might be. Prioritisation of some samples over others may prove useful, for example, prioritising one genome from each household may be (in the short-term) more useful than sequencing all contacts in a single household where the transmission trail is already established under time-pressure.

(vi) ***What is the transmission risk on aeroplanes and/or airports?*** There remains an unquantified transmission risk on aeroplanes and in airports. Studying the genomes of people that are suspected to have been positive when flying (e.g. a positive day-3 test), there is the chance to study how many other passengers might have contracted that exact viral lineage. While some recent reports downplay the risks[10], a surveillance programme such as this may help to quantify the risk while travelling. Importantly, it may inform whether protective equipment and on-plane/airport prevention measures are optimal (this research is currently being considered) – in so doing, it may help to inform the public on the relative risks.

(vii) ***What can genetic/genomic testing of wastewater tell us?*** The positive detection of virus from wastewater remains a promising area to explore[11]. The successful recovery of a viral genome from a sample (a recent ESR result) provides a level of intelligence well above that of that of the original intent around wastewater testing – which was to solely provide a PCR-derived yes/no answer. Recent evidence has emerged that SARS-CoV-2 can affect the intestinal tract, even when nasal/throat swabs test negative[12]. Accordingly, genomic research on wastewater (or 'chasing' relevant mutations if RNA preservation is not sufficient) should remain a research priority. Related to this, it may be of benefit to prioritise the collection of stool samples in some of the MIQ surveillance testing, as it may help prevent gaps developing in the genomic surveillance net.

(viii) ***How is the virus evolving? Are we seeing new strains and/or recombinant viruses?*** – SARS-CoV-2 will continue to accumulate mutations. Surveillance of circulating lineages around Aotearoa New Zealand will enable a close watch on evolution of the virus. This

[10] https://www.newscientist.com/article/2252152-how-likely-are-you-to-be-infected-by-the-coronavirus-on-a-flight/

[11] https://www.sciencedirect.com/science/article/pii/S0048969720325936

[12] https://www.gastrojournal.org/action/showPdf?pii=S0016-5085%2820%2930282-1

molecular surveillance may ultimately help to inform vaccine strains, and watch for any recombination events that might occur.

17. Given the many questions where genomics can add valuable intelligence, there is a risk that the questions (listed above) become conflated with each other.

    Recommendation: Decision-makers should be explicit in which question(s) they would like addressed by genomic/genetic analyses. Clarity from stakeholders regarding primary and secondary questions will result in more targeted reporting and work prioritisation.

How should genomic data be integrated with epidemiological data - what should or shouldn't be presented?

18. While our health system is familiar with reporting positive/negative test results from genetic studies the rapid incorporation of genomic data and phylogenetic analysis is a new area that requires considered approaches, especially in how the data is graphically presented and communicated. The reporting of genomic data can, and should, be improved and there are short- and long-term solutions to this.

    Key Recommendation: A common language and consistency of data presentation is needed when exploring genomic data that acknowledges the strengths, weakness and precision of different data types. This will involve genomic experts generating text and graphics that health professionals and decision-makers can easily and accurately comprehend (often in a time-sensitive setting). Critically, reports should also convey 'time & space' as a cluster expands or is contained. It is vital that these same professionals and decision-makers attain a level of genetic literacy so they can ask the 'right questions' of the data - case studies and educational materials could help expedite this process and build capacity.

19. The genomic data presented within interim and weekly reports is factually accurate and concise. However, the staff at ESR readily conceded that there is room to improve and to tailor the reports for different audiences. The time pressure surrounding Aotearoa New Zealand's 2nd wave, understandably, saw ESR prioritise genomic data generation over data integration/presentation. However, interviews with ESR and MoH personnel demonstrated an intrinsic drive and willingness to explore options (e.g. integrated dashboards) and epidemiological maps that incorporate genomics. Anytime 'downtime' between pandemic waves should be spent carefully fostering collaboration and optimising data presentation. Such work should be seen as capacity building for *any* disease response – it is not limited solely to the COVID-19 pandemic.

    Recommendation: There needs to be an ongoing iterative process to improve presentation of genomic data alongside, or integrated with, traditional epidemiological data. Above current capacity, at least one dedicated staff member embedded at the interface of ESR/MoH that actively seeks feedback, investigates processes used in other countries, innovates and trials different ways to present the data should be a priority –

this position is needed until stability in reporting is achieved. This staff member could be charged with developing relevant case studies.

20. At the present time there are two key areas that stand out as needing attention in order to better coordinate genomics into the holistic health response - these are: (i) naming conventions and (ii) metadata:

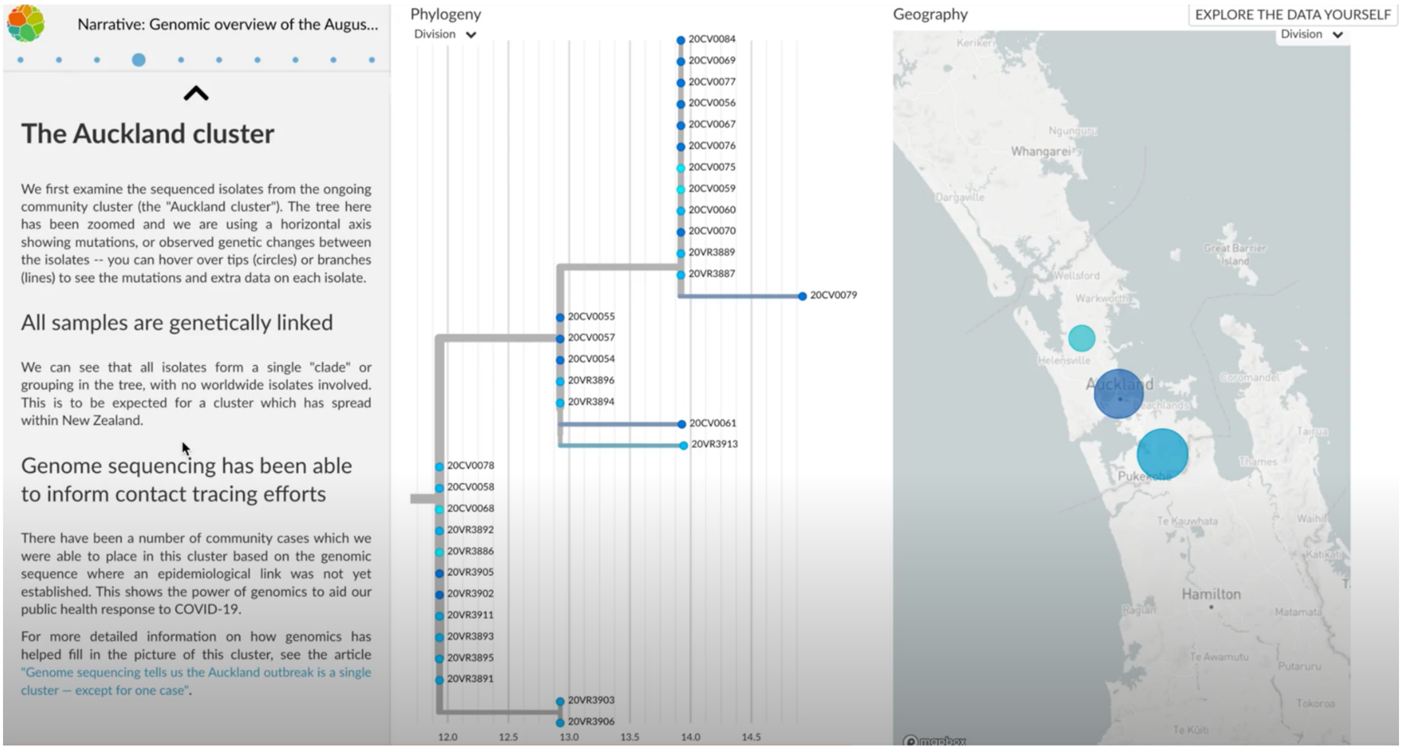
    (i) ***Naming conventions***: There is an international effort underway to harmonise the naming of SARS-CoV-2 genomes. Despite the set of guidelines that should describe how a viral lineage is named, the authoritative paper on this topic openly notes[13] "*we do not see it as exclusive to other naming systems, particularly those that are specifically intended to track lineages circulating within individual countries for which a finer scale will be helpful*". Accordingly, Aotearoa New Zealand needs to rapidly adopt a unified naming convention that encompasses mutations that occur around a core viral lineage, to be used in contact tracing and cluster analysis. Ambiguity around naming of mutations (or sub-clusters) that surround the core B.1.1.1 lineage were a point of confusion during the 2nd wave outbreak in Auckland. Work is underway to rectify this in order to aid interpretation - a naming convention has been proposed (e.g. Auckland outbreak variant 1 in the B.1.1.1 lineages would be assigned a sub-lineage AO.1.)

    (ii) ***Access to metadata***. The flow or metadata is an area that warrants attention. Against a backdrop of privacy issues (and some commercial sensitivities) the flow of information from testing laboratories, to ESR then into MoH/PHU and decision-makers has been sub-optimal. Without going into details dynamic processes need to be established to enable the correct people to see relevant data in real-time. Delays in reporting and contact tracing will be inevitable unless a smoother flow of metadata is intrinsically linked to each genome.

21. One area that may help decision-makers better comprehend how genomics, contract tracing and cluster analysis all dovetail is to recognise the severe limitation of static images. There may be a step-changes in comprehension if animations are embraced which can, to most people, better convey the spread and interconnectedness of data/patients and risk in both time and space. Reports might directly link to secure videos that show clusters grow and infections shrink and/or expand.

22. One example of live genomic data that is animated and accompanied by a narrative is Hadfield et al.'s current work, which seeks to harmonise complex genomic data with a narrative overlay. A screen capture is shown below (Figure 3). The salient point with this form of dynamic reporting is that it can updated in real-time and respond as narratives change, thus overcoming the need for paper-based reports which rapidly become outdated. This innovative work demonstrates first-hand the speed at which data visualisation solutions can be developed when collaborative networks are formed (in this case Otago University, ESR, University of Auckland and NextStrain).

[13] https://www.nature.com/articles/s41564-020-0770-5



Narrative: Genomic overview of the Augus...

The Auckland cluster

We first examine the sequenced isolates from the ongoing community cluster (the "Auckland cluster"). The tree here has been zoomed and we are using a horizontal axis showing mutations, or observed genetic changes between the isolates -- you can hover over tips (circles) or branches (lines) to see the mutations and extra data on each isolate.

All samples are genetically linked

We can see that all isolates form a single "clade" or grouping in the tree, with no worldwide isolates involved. This is to be expected for a cluster which has spread within New Zealand.

Genome sequencing has been able to inform contact tracing efforts

There have been a number of community cases which we were able to place in this cluster based on the genomic sequence where an epidemiological link was not yet established. This shows the power of genomics to aid our public health response to COVID-19.

For more detailed information on how genomics has helped fill in the picture of this cluster, see the article "Genome sequencing tells us the Auckland outbreak is a single cluster – except for one case".

Figure 3. Screen capture of a narrative-type report layout developed by Hadfield et al. An editable narrative (left) is overlayed on graphics (right) – it is envisaged this format will update 'on the fly' as new data and information becomes available.

23. Despite developments, there still exists a number of challenges when trying to present complex genomic and epidemiological data in tandem. With regard to the genomic data there is likely need for a number of reporting formats (some interim and others more comprehensive) to address the question(s) that are being asked of the analysis.

24. The representation of viral lineages (and their mutational variants) needs to be overlayed on traditional epidemiological maps (this work is underway at ESR). The reverse is also true, where clusters (or sub-clusters) can be colour-coded and overlaid on viral trees (or spanning networks). In the example below (Figure 4) the hypothetical viral linages (B.1.1.1 v1-v3) are overlaid over an epidemiological map where (close) contacts between Households, Churches and Workplaces. While 'time' is missing from this picture these graphic overlays enables the viewer to rapidly investigate possible infection pathways. For example, in this hypothetical scenario "Unknown 1" and "Household 9" might be linked to "Household 3" and "Church 1". If provided in a timely manner this type of genomic intelligence can help either include or exclude contact scenarios and even the possible direction of infection. A more nuanced execution of the map might incorporate the geographical location in the Auckland area – for example maybe "Household 9" and "Household 13" are in the same suburb. As noted throughout this report, these connections can be drawn manually but could be automated and updated in real-time with data science solutions.

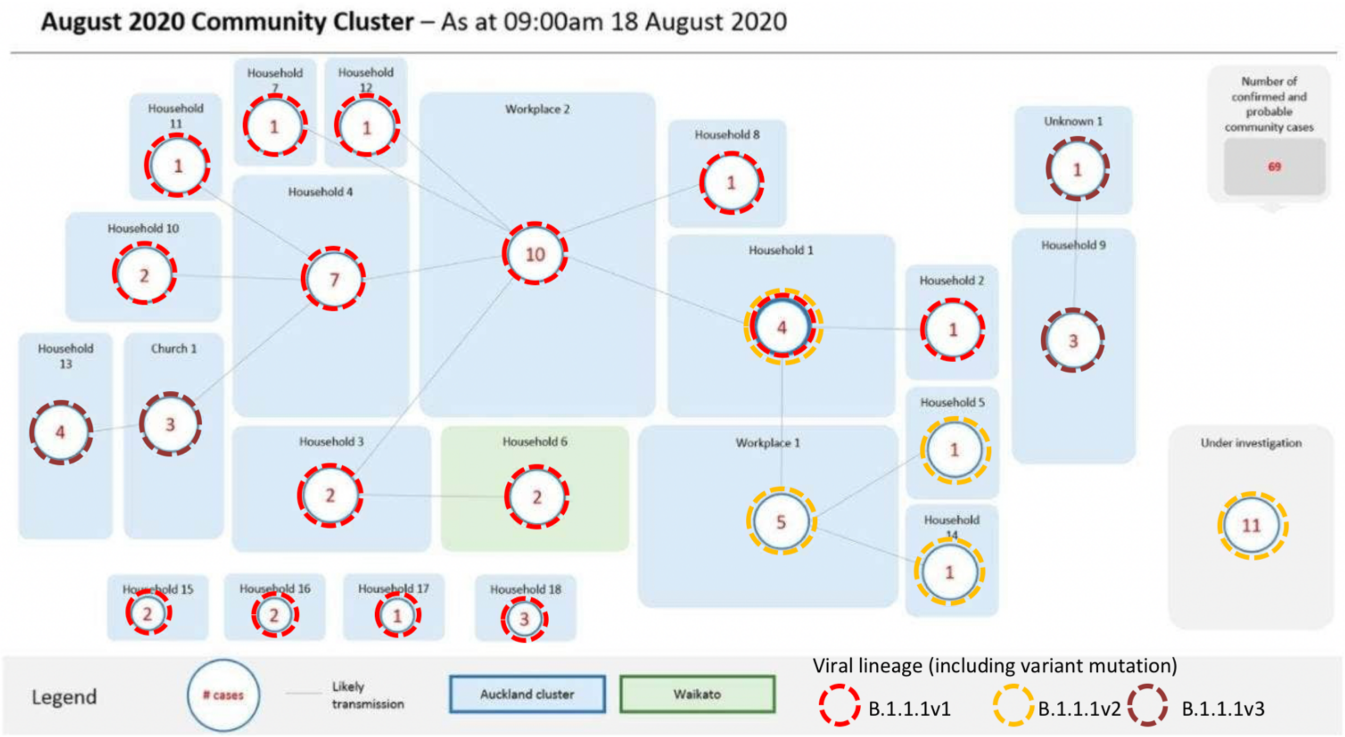


Figure 4: Scenario where, for illustration purposes, hypothetical genomic data is overlaid on an epidemiological map of the Auckland Cluster on 18th August 2020.
Source: https://www.stuff.co.nz/national/health/coronavirus/300086391/coronavirus-mapping-aucklands-new-cluster-now-one-of-the-countrys-biggest

25. ESR scientists have been rapidly developing an alternative visualisation called Vis-Network. The 'R'-based application draws in data from Epi-Serve and MicroReact to model and visualise a cluster map. In Vis-Network bubble diagrams are formed that can link cases by genomic variant and contacts. Importantly, this application is able to (where data allows) infer direction of infection and infection hierarchies. The interface will become important in that it will enable users to the exploration data and filter/sort data according to various metadata or timelines.

26. With the view to providing rapid high-level overviews of the overall status of clusters or regional outbreaks, ESR/MoH might consider adopting some visualisation tool(s) that attempt to aggregate the overall status of clusters, showing strengths and weaknesses of the overall response status. One option to aggregate these data are to present them in a 'snowflake' diagram where the strength of a response could be ranked on a multi-point scale (e.g. where a grade of '1' represents a response that is comprehensive and/or complete and where a grade of '5' might represents a response that is poor/preliminary or incomplete). While there are limitations to this form of data presentation, it has the benefit of seeing an overall status of the various 'arms' of the pandemic response and provide a green, amber or red status if any one of the response arms is struggling. A mock-up to explain this concept is included below (Figure 5).

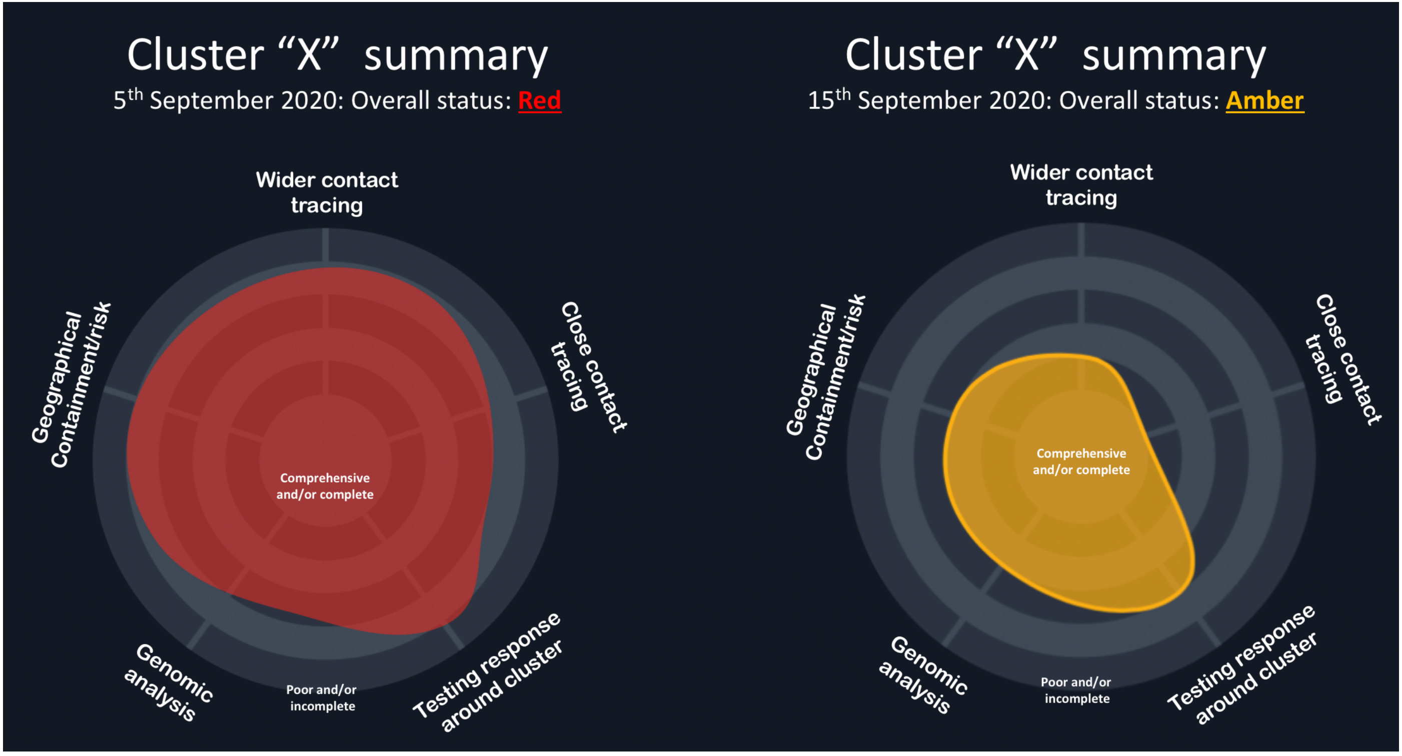


Figure 5. Hypothetical concept figure to represent a high-level overview of the various arms of the pandemic response. The axes of this figure are for demonstration purposes only and need careful consideration on what best to represent and how these data sit on a simplified multi-point scale. The visualisations may wish to consider placing a series of 'snowflakes' from different timepoints to track progress on cluster management. An overall status of; Red, Amber or Green may help to convey the overall level of cluster management.

What is needed in the short/medium/long term to develop integrated data science solutions?

27. Hard problems require innovative solutions and the COVID-19 pandemic is an example where data-science limitations and visualisations have come to the forefront. While it is possible to continue approaches that manually curate genomic and epidemiological data, the more strategic response is to explore 'at pace' how data and predictive models can operate in real-time, and in ways that enhance rapid decision-making under pandemic scenarios.

> ***Key Recommendation:*** Investment and collaboration is needed to develop bioinformatics and data science solutions (software and databases) for real-time reporting. While manual curation of data will be required in the short term, this cannot be conducted at pace or at scale, nor does it position Aotearoa New Zealand well for future outbreaks/pandemics. There are options available within Aotearoa New Zealand to 'prime' this area.

28. After a rapid evaluation of the data science landscape across Aotearoa New Zealand the best existing initiative (in terms of capability for responding to the data science challenges posed by

both genomics and epidemiological data) I discovered was the MBIE-funded SSIF "Beyond Prediction" collaboration (University of Auckland). A core theme of this e-science initiative is:

"*When all the steps in the process of analysis -- from data discovery to application -- are made transparent, auditable and reactive to change (via continuous automated integration of new data, models and methods) we close the gap between doing research and its effective and timely communication.*" (M Gahegan *et al.* 2019).

29. In my opinion, academic capacity and innovation is required to tackle the underlying data-science (and modelling) challenges that COVID-19 has unexpectedly placed on our doorstop. With regard to genomics, initiatives such as NextStrain and Microreact have rapidly been adopted. However, Aotearoa New Zealand can, and I would argue should, seek to develop its own bespoke live-data integration platforms that integrate the subtleties of the pandemic response within Aotearoa New Zealand. The first and second pandemic waves to reach Aotearoa New Zealand provide an ideal time-stamped dataset to retrodict how our responsiveness could have been enhanced.

> ***Recommendation:*** Given the lag-time associated with forming collaborations and capacity, the existing SSIF-funded 'Beyond Prediction' collaboration is in an excellent position to respond to the challenges presented by the genomics and epidemiological data streams. The HRC-funded project 'Predict and Prevent' initiative (which overlaps in personnel with 'Beyond Prediction' is also rapidly developing capacity in this area. The 'Collaboratory' model is well positioned to respond to the COVID-19 data problems, as it already has the expertise in data science, coding, visualisations, genomics and modelling. Allocating resource of a COVID-19 "module" to this existing SSIF or MRF-funded programmes of work would likely yield tangible benefits within the next 3-12 months. Researchers within such consortiums have already rapidly published innovative data-science solutions modelling COVID-19 pandemic scenarios[14] (live applications for funding into various COVID-19 related workstreams may already be in progress). It is my view that the innovation and software in this area will come primarily from the academic sector that should seek to draw key skills from CRI's and public health experts.

Engagement & education strategies for genetics and genomics

30. The high public and political interest in COVID-19 means there is a need to develop tools for; education, communication and engagement.

31. Our public heath responders, testing facilities and decision-makers must develop a better appreciation for the role genetics/genomics in a phylogenetic context can play in our pandemic response and preparedness. The write-up of case studies from clusters that explain how genomics is integrated into traditional epidemiological work is needed to better support this work. The work of Souxsie Wiles & Toby Morris is an excellent example of this (e.g. see[15]).

[14] https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12660

[15] https://sciblogs.co.nz/infectious-thoughts/2020/08/13/how-genome-sequencing-could-crack-the-case-of-the-nz-covid-comeback/

> ***Key Recommendation:*** Resources such as video, graphics, animations and webinars should be developed to explain the science behind COVID-19 genomics and how it is being implemented at our front lines. This should build on the foundational communication work by ESR and University of Otago. Visual tools must accurately represent the data and be cognisant of cultural and privacy concerns. Given the public interest in this science there is an opportunity to explain and educate about the wider benefits of genomic testing/technologies including how vaccines are made and tested[15].

32. The hesitancy with which the Aotearoa New Zealand public can view some genetic modification technologies (e.g. a vaccine that might be generated by genetic modification) may come to the fore during this pandemic. I would advocate that any opportunity to promote accurate knowledge and information in this area may help to socialise the underpinning science of vaccines to enable the public to make informed choices.

Horizon scanning – the long-termism needed in studying disease genomics.

33. While studying the genomes of SARS-CoV-2 is proving to be of value in managing outbreaks and clusters – it is important to consider the longer-term benefits of genomics strategies for the identification and control of disease.

34. The novelty of the SARS-CoV-2 virus in human hosts means the virus has not experienced significant genomic selection – the ~2000 or so variations observed have been largely 'random' and are scatted across the genome. This will change when vaccines and medicines start to select for viral lineages.

> ***Key Recommendation:*** Aotearoa New Zealand needs to embrace genomic tools and analyses for long-term monitoring of viral evolution. This is not simply an academic exercise, rather there is a pressing need to monitor the viral lineages that are circulating (akin to seasonal influenza tracking). This is vitally important once a vaccine is deployed in Aotearoa New Zealand and border restrictions are relaxed. Genetic characterisation of circulating lineages needs to become routine practice. Capacity that is built in this area can easily be adapted to other disease outbreaks in people as well as animals (e.g. *mycobacterium bovis*).

35. The rapid advances in genome sequencing capabilities in the last decade mean that it is fast becoming best-practice to sequence genomes of viruses and bacterial as part of routine clinical practice. Investment and training in technicians, laboratory infrastructure, data scientists and modellers will become increasingly useful in the coming decade. In much the same way as some infrastructure projects are 'shovel-ready', there are many genetics/genomics science projects that are 'pipette-ready' [16]. Indeed, Aotearoa New Zealand might wish to revisit the 'One Health' framework that seeks to remove traditional barriers that have siloed medical, animal and plant disease.

[16] https://www.stuff.co.nz/science/300098921/pandemic-recovery-means-big-money-for-shovelready-projects-but-what-about-science

Final Comments:

36. In completing this rapid review, I conclude Aotearoa New Zealand's genomics response to the COVID-19 has been world-leading. However, there remains room for improvement. The rapid deployment of diagnostic PCR assays across the country was backed up by a group of committed scientists that appreciated, early in the piece, the role that genomics would play in the pandemic response, and advocated for funding and operational relevance. The data they have generated has notably enriched the public health response.

37. With the value of hindsight, there are aspects the genomic researchers could have done better during the 2nd wave pandemic in Auckland. This is acknowledged, and is to be expected given that this is the first-time genomic data has been integrated into epidemiological operations in real-time. Importantly, there is a willingness to adapt, modify and innovate in order to better prepare for any future challenges.

38. The genomics 'arm' of Aotearoa New Zealand's pandemic responses will need support in the next phase of deployment as it builds depth (multiple sites), surge capacity, bioinformatics expertise and data visualisation tools. Attention needs to be paid to unifying public response across public heath units, CRI's, Universities and commercial/public testing laboratories. In the interim this collaboration is acting as a virtual 'CDC' equivalent for Aotearoa New Zealand.

39. This rapid-review was conducted 'at pace' to generate a point-in-time appraisal of Aotearoa New Zealand's genomics capacity in response to COVID-19. Some of the recommendations in this report are already being enacted, others should be considered.

40. Finally, I am aware that the WHO is set to shortly release guidelines on genomic testing. These guidelines, in addition to Aotearoa New Zealand's own learnings for the COVID-19 genomics work (to date), should form the basis of a revised national COVID-19 testing strategy. In the longer term, this strategy should be formally integrated into Aotearoa New Zealand's pandemic plans.

Acknowledgments:

I would like to acknowledge a large number of individuals that gave up their time to share knowledge and ideas. Every conversation I had filled me with a sense of collaboration and willingness to make a difference in the COVID-19 battle – this is science done at pace – it may be "its finest hour" in the fight to control this pandemic across Aotearoa. I have tried to interweave multiple perspectives, thoughts and ideas into this rapid review but fully acknowledge that the synthesis view expressed in this document are my own and may not be endorsed by any of those to whom I interviewed. Thanks to; Eddie Holmes, Jemma Geoghegan, Joep de Ligt, Lisa Oakley, Ben Waite, ESR laboratory staff, Alexei Drummond, Mark Gahegan, Ryan Lister, Colin Simpson, Marie Estcourt, Niki Stefanogiannis, Jill Sherwood, Andrea McNeill, Hamish Spencer, Gary Evans, James Ussher, David Murdoch, Peter Dearden, James Hadfield, Hannah Davidson, George Slim, Juliet Gerrard and Ian Town. Finally, I'd like to thank my employer (the Environmental Protection Authority) for making time available to conduct this rapid review.

APPROVED FOR PUBLIC RELEASE 21st SEPTEMBER 2020

Glossary of key terms and acronyms as they relate to COVID-19, PCR and genomics:

COVID-19: The name given to the disease pandemic caused by the SARS-CoV-2 virus that originated in the Wuhan Provence in China in late 2019.

C_T or C_q: these are relative values generated during qPCR. These metrics (typically a number between 1-40) are used to tell the operator how much viral RNA is present. Somewhat counterintuitively the higher the number the less viral RNA is present.

False positive: When a test for COVID-19 comes up as positive when in fact the patient is not carrying the virus or has been previously exposed.

False negative: When a test for COVID-19 comes up as negative when in fact the patient is carrying virus.

Gene: A region of the SARS-CoV-2 virus that codes for a part of the virus's protein core.

Genome: the genetic material carried by an organism or cell. In the case of SARS-CoV-2 this is a single stranded RNA molecule comprised of nearly 30,000 bases.

Library: the laboratory process by which the virus is converted into labelled pieces of DNA that are ready to be sequenced.

Illumina: A DNA sequencing platform that is able to determine the genome of the SARS-CoV-2 virus

Limit of Detection (LOD): The limit of detection in the context of COVID-19 PCR refers to the lowest number of viral copies that can be detected by the test. Different PCR tests will exhibit different LOD's.

Lineage: Term given to a distinct group of the SARS-CoV-2 viruses that differ significantly from other branches of the viral tree. In September 2020 there were ~80 distinct lineages that had been defined.

Mutation(s): The process by which the underlying genetic code of the virus changes. Mutation can be random or it can occur at key spots in the genome. Mutation in viruses is heavily influenced by external drivers such as medicines and vaccines which place pressure on the virus to 'escape'.

Nanopore: A DNA sequencing platform that is able to determine the genome of the SARS-CoV-2 virus

NGS: next generation sequencing – a powerful platform whereby DNA or RNA is rapidly sequenced to determine the underlying genetic code.

PCR: 'Polymerase Chain Reaction'; a test by which DNA (or RNA) is photocopied. This is the core test for COVID-19 virus as it is very sensitive and can be rapidly implemented.

Phylogeny. Viral genomes are assembled into a 'tree' (or phylogeny) that is able to represent the history of the virus as it mutates and/or evolves.

Phylogenetic analyses: The analyses where viral genomes are modelled to better understand their evolutionary history and trajectory.

Primer: Small pieces of DNA (also called oligos) that are used to target a viral gene during PCR – the sequence of the primers provides the key specificity of the PCR test.

Probe: A small piece of DNA with a fluorescent dye on one end that is used to measure the accumulation of viral template within a PCR. Probes can also increase specificity of the PCR.

qPCR: quantitative PCR, refers to PCR that is tracked in 'real time' via the fluorescent probes that bind to new copies of the viral targets. qPCR is very commonly employed when testing for infectious diseases.

qRT-PCR: quantitative reverse-transcriptase PCR. A form of PCR that first converts viral RNA to DNA then amplifies viral genes using qPCR. COVID-19 testing typically tracks two viral genes in tandem.

Recombination: the process by which closely related viruses might 'shuffle' their genetic material to constitute a new viral particle.

Reverse transcriptase. An enzyme that converts RNA to DNA, it is a step used in during PCR testing. It is vital as the COVID-19 virus has an RNA genome so conversion to DNA is required when testing using PCR.

RNA: A form of genetic material that the COVID-19 virus uses.

RNA extraction kit: A kit purchased commercially that contains all the 'ingredients' needed to isolate viral RNA from human samples for use in PCR tests.

RNAseq: a technique by which all the RNA in a sample is sequenced, such approaches have been used in viral discovery and diagnostics.

SARS-CoV-2: The virus that is responsible for the COVID-19 pandemic.

Sensitivity (analytical): Refers the ability of a COVID-19 test to pick up traces of the virus under laboratory conditions. PCR tests are not all created with equal sensitivity and can have different Limits of Detection (LOD).

Sensitivity (Clinical): Refers to the ability of a COVID-19 test to detect virus in actual patient samples following infection. Early in an infection patients may have low levels of the virus and thus test as negative, likewise swabbing technique may not be optimal and result in negative tests.

Virion(s): The scientific name given to describe virus particles.

WGS: acronym for Whole Genome Sequencing which is equivalent to determining the viral genome.