Enhancing public health surveillance: integrating genomic and epidemiologic data to inform public health action and One Health progress

Hanna Oltean-Parke

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Janet Baseman, Chair

Trevor Bedford

Scott Lindquist

Program Authorized to Offer Degree:

Epidemiology

University of Washington

Abstract

Enhancing public health surveillance: integrating genomic and epidemiologic data to inform public health action and One Health progress

Hanna Oltean-Parke

Chair of the Supervisory Committee:

Janet Baseman

Department of Epidemiology

Pathogen genomic data can provide highly useful information for public health practice, particularly when combined and analyzed with epidemiologic data in real time. Likewise, a One Health approach pushes our current health surveillance systems beyond their siloed views to consider balancing and optimizing health outcomes across human, animal, and environmental domains. Implementation of genomic epidemiology in public health practice alongside a One Health approach holds promise for early and more specific outbreak detection, improved understanding of health risks, increased hypothesis generation for research, and proactive public health action to prevent health threats.

This dissertation focuses on genomic data integration and use within public health practice, highlighting the systems changes required for successful implementation, demonstrating population-level genomic-epidemiologic analyses for the purpose of public health action, and discussing expansion of these concepts to encompass a One Health approach. In the chapters that follow, I first describe implementation of a comprehensive system for large-scale genomic data capture and linkage to epidemiological data and an evaluation of this system. This study

identifies key areas of success for this system as well as areas for improvement to enable real-time genomic-epidemiologic analyses. Next, I apply genomic-epidemiologic methods, demonstrating the utility of genomic data produced at the population-level to add information for public health action over the course of the SARS-CoV-2 pandemic. Given the available data, computing infrastructure, workforce, and tools, I outline which genomic-epidemiologic methods are most applicable for ongoing or routine data analysis given the system's current state, as well as recommendations for improved data capture to support additional methods. Finally, I outline requirements for operationalizing One Health data integration through the development of a framework and possible approaches to One Health genomic data storage and co-analysis. This framework is developed to support data integration across One Health domains, expanding our joint ability to prevent and control disease. Together, this work envisions a more holistic approach to infectious disease surveillance, considering data generated from pathogens, hosts of all species, and the environment to better prepare our public health system to face emerging and endemic health threats.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible without the contributions and support of many individuals, including each of my co-authors. I thank each of my committee members: Janet Baseman, for your support with project development from the very beginning, your ongoing encouragement, and unfailing optimism that I would finish this undertaking; Scott Lindquist for your positivity and assurances of the importance of this work; Trevor Bedford for your expertise, constructive feedback, and support through the challenge of learning genomic epidemiology techniques; and Jim Hughes for your ever-timely feedback and helpful review. I thank the following colleagues for your critical support through these projects: Krisandra Allen, Alli Black, and Beth Lipton. Without each of your intellectual contributions, these projects would be greatly diminished. To my team and colleagues at the Department of Health – you have given me space and flexibility to pursue this work and encouraged me through the hard times. Our efforts together during an unprecedented global pandemic, including those detailed herein, were truly career-defining. And finally, to the family that made everything possible: to David, for your unfailing support and patience through this long effort, as well as your Word formatting and code troubleshooting skills, to Eric for your patience as a Python teacher, and to Levi for opening my eyes to new curiosity in this world.

Chapter 1. INTRODUCTION

1.1 *Public Health Surveillance Implementation and Evaluation*

Public health surveillance is the ongoing and systematic collection of samples or data, data storage, analysis, interpretation, and dissemination or outcome communication regarding a health event for the purpose of public health action[1]. Surveillance systems vary widely in their scope, objectives, methods, and platforms available for data storage, linkage, and analysis. Public health information system complexity ranges from paper data collection to sophisticated systems with standardized reporting. The need for increased informatics capacity and data modernization across public health is clear. Gaps in public health informatics capacity can be identified when attempting to make use of new data types or structures, such as pathogen genomic data. Pathogen genomic data can provide highly useful information for public health practice, particularly when combined and analyzed with epidemiologic data in real time. Pathogen genomic data incorporated into public health surveillance systems can allow for earlier detection and more precise investigation of outbreaks[2]. Characterization of microbes through sequencing provides insights into transmission dynamics and evolution[2]. However, the incorporation of sequence data into public health surveillance systems presents challenges[3]. First, capacity for sequence generation and a workforce with training in bioinformatics are required to generate and assemble genomes and perform quality checks on the data. Second, sequence data storage requires intensive computing infrastructure such as cloud computing, which may not be readily available to public health systems. Third, methods for linking pathogen sequence data to traditional epidemiological data are needed, requiring either co-storage of data or identifiers available to link between systems. And finally, a workforce trained in genomic epidemiology and

bioinformatics methods is needed to make use of paired genomic and epidemiological data. To arrive at a surveillance system that incorporates each of these elements in a way that allows for real-time data analysis is a major challenge facing the public health practice sector.

Despite these barriers in the development of integrated systems with capacity for joint epidemiologic and genomic analyses, the production of such systems has the potential to improve disease prevention and control efforts. Goals of establishing genomic surveillance can include: monitoring circulating and emerging variants, detecting and characterizing outbreaks, describing spatiotemporal patterns of pathogen transmission, supporting epidemiological and genomic characterization of pathogens, and pinpointing sources that may be identified as risk factors[4]. Established systems can translate data into public health interventions to prevent disease, control spread, and mitigate outbreaks. These interventions could include preparedness planning based on emerging variant characteristics, changing use of therapeutics and non-pharmaceutical interventions, and targeted recommendations for control based on outbreak or source characteristics.

Implementation of new surveillance systems requires decisions around target population and methods of active versus passive data collection. Attributes of surveillance systems that are considered in design and evaluation include the data quality, data timeliness, system flexibility, system simplicity, stability, sensitivity, predictive value positive, representativeness, and acceptability[1]. In general, the attributes of the system should be balanced with the system's objectives, considering those that are the highest priority. When considering implementation of genomic-epidemiologic surveillance, the data quality, timeliness, and representativeness are priorities to support analysis and interpretation of findings.

1.2 *Genomic Epidemiology in Public Health Practice*

Methods for conducting genomic epidemiology analyses are broad, varied in complexity, and increasingly expanding. Two of the most widely used methods underpinning genomic-epidemiologic analysis are maximum likelihood-based methods (ML), which relies on frequentist statistics, and Bayesian phylogenetic inference. Selection of the appropriate method requires an understanding of the data collection methods, whether prior knowledge of the analysis in question is available, limitations in computational resources, and desired inferences. Specifically, ML methods may produce biased results in situations where data is sparsely sampled, or where sampling bias is a concern[5]. While Bayesian methods are more robust to sampling, informative prior knowledge to allow predictions on sampled data is required; use of unreliable priors can lead to biased models. A background understanding of available models and model parameterization is needed. Additionally, Bayesian methods often require extensive computational resources to complete the iterative analyses required. Bayesian methods for phylogenetic analysis generate a distribution of trees, which can more effectively capture uncertainty. In contrast, ML methods generate a single most-likely tree, which does not allow for robust assessment of uncertainty.

The most basic genomic-epidemiologic analysis is phylogenetic tree visualization, overlaying epidemiologic data. Phylogenetic visualization of samples of interest can be accomplished through phylogenetic placements or through phylogenetic tree inference[6]. Commonly used tools for phylogenetic placement, which involves placing samples of interest onto a previously-inferred tree, include UShER and Nextclade[7,8]. These tools are limited to a set of viral pathogens for which previously inferred trees are maintained, currently including SARS-CoV-2, monkeypox virus, influenza A and B viruses, and RSV. Alternatively, phylogenetic trees can be

constructed using ML or Bayesian methods, as well as less-recommended distance matrix methods such as unweighted pair group method with arithmetic mean (UPGMA). A common tool for ML tree inference is Nextstrain Augur (an analysis pipeline which runs IQ-TREE by default or RAxML or FASTTREE if specified); alternatively, a custom analysis pipeline can be developed using these ML tools or others[9]. A common tool for Bayesian tree inference is BEAST/BEAST2, an analysis tool used in conjunction with additional programs to format data, produce a summary tree from the distribution of trees, reduce sampling frequency, summarize results, and view trees[10]. The final tree file from any of these placement or inference programs can then be visualized and overlayed with epidemiological data (such as demographic information, linkages to people or locations, clinical data, etc.). Common tools for overlaying epidemiological data onto phylogenetic trees include Nextstrain Auspice, Microbetrace, and MicroReact[9,11,12].

Beyond phylogenetic tree inference and visualization, additional methods employed in genomic epidemiology commonly include phylogeographic analysis, transmission tree reconstruction, and cluster detection, among others. Phylogeographic analyses aim to infer migration trends, transmission dynamics, and the history of sampled lineages from genomic data[5]. One method for phylogeographic analysis is discrete trait analysis (DTA), which treats migration of lineages between locations as if the location were a discrete trait. TreeTime is a common tool for DTA analysis, which can be specified within Nextstrain Augur[13]. This method is computationally efficient but also sensitive to sampling bias[5]. Alternatively, Bayesian Structured Coalescent Approximation (BASTA), implemented in BEAST2, is proposed as an analysis method to address model problems identified with DTA[5]. Transmission tree reconstruction aims to reconstruct individual transmission events in a disease outbreak, informing risk factor analysis,

infection prevention breaches, and allowing for evaluation of control measures[14]. Transmission tree reconstruction methods rely on within- and between-host genetic diversity to resolve likely transmission events; this reconstruction suffers when diversity has not accumulated and genomes are identical, as may often be the case in rapidly spreading viral outbreaks with slow mutation rates. Common tools include TransPhylo, which assesses genetic data alone, and outbreaker2, which utilizes both genomic and epidemiologic data in reconstruction[14,15]. Finally, many cluster detection methods have been developed to use either sequence data alone or paired genomic and epidemiologic data to group cases into clusters, either using distance-based thresholds or threshold-free methods. These methods can replace or supplement traditional space-time scan statistics commonly employed in public health surveillance.

Genomic epidemiologic methods can be applied to a single case (to understand possible sources of infection or determine linkage to a known cluster), suspected outbreaks (to elucidate sources, transmission dynamics, or to differentiate non-outbreak cases), or on a broader scale (to understand transmission dynamics, geographic spread, identify clusters, or other population-level questions). Most examples of genomic epidemiology studies in the literature constitute retrospective analyses of outbreak-level data. Moving toward population-level analysis and real-time data consumption for public health purposes will require improved surveillance systems for timely capture and linkage of data as described above, improved computing infrastructure to allow selection of the methods most appropriate for the question at hand, a workforce trained in the concepts and theory underpinning frequentist and Bayesian analysis methods, and collaboration between surveillance epidemiologists with understanding of data collection and sampling partnering with genomic epidemiologists to select and carry out appropriate studies.

1.3 *One Health*

The use of pathogen genomic sequencing and analyses in support of infectious disease surveillance holds promise not just for traditional communicable disease surveillance or emerging disease surveillance, but also for the field of One Health. One Health is an integrated, unifying approach to balance and optimize the health of people, animals, and the environment[16]. This approach pushes our current health surveillance systems beyond their siloed approaches to coordination and integration across the surveillance pathway. Pathogen genomic sequencing is host-agnostic, and phylogenetic analysis of resulting data allows for assessment of transmission dynamics at the human-animal-environment interface. This technology can be applied across bacterial, viral, fungal, and parasitic pathogens. A One Health implementation allows for early outbreak detection and improved understanding of pathogen reservoirs, evolution, and vehicles of transmission, enabling proactive prevention of One health threats[17].

However, the challenges in surveillance system implementation outlined in section 1.1, including need for increased informatics capacity and data modernization, are amplified when considering the multiple sectors and corresponding agencies that comprise One Health domains. Moving from a single-sector surveillance system to a One Health surveillance system requires intensive coordination and collaboration. Like implementation of pathogen genomic surveillance, One Health surveillance has potential to improve prevention and control efforts[18]. Effective integration across domains to develop One Health surveillance systems, including for pathogen genomic data, requires implementation of modern technologies and database infrastructures, including application programming interfaces (APIs), artificial intelligence (AI), machine learning (ML), and alternative data systems[3,19].

1.4 *About this dissertation*

This dissertation focuses on genomic data integration and use within public health practice, highlighting the systems changes required for successful implementation, demonstrating population-level genomic-epidemiologic analysis for the purpose of public health action, and discussing expansion of these concepts to encompass a One Health approach. In Chapter 2, we describe implementation of the first comprehensive system developed at the Washington State Department of Health for large-scale genomic data capture and linkage to epidemiological data and an evaluation of this system. This study identifies key areas of success for this system as well as areas for improvement to enable the types of real-time genomic epidemiologic analyses described above. This chapter highlights the importance of understanding sampling and quantifying sampling bias in phylogenetic studies, as well as ongoing evaluation and improvement of public health surveillance systems. Notably, by performing this evaluation, we support understanding the population of sampled cases and limitations on inference affecting the use of this genomic data, setting up for the genomic epidemiologic analyses preformed and described in Chapter 3.

In Chapter 3, we use the above-described genomic-epidemiologic methods to assess the utility of genomic data produced for long-term care facility-associated (LTCF) cases to add information for public health action over the course of the SARS-CoV-2 pandemic. We use population-level data to assess transmission dynamics within and between LTCFs in a setting of changing guidance and policy. Given the available data, computing infrastructure, workforce, and tools, we outline which genomic-epidemiologic methods are most applicable for ongoing or routine data analysis, as well as recommendations for improved data capture to support additional methods.

Finally, in Chapter 4, the lens widens beyond traditional public health practice to consider data integration, including consideration of pathogen genomic data, across One Health sectors. Use of an integrated One Health approach to genomic epidemiology has been most commonly applied in the area of food-borne disease, with the collection of genomic and epidemiological data from human, veterinary, food, and environmental domains in systems such as PulseNet, GenomeTrakr, and NCBI in the United States and the EFSA One Health WGS System in the European Union[20,21]. Expansion of a One Health approach to integrated genomic surveillance has not yet been widely extended to zoonotic or vector-borne disease pathogens, despite clear risk for impact of these pathogens on a global scale and clear benefits for understanding transmission at the human-animal-environmental interface. In this chapter, we explore requirements for operationalizing One Health data integration through the development of a framework and understanding possible approaches to One Health genomic data storage and co-analysis.

In sum, this dissertation represents the evolution of incorporating genomic data and analysis into epidemiology practice, from the initial conceptualization of a new surveillance system during a global pandemic to future-thinking about expansion to One Health domains. While much work and resourcing are still required to develop infrastructure, build a capable workforce, and implement genomic epidemiology as a core component of communicable disease surveillance across all pathogens, the work represented herein moves our public health system forward toward the future of integrated surveillance. As we move toward the improved precision promised through genomic epidemiology and the holistic solutioning available through a One Health approach, the public health system will increasingly be able to effectively and efficiently address present and emerging threats, improving health for all.

Chapter 2. IMPLEMENTATION AND EVALUATION OF A SENTINEL SURVEILLANCE SYSTEM FOR SARS-COV-2 GENOMIC DATA: WASHINGTON STATE, 2020-2021

2.1 *Background*

Genomic data can provide highly useful information for public health practice, particularly when combined with epidemiologic data in real time. Goals of establishing genomic surveillance can include: monitoring circulating and emerging variants, detecting and characterizing outbreaks, describing spatio-temporal patterns of viral transmission, supporting epidemiological and genomic characterization of variants, and pinpointing introduction sources that may be identified as risk factors[4]. Information from a paired genomic and epidemiologic surveillance system can then be translated into public health interventions to prevent disease, control spread, and mitigate outbreaks. These interventions could include preparedness planning based on emerging variant characteristics, changing use of therapeutics and non-pharmaceutical interventions, and targeted recommendations for control based on outbreak characteristics. To ensure generalizability and equity when using paired genomic and epidemiologic data for public health purposes, the methods for capturing this data must ensure a representative sample from the underlying population of interest[1,22].

Ongoing high circulation of SARS-CoV-2 globally and repeated emergence of new variants indicate the need for robust genomic surveillance to inform public health response[23]. In Washington State, surveillance for SARS-CoV-2 is passive, and therefore focused on cases of COVID-19 disease in persons seeking testing. Additionally, the currently available methods for conducting next-generation sequencing (NGS) introduce limitations on sampling; namely,

specimens must contain viral RNA present at high enough quantities for sequencing efforts to be successful. Therefore, persons with mild illness, delayed test-seeking, reinfection or other characteristics that may result in lower viral load, are less likely to be represented in sequencing data. Knowing these limitations, the Washington State Department of Health (WA DOH) sought to establish a genomic sentinel surveillance system for SARS-CoV-2 in March 2021.

Prior to the initiation of sentinel surveillance, large amounts of genomic data were produced by academic and clinical laboratories in Washington and shared publicly via the GISAID EpiCoV database[24-26]. Studies utilizing these data to rapidly produce critical viral transmission and evolution information were published early in the course of the pandemic; however, the underlying population captured is unknown[27-31]. Sampling bias, or systematic differences in sample characteristics between sequenced and the total population of COVID-19 cases, is a concern. The use of large datasets from a limited number of geographically sparse institutions threatens to produce inaccurate phylogenetic representations of the distribution and migration of the virus in the population[32,33]. Specifically, models relying on discrete traits analysis (DTA), a type of phylogeographic analysis that treats migration of lineages between locations as if the location were a discrete trait, assume sample sizes across subpopulations are proportional to their relative size, with random sampling occurring in these populations[5]. If one population is oversampled compared to another, large biases are expected in model output[5]. This concern extends beyond state or country borders, as representative sampling is often assumed for contextual data, which provides the backdrop upon which phylogenetic inference is based.

Herein, we describe implementation of a sentinel surveillance system to enable pairing of genomic and epidemiologic data. Additionally, we assess representativeness and timeliness of genomic data generated before and after system implementation. By performing this evaluation,

we support understanding the population of sampled cases and limitations on inference affecting the use of this genomic data. We identify population subgroups that may be systematically excluded from sequencing surveillance to support planning efforts to obtain a more equitable and representative sample. More broadly, our description and analysis raise awareness regarding sampling bias in convenience-based genomic surveillance systems and support development of robust genomic surveillance systems in additional jurisdictions.

2.2 *Methods*

2.2.1 *Sentinel Surveillance System Design*

In March 2021, WA DOH partnered with multiple laboratories to establish a sentinel surveillance program, monitoring the genomic epidemiology of SARS-CoV-2 in the state. Partner laboratories were selected to maximize geographic coverage and specimen numbers. The initial proportion of randomly selected positive specimens submitted for sequencing was designed to balance geographic coverage regionally and match available sequencing capacity; statewide case coverage varied from 8% to 25% during the study period[34]. In addition to the Washington State Public Health Laboratories, the six sentinel laboratories include: Atlas Genomics, Confluence Health/Central Washington Hospital, Interpath Laboratories, Incyte Diagnostics Spokane, Northwest Laboratories, and University of Washington Virology. Cycle threshold (C_t) value is capped at 30 for this surveillance system. This program is supplemented by a national surveillance effort supported by the Centers for Disease Control and Prevention which includes multiple commercial laboratories sequencing randomly selected specimens. Methods for NGS vary across laboratories, but >90% are generated using an Illumina platform; assembly methods also vary.

2.2.2 *Study Population for Evaluation*

All confirmed COVID-19 cases (detection of SARS-CoV-2 RNA by molecular amplification) reported among Washington residents between January 21, 2020 and December 31, 2021 in the Washington Disease Reporting System (WDRS) as of January 31, 2022 were included. Sequences uploaded to the GISAID EpiCoV database between January 21, 2020 and January 31, 2022 and indicating Washington State in their geographic tag were linked to these cases using laboratory accession numbers or patient demographics. Cases were classified as pre-sentinel surveillance if they had specimens sequenced prior to March 1, 2021. Cases were classified as sentinel surveillance if they had specimens sequenced on or after March 1, 2021 and submitted through the WA DOH sentinel surveillance program or if the sequencing laboratory indicated that specimens were randomly selected. Specimens specifically selected for targeted sequencing as part of outbreak investigations, due to travel history, known vaccine breakthrough status, or S-gene target failures were not considered sentinel surveillance if sampled outside of the random selection process. The Washington State and University of Washington Institutional Review Boards determined this project to be surveillance activity and exempt from review.

2.2.3 *Data Analysis*

Representativeness of data pre- and post-implementation of sentinel surveillance was assessed by comparing cases with sequencing performed to all cases during the same time-period based on the following characteristics: sex, age, race, ethnicity, language, long-term care facility (LTCF) association, occupation, county of residence, outbreak association, travel history, hospitalization, and death. All epidemiological data analysis was performed in R version 4.0.3[35]. Categorical data were compared using Pearson's chi-squared and by calculating $\Sigma(|E-O|)/E$, where expected counts, E, were calculated by standardization to overall reported cases during the same time-

period. Geographic comparisons were visualized by mapping standardized ratios of observed versus expected cases at the county level. The percentage of cases sequenced was graphed by county and month to visualize spatio-temporal sampling. Areas with high pre-sentinel sequencing coverage were investigated to further understand representativeness, as data from these areas may enable robust phylogeographic studies. Similarly, counties with high and low sentinel sequencing coverage were investigated.

To understand and exemplify variability in the structure of the genomic data, we constructed phylogenetic trees of four scenarios using the Nextstrain[9] pipeline for SARS-CoV-2: 1) pre-sentinel surveillance high coverage, low representativeness, 2) pre-sentinel surveillance high coverage, high representativeness, 3) sentinel surveillance, high coverage, high representativeness, 4) sentinel surveillance, low coverage, low representativeness. A rarefaction analysis was performed to examine how sampling affected the diversity of sequences captured in each of these four scenarios. For each value from 1 to n where n is the total number of available sequences for a location/timeframe of interest, we generated 10 subsampled datasets (sampling without replacement). The number of unique haplotypes was counted and plotted as a function of the number of sequences sampled.

Timeliness of data was assessed by comparing the interval between initial specimen collection and genomic data upload to GISAID. Median timeliness by month was assessed, as well as categorical comparisons of data uploaded within <14 days, 14-<28 days, and 28 days+ after specimen collection.

2.3 ***Results***

2.3.1 *Sequencing coverage and representativeness*

During the pre-sentinel surveillance period, 10,653 (3.3%) cases had sequencing information available, compared to 56,106 cases sampled (12.1%) during sentinel surveillance. For all categorical comparisons using Pearson's chi-squared, statistically significant differences were observed between pre-sentinel sequenced cases and sentinel cases. To avoid having a single large discrepancy dominate the measure of representativeness, the calculation $\Sigma(|E-O|)/E$ was used instead of Pearson's chi-squared to directly compare representativeness between these populations (Table 2.1).

Table 2.1: Comparison of demographic characteristics between COVID-19 cases with sequenced specimens and all confirmed COVID-19 cases*

	Presentinel period†				Sentinel period‡			
Variable	Overall	Sequenced	O/E	Σ(|E-O|)/E§	Overall	Sequenced	O/E	Σ(|E-O|)/E§
Total no.	326,850	10,653			463,639	56,106		
Sex				0.73				0.55
Female	159,460 (48.8)	5,326 (50.0)	1.02		230,524 (49.7)	27,163 (48.4)	0.97	
Male	157,133 (48.1)	4,932 (46.3)	0.96		223,711 (48.3)	27,916 (49.8)	1.03	
Other	287 (0.1)	NA¶	0.53		331 (0.1)	55 (0.1)	1.37	
Missing	9,970 (3.1)	390 (3.7)	1.20		9,073 (2.0)	972 (1.7)	0.89	
Age Group, y				2.36				1.58
0–4	7,802 (2.4)	211 (2.0)	0.83		18,499 (4.0)	2,188 (3.9)	0.98	
5–17	32,121 (9.8)	932 (8.7)	0.89		77,782 (16.8)	9,815 (17.5)	1.04	
18–44	165,920 (50.8)	5,128 (48.1)	0.95		224,380 (48.4)	28,909 (51.5)	1.06	
45–64	83,046 (25.4)	2,628 (24.7)	0.97		102,215 (22.0)	11,309 (20.2)	0.91	
65–79	26,724 (8.2)	1,073 (10.1)	1.23		32,000 (6.9)	3,052 (5.4)	0.79	
≥80	10,998 (3.4)	680 (6.4)	1.90		8,591 (1.9)	832 (1.5)	0.80	
Unknown	239 (0.1)	NA¶	0.13		172 (0.0)	NA¶	0.05	
COVID-19 deaths	5,134 (1.6)	448 (4.2)	2.68	1.68	4,568 (1.0)	452 (0.8)	0.82	0.18
Hospitalized for COVID-19	18,992 (5.8)	891 (8.4)	1.44	0.44	25,060 (5.4)	1,721 (3.1)	0.57	0.43
Outbreak-associated	49,165 (15.0)	2,350 (22.1)	1.47	0.47	25,902 (5.6)	4,281 (7.6)	1.37	0.37
LTCF-associated	19,899 (6.1)	1,614 (15.2)	2.49	1.49	7,317 (1.6)	1,105 (2.0)	1.25	0.25
Symptoms				0.61				0.80
Yes	172,070 (52.6)	6,860 (64.4)	1.22		173,363 (37.4)	27,140 (48.4)	1.29	
No	24,182 (7.4)	701 (6.6)	0.89		44,731 (9.6)	3,430 (6.1)	0.63	
Unknown	130,598 (40.0)	3,092 (29.0)	0.73		245,545 (53.0)	25,536 (45.5)	0.86	
Race/Ethnicity				1.95				1.35
Hispanic	70,020 (21.4)	2,671 (25.1)	1.17		53,221 (11.5)	9,285 (16.5)	1.44	
Non-Hispanic, American Indian, or Alaska Native	3,953 (1.2)	161 (1.5)	1.25		5,455 (1.2)	685 (1.2)	1.04	
Non-Hispanic Asian	16,321 (5.0)	755 (7.1)	1.42		21,787 (4.7)	3,261 (5.8)	1.24	
Non-Hispanic Black	14,863 (4.5)	548 (5.1)	1.13		19,812 (4.3)	2,429 (4.3)	1.01	
Non-Hispanic multiracial	5,575 (1.7)	217 (2.0)	1.19		7,707 (1.7)	1,173 (2.1)	1.26	
Non-Hispanic Native Hawaiian or other Pacific Islander	5,338 (1.6)	203 (1.9)	1.17		6,432 (1.4)	704 (1.3)	0.90	
Non-Hispanic White	133,224 (40.8)	4,174 (39.2)	0.96		229,100 (49.4)	24,039 (42.8)	0.87	
Non-Hispanic, other race	3,211 (1.0)	138 (1.3)	1.32		3,271 (0.7)	345 (0.6)	0.87	
Unknown	74,345 (22.7)	1,786 (16.8)	0.74		116854 (25.2)	14185 (25.3)	1.00	
Language				0.94				2.15
English	104,984 (32.1)	3,357 (31.5)	0.98		138,437 (29.9)	18,484 (32.9)	1.10	
Spanish	23,408 (7.2)	884 (8.3)	1.16		9,849 (2.1)	2,474 (4.4)	2.08	
Other	5,137 (1.6)	239 (2.2)	1.43		1,745 (0.4)	337 (0.6)	1.60	
Unknown	12,519 (3.8)	273 (2.6)	0.67		9,261 (2.0)	1,434 (2.6)	1.28	
Missing	180,802 (55.3)	5,900 (55.4)	1.00		304,347 (65.6)	33,377 (59.5)	0.91	

*Values are no. or no. (%). We included all confirmed COVID-19 cases (SARS-CoV-2 RNA detected by molecular amplification) reported among Washington residents from January 21, 2020, through December 31, 2021, in the Washington Disease Reporting System. E, expected counts; LTCF, long-term care facility; NA, not applicable; O, observed counts.

†Cases were classified as presentinel if specimens were sequenced before March 1, 2021.

‡Cases were classified as sentinel if specimens were sequenced on or after March 1, 2021 through the sentinel surveillance program

§Formula used to directly compare representativeness between populations.

¶Counts <10 are censored.

Both pre-sentinel and sentinel sequenced cases were generally representative of sex at birth. During the pre-sentinel surveillance period, sequencing data overrepresented older age groups and hospitalized persons; persons who died of COVID-19 were overrepresented by almost three-fold among pre-sentinel sequenced cases compared to cases that were not sequenced. Sentinel surveillance implementation resolved overrepresentation of decedents but under-represents hospitalized cases and persons 65 and older.

In the early pandemic, known outbreak-associated cases were more commonly sequenced, likely reflecting preferential sample selection of these cases for studies. Similarly, LTCF-associated cases were enriched by 2.5x among sequenced cases. Implementation of sentinel surveillance decreased but did not completely resolve the enrichment for outbreak-associated cases, whereas LTCF-associated enrichment was greatly resolved.

Pre-sentinel sequenced cases had more complete symptom information when compared to all cases, and more often reported symptoms. Similarly, sentinel sequenced cases more often reported symptoms.

Persons reporting minority classifications of race or ethnicity were generally overrepresented among pre-sentinel sequenced cases, and race/ethnicity data was less likely to be missing among sequenced cases than overall. After implementation of sentinel surveillance, persons reporting Hispanic ethnicity or a preferred language of Spanish were overrepresented among sequenced cases. The differential missingness of race data was resolved after implementation of sentinel surveillance.

Industry information was missing for most cases. Comparing cases with industry information available, agriculture, forestry, fishing and hunting, and healthcare and social assistance were overrepresented among sequenced cases. Industry information was missing for >90% of cases

during the sentinel surveillance time-period, so industry representation was not assessed in this evaluation.

More pre-sentinel sequenced cases traveled out of country than expected, indicating likely enrichment for international travelers. Travel information was missing for >95% of cases during the sentinel surveillance time-period, so traveler representation was not assessed in this evaluation.

Reinfection data were captured starting September 1, 2021, so case-level data is not available for most of the study period. From September-December 2021, reinfection cases were underrepresented in sequence data, which may reflect higher average C_t values in this population.

Prior to implementation of sentinel surveillance, geographic coverage was variable and focused on Western Washington (Figure 2.1), with high coverage in King, San Juan, Pacific, and Yakima counties. Some areas of the state had little to no data available. After implementation of sentinel surveillance, geographic coverage equalized regionally across the state, with variable coverage due to service areas of sentinel laboratories, as expected (Figure 2.1).

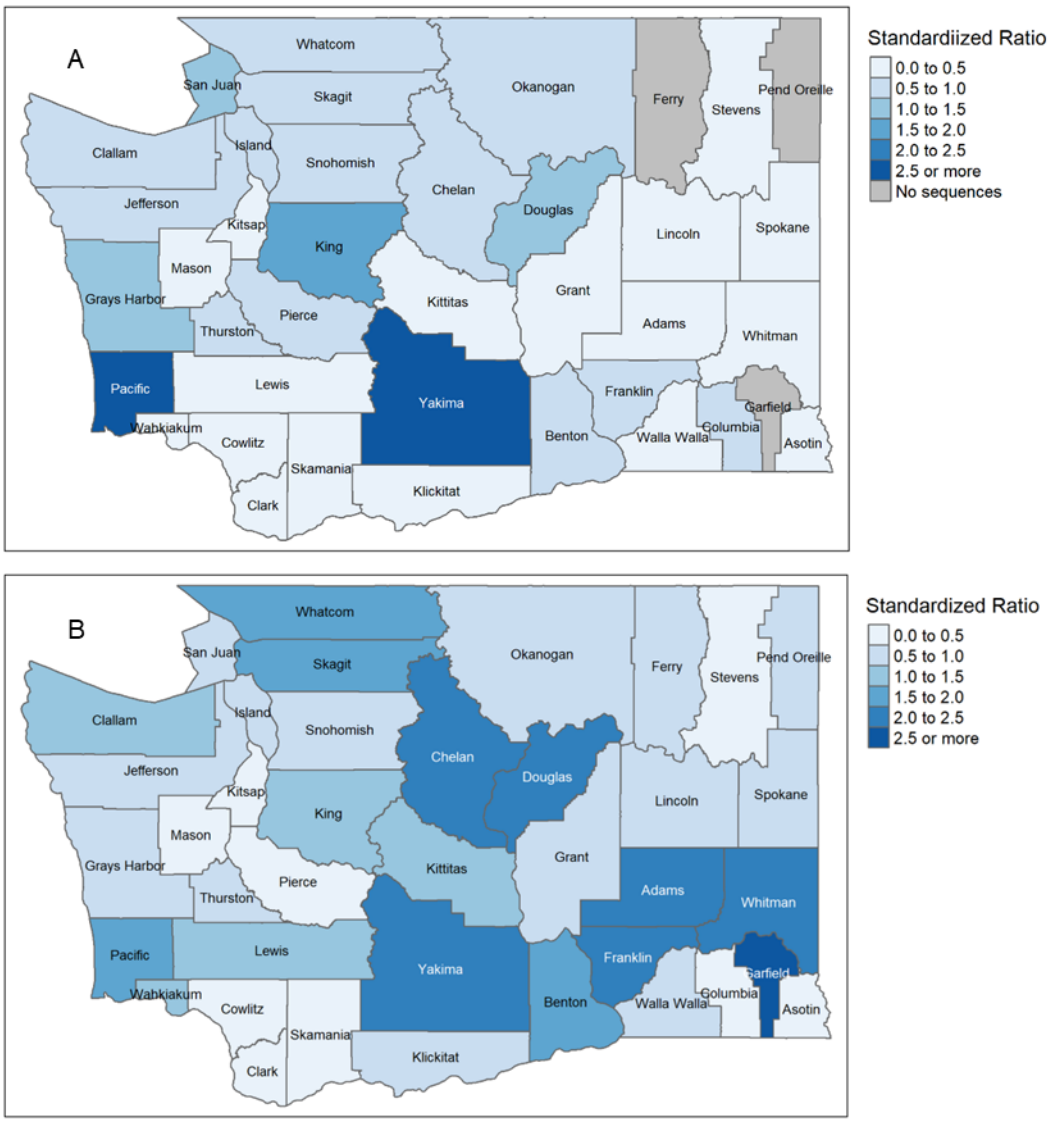


Figure 2.1: Standardized ratio of sequenced cases to overall cases by county.

Map showing extent of sequencing data available for COVID-19 cases A) Presentinel surveillance. B) Sentinel surveillance. Standardized ratios (observed/expected counts) of cases with sequenced specimens are indicated by county. Cases were classified as presentinel if specimens were sequenced before March 1, 2021, and classified as sentinel if specimens were sequenced on or after March 1, 2021 through the sentinel surveillance program. No sequence data were available for 3 counties during the presentinel period.

2.3.2 *County-level assessment of sequencing*

Areas with high pre-sentinel sequencing coverage and high cases numbers were investigated to further understand representativeness (Figure 2.2). During March 2020-June 2020, Yakima County had 19%-30% sequencing coverage across all COVID-19 cases, with high-quality genomic data available for 1,696 cases. This high coverage was partially driven by sequencing of LTCF-associated cases, with 25% of sequenced cases affiliated with LTCF compared to 11% of all cases during this time-period. Sequenced cases were more commonly 65 years or older and less commonly Hispanic or Spanish-language preference.

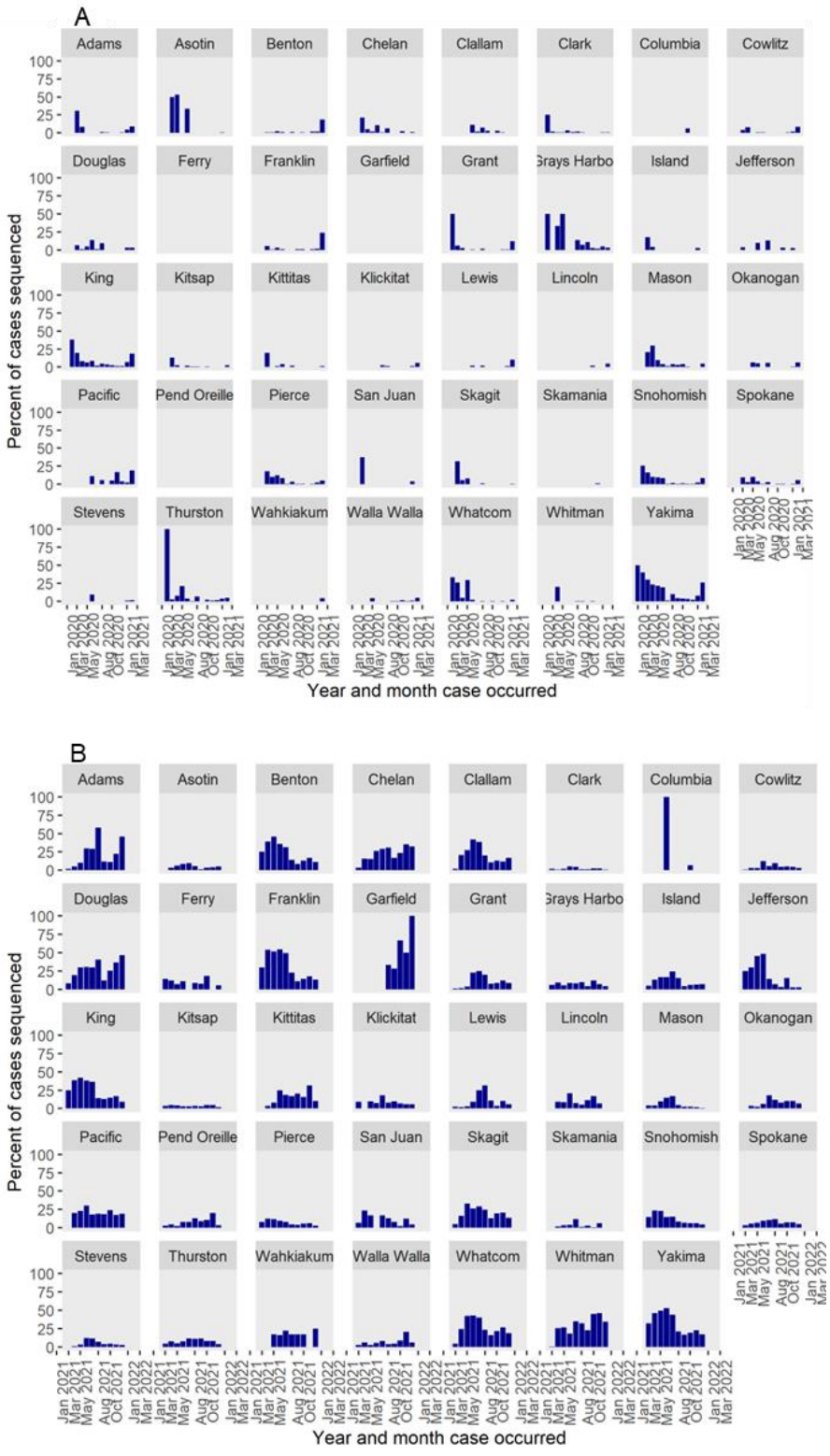


Figure 2.2: Percentages of COVID-19 cases with sequenced specimens by county, month, and year.

Presentinel specimens were sequenced before March 1, 2021; sentinel specimens were sequenced on or after March 1, 2021, through the sentinel surveillance program. A) Presentinel surveillance. B) Sentinel surveillance.

2.3.3 *Phylogenetic comparison of pre-sentinel and sentinel time-points*

A phylogenetic tree placing all sequenced specimens from Yakima cases with onset dates during March 2020-June 2020 is shown in Figure 2.3A**.** During this time-period, most sequences (63%) were classified as Nextstrain clade 20B (Pango lineage B.1.1); 23% were 19B (Pango lineage A), 9% were 20A (Pango lineage B.1) and 5% were 20C. Comparatively, looking at all of Washington State during the same time-period, 20C and 19B (Pango lineage A) were most prevalent.

Sequencing coverage was also high in Yakima County in February 2021, with 26% coverage across all COVID-19 cases, and high-quality genomic data available for 271 cases. During this time-period, smaller differences between sequenced cases and all cases were observed in ethnicity and outbreak-association; otherwise, sequenced cases were largely representative of all cases during this time. Phylogenetic analysis of cases from Yakima during February 2021 is shown in Figure 2.3B**.** The most common lineage identified was 21C (Pango lineage B.1.427/429 or Epsilon), representing 33% of sequences, followed by 20G (Pango lineage B.1.2) at 29%, and 20A-C (13%, 9%, and 15%, respectively). In Washington State, 21C/Epsilon similarly represented 30% of sequences in GISAID in February 2021.

After implementation of sentinel surveillance, variability in geographic coverage was diminished regionally but persisted at the county-level. Counties with high and low sentinel sequencing coverage were investigated to further understand impact of variable sentinel specimen sampling. We specifically compared a county with high coverage from a sentinel laboratory (Whatcom County), to a county with low coverage (Clark County).

During the sentinel surveillance period, sequenced cases from Whatcom County were representative of all cases from Whatcom County by age, sex, race, death due to COVID-19, and LTCF-association. Hospitalized cases were underrepresented among sentinel surveillance cases, reflecting statewide findings. Outbreak-associated and symptomatic cases were slightly overrepresented among sentinel surveillance cases. Phylogenetic analysis of cases from Whatcom County during the sentinel surveillance period is shown in Figure 2.3C, showing a transition from 20I (Alpha) to 21A/21I/21J (Delta) dominance, similar to what is seen in Washington overall.

Clark County had very low sequencing coverage over the sentinel surveillance period, ranging from 0.8% of cases in April 2021 to 4.9% of cases in June 2021. Persons under 45 years of age and outbreak-associated cases were overrepresented among sequenced cases, and hospitalized cases were underrepresented. Phylogenetic analysis of cases from Clark County during the sentinel surveillance period is shown in Figure 2.3D. Despite limited coverage, a similar variant profile is seen compared to Whatcom County and Washington overall.

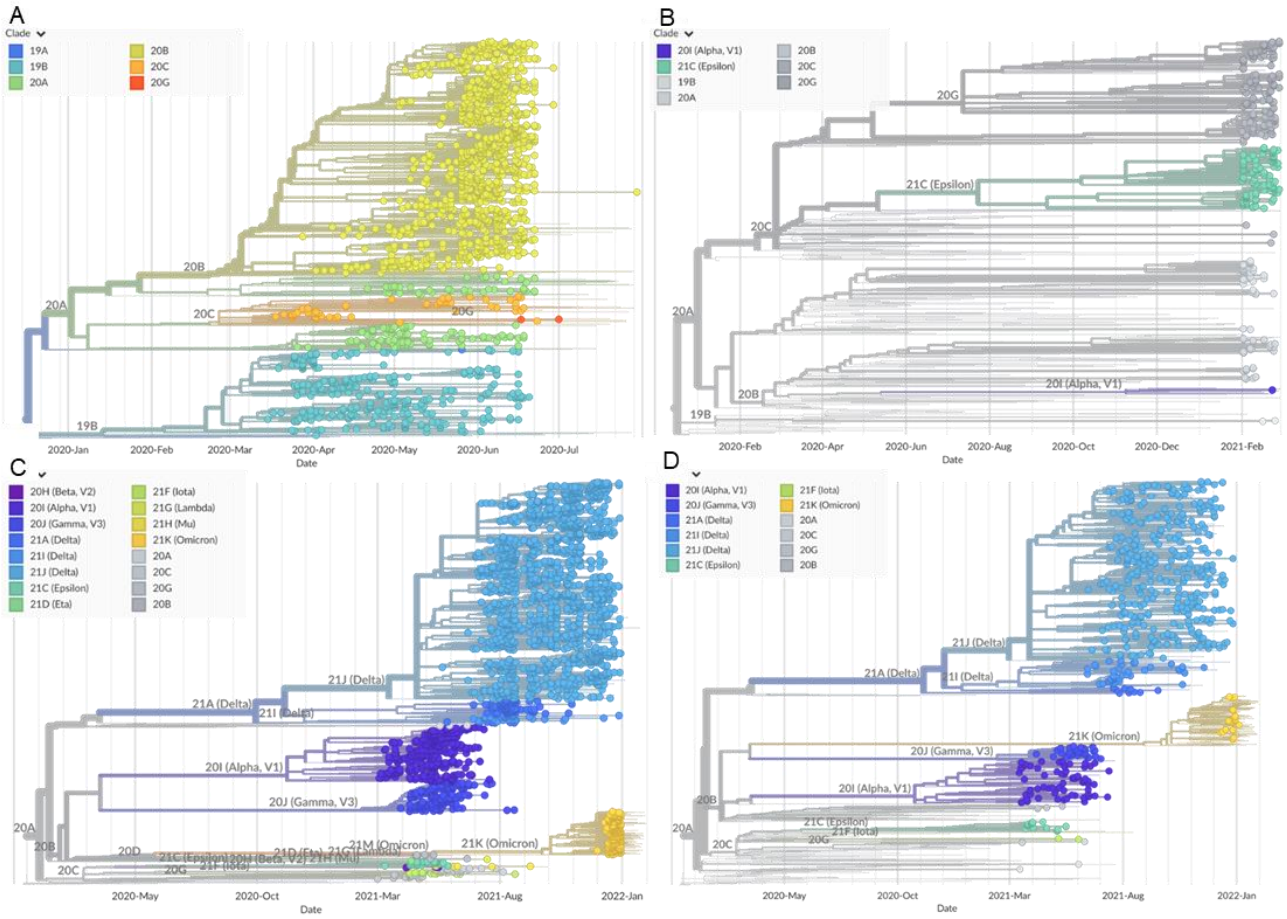


Figure 2.3: Time-scaled phylogenetic analysis of sequence data from Yakima, Clark, and Whatcom Counties.

Trees show SARS-CoV-2 specimens from presentinel COVID-19 cases in Yakima County (2 timepoints) and from sentinel COVID-19 cases in Clark and Whatcom Counties. Presentinel specimens were sequenced before March 1, 2021; sentinel specimens were sequenced on or after March 1, 2021, through the sentinel surveillance program. A) Yakima County, March–June 2020. B) Yakima County, February 2021. C) Whatcom County, March– December 2021. D) Clark County, March–December 2021.

2.3.4 *Rarefaction analysis*

We performed rarefaction analysis and found sentinel sampling from Clark and Whatcom counties displayed higher viral diversity than Yakima County at 2 presentinel timepoints (Figure 2.4). Additional sampling will be required in all scenarios to fully capture circulating viral diversity.

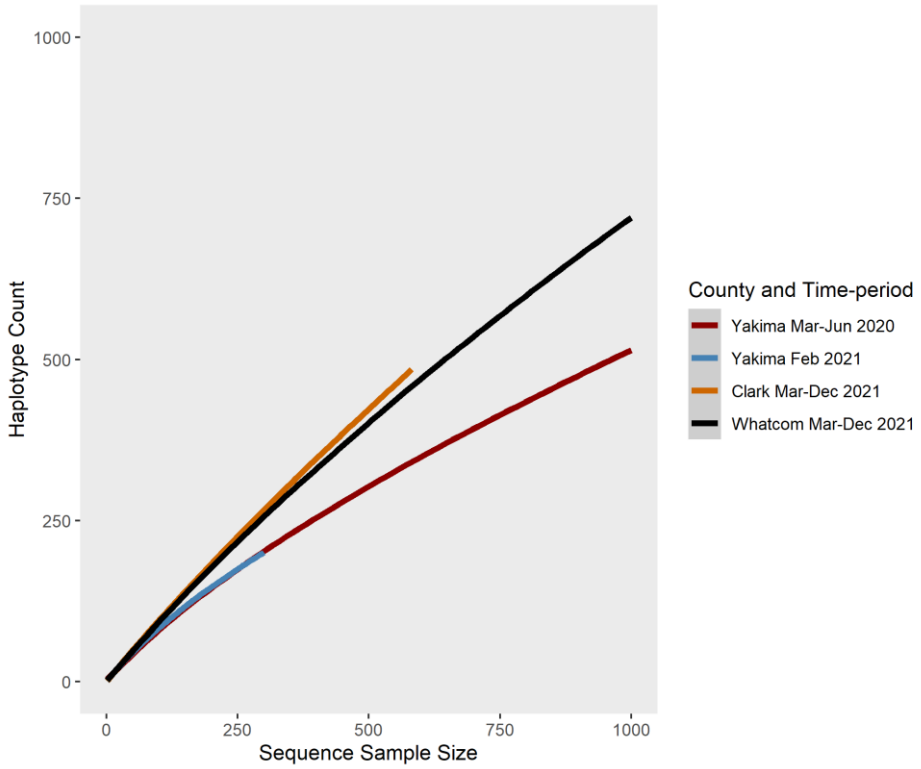


Figure 2.4: Rarefaction analysis of virus haplotype diversity in Yakima, Clark, and Whatcom Counties.

Presentinel COVID-19 cases with sequenced specimens from Yakima County (2 timepoints) were compared with sentinel COVID-19 cases with sequenced specimens in Clark and Whatcom Counties. Presentinel specimens were sequenced before March 1, 2021; sentinel specimens were sequenced on or after March 1, 2021 through the sentinel surveillance program. Haplotype count indicates virus diversity.

2.3.5 *Timeliness of genomic data*

Timeliness of the availability of genomic data in GISAID was variable over the study period (Figure 2.5). During the pre-sentinel period, median timeliness ranged from 23-98 days in February 2020 and October 2020, respectively, with ≥50% of sequences uploaded to GISAID 28 days+ after specimen collection for most months. During the sentinel period, median timeliness ranged from 15 days in December of 2021 to 26 days in August of 2021; most sequences were uploaded to GISAID <28 days after specimen collection in all months after implementation of sentinel surveillance.

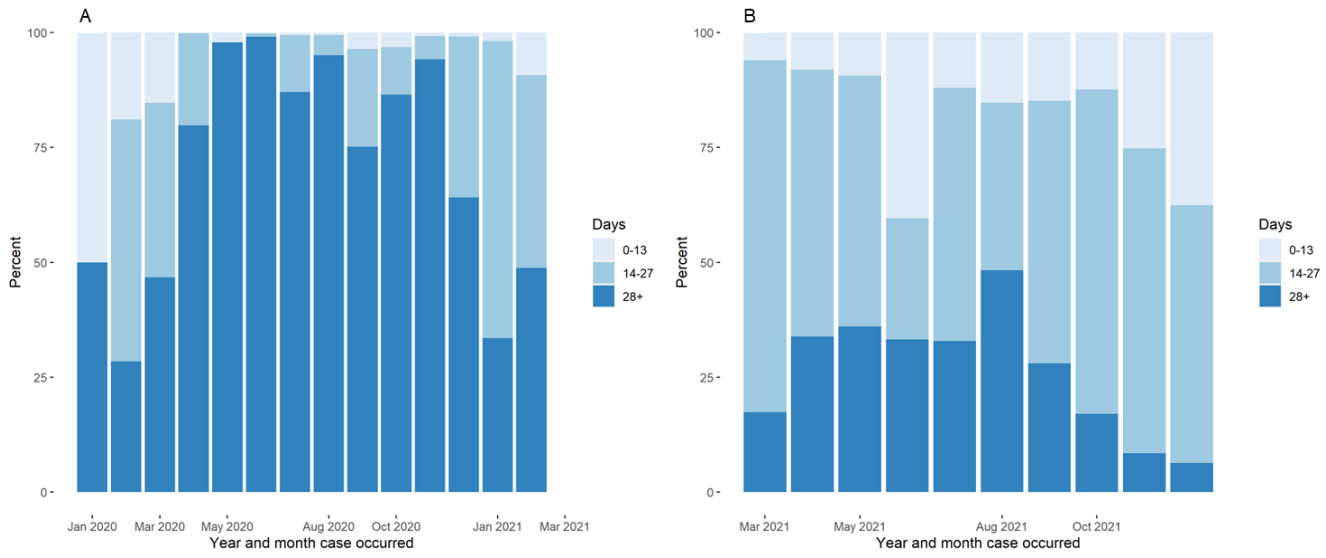


Figure 2.5: Timeliness of sequence data availability.

Graph shows percentages of COVID-19 cases with sequenced data uploaded to the GISAID database (https://www.gisaid.org) within 0–13, 14–27, and ≥28 days after specimen collection. A) Presentinel surveillance (specimens sequenced before March 1, 2021). B) Sentinel surveillance (specimens sequenced on or after March 1, 2021 through the sentinel surveillance program).

2.4 *Discussion*

Data available following implementation of a sentinel surveillance system for sequencing SARS-CoV-2 specimens in Washington state were more epidemiologically and genomically representative and timelier than that available prior to implementation. Specifically, representativeness improved by age, death due to COVID-19, outbreak association status, LTCF-affiliated status, and geographic coverage; increased viral diversity was also noted. Prior to implementation of surveillance, we were unable to identify a county/time-period with representative sampling, with the single exception of Yakima during February 2021. Following implementation, representativeness improved across multiple areas. This increase in representativeness is a critical achievement, as genomic data is routinely displayed to public health leaders and decision-makers; ensuring equitable sampling coverage has important implications for response planning and interventions. Measuring the impact of genomic

surveillance on the public health response in WA was not included as part of this analysis; however, methods for measuring and evaluating effectiveness should be explored. Overrepresentation of older ages in pre-sentinel data was partly driven by selection from LTCF-associated cases and cases resulting in hospitalization or death. The post-implementation decrease in representation of persons aged 65+ improves overall representativeness but actually under-samples these ages, possibly indicating poor coverage of facilities where this population seeks care. Indeed, the implemented system underrepresents hospitalized cases; further consideration to how to better capture both inpatient and outpatient cases in sampling is needed. Prior to implementation of sentinel surveillance, outbreak-associated cases and symptomatic cases were oversampled. Following implementation, overrepresentation of these cases decreased but was not resolved. There are at least three possible explanations for these findings: 1) symptomatic cases may be more likely to be sequenced due to, on average, higher viral loads leading to successful sequencing, 2) asymptomatic cases may be detected through screening programs not associated with sentinel laboratories, 3) outbreak-associated specimens may be sent to sentinel laboratories to ensure sequencing for investigation purposes. Random sampling among specimens received at sentinel laboratories could thereby still lead to biased samples. Minority race and ethnicity were more commonly reported among pre-sentinel sequenced cases; however, data were also more complete among sequenced cases. It is unclear whether this was true overrepresentation or whether data was differentially missing by race among all cases. After implementation of sentinel surveillance, persons reporting Hispanic ethnicity and Spanish language were overrepresented compared to overall cases statewide. This likely reflects the catchment areas of the sentinel laboratories. Variability was identified in geographic coverage pre- and post-implementation of sentinel surveillance. Pre-implementation, coverage was

focused in Western Washington, reflecting coverage areas of laboratories connected to sequencing capacity. Implementation of sentinel surveillance opened access to sequencing for additional laboratories and ensured more equitable regional coverage, although variability at the county and sub-county levels remains. Variable coverage and representativeness at the sub-statewide level should be considered when using data for specific analyses. Increasing geographic coverage will require addition of sentinel laboratories contributing specimens from areas of low coverage.

Other epidemiologic information was of interest in assessing representativeness, including industry and occupation, travel history, and reinfection status. However, data on these variables was highly incomplete, limiting their utility. As public health systems pivot away from capturing data through individual case interviews, consideration should be paid to the dataset available for assessing sampling of specimens for sequencing. Only with epidemiological metadata to pair with genomic data can the full potential of public health impacts through genomic epidemiologic surveillance be realized.

Genomic diversity capture was assessed using rarefaction; viral diversity has been and continues to be dynamic over the course of the pandemic. Measurement of true viral diversity requires random or complete sampling. It is likely that actual circulating diversity differed across the locations and timepoints included in our analysis; if circulating diversity generally increased over time, this would bias our conclusions toward assumption of improved capture due to surveillance.

Other states and countries have implemented variable practices for selection of SARS-CoV-2 specimens for sequencing. Methods that rely on convenience samples, as did our pre-sentinel system, likely have sampling biases that impact phylogenetic inference. In these settings,

methods to weight cases for inclusion in estimates using selection probabilities may help to enable bias correction. Alternatively, approaches to correct for non-representative sampling during analysis, such as inverse probability weighting, should be considered. Even after implementation of a sentinel surveillance system, some biases, such as under-sampling of hospitalized cases, remain and should be corrected for through diversifying sources of specimens. Ongoing evaluation and improvement of systems is necessary, especially in the context of performing epidemiological studies. Many epidemiological studies on COVID-19 have inclusion criteria including availability of genomic data; if sampling biases are not understood, biased conclusions may be drawn. These findings underscore the need for the co-development of genomic epidemiology programs alongside bioinformatics programs in public health departments, as epidemiological and phylogenetic analyses are best performed after consideration of sampling methods and data limitations.

Although representativeness and timeliness were the focus of this evaluation, other features are important to consider in the design of surveillance systems, including simplicity, flexibility, sensitivity, and stability, among others[1]. Sentinel surveillance systems require ongoing coordination with laboratory partners and are not simple; stability requires public health resources. Alternative systems to allow for representativeness and timeliness while increasing simplicity and stability could include requirements for specimen submission, such as those commonly employed for foodborne pathogens and other notifiable conditions. Sensitivity is an important topic in the context of surveillance system goals for rare variant detection and timely surveillance for circulating variant proportions. Right-size sampling, such as that performed for influenza surveillance, should be considered[36,37].

Even after careful consideration to design of a surveillance system for pathogen sequencing and pairing with epidemiologic data, limitations remain due to specimen requirements for sequencing. Any studies utilizing surveillance sequencing data should report the following limitations: 1) selection into laboratory-based diagnostic testing may depend on many factors, which are difficult to assess and grow increasingly complex with improved availability of at-home testing, 2) among positive test results, those with low C_t values are more likely to be sequenced, and thereby the representativeness of sequencing data is inherently limited.
A pre- and post-implementation assessment of representativeness such as this is limited in the causal inferences that can be drawn. Other concurrent factors may also affect representativeness and timeliness during this study period. Importantly, CDC surveillance efforts were also increased during this timeframe; samples sequenced under CDC surveillance were coded as sentinel and were analyzed as part of the implemented sentinel surveillance system.

2.5 *Conclusions*

In conclusion, implementation of a sentinel surveillance system for sequencing SARS-CoV-2 specimens in Washington state was associated with improved genomic and epidemiological representativeness and timeliness of available sequence data for our state. Ongoing evaluation and improvements are necessary to ensure representative capture of in-patient settings. As public health leaders discuss changes to COVD-19 surveillance systems nationally, consideration should be paid to the dataset required to assess representativeness of sampling for sequencing. Cross-jurisdictional sampling bias is a concern to the validity of application of phylogeographic methods; attention to sampling will improve the usefulness of these datasets for public health practice.

Chapter 3. CHANGING GENOMIC EPIDEMIOLOGY OF COVID-19 IN LONG-TERM CARE FACILITIES DURING THE 2020-2022 PANDEMIC, WASHINGTON STATE

3.1 *Introduction*

The COVID-19 pandemic disproportionately impacted residents of long-term care facilities (LTCFs), who have suffered higher mortality rates than the general population; in Washington State (WA), LTCF-associated cases represent 3% of cases, but 30% of deaths due to SARS-CoV-2.[38] This impact materialized in WA and across the US despite early recognition of LTCFs as high-risk settings due to residents' advanced age, chronic underlying health conditions, congregate living, asymptomatic transmission, and movement of healthcare personnel.[39-41] Based on these concerns, Centers for Disease Control and Prevention (CDC) developed recommendations over the course of the pandemic for infection prevention and control (IPC) in LTCFs, including training, use of personal protective equipment (PPE) and hygiene measures, visitor restrictions, resident distancing and cohorting, environmental cleaning and disinfection, testing and reporting to public health jurisdictions, and provision of staff sick leave.[42] Similarly, WA's governor, secretary of health, and Department of Health (DOH) developed and instituted regulations and guidance governing prevention efforts.[43,44] Centers for Medicare and Medicaid Services (CMS) outlined rules for testing staff and residents of LTCFs.[45] Changes in these rules, regulations, and guidance over time are expected to have impacted transmission dynamics in LTCF settings.

One key tool for understanding transmission dynamics in-place is pathogen genomic sequencing and analysis, particularly phylogeographic analysis. Understanding sampling methodology is

important for describing potential bias in this type of analysis.[5,32,46] Systems for sequencing SARS-CoV-2 specimens have changed over time. Prior to March 2021, sampling for sequencing from WA residents was convenience- or research-based. In March 2021, a sentinel surveillance system was implemented in WA to support representative sampling.[46] The population of WA LTCF-associated cases with genomic data available is as-yet undescribed. Additionally, the utility of the existing surveillance system for adding insight and actionable data for public health practice has not been completely explored.

Multiple examples of genomic epidemiology studies of single outbreaks or facilities exist in the literature, including from WA. A previous study documented the utility of targeted genomic surveillance during two SARS-CoV-2 outbreaks in LTCFs in WA.[47] Likewise, a study of a single LTCF-associated outbreak in WA early in the pandemic utilized genomic epidemiology to understand phylogenetic clustering of cases within the facility.[48] Fewer studies have leveraged pathogen genomic data to describe how transmission dynamics changed over the pandemic or describe the impact of sequence data availability on public health action. A review article assessing published genomic epidemiologic investigations during 2020 documented the value of this type of analysis for identifying independent clusters of infections but found that large-scale sequencing of outbreaks added limited value after sequencing initial cases, focusing on individual outbreak- or facility-level studies.[49] An analysis of all care-home linked cases in the east of England used genomic epidemiology to explore large-scale transmission dynamics in nearly 300 facilities; however, this analysis was limited to a 3-month study period.[50]

Here, we aim to assess the utility of genomic data produced for LTCF-associated cases to add information for public health action over the course of the SARS-CoV-2 pandemic, from 2020-2022. We pair patient-level epidemiological and pathogen genomic data to understand variations

in transmission patterns over time. Specifically, we address the following questions of public health concern: is available genomic data obtained from LTCF-associated cases representative of all LTCF-associated cases? Do temporal changes in guidance or policy apparently impact intra-facility transmission patterns? Given available data, which genomic-epidemiologic methods are most applicable for ongoing or routine data analysis? And finally, what changes are needed to ensure the ongoing use of genomic data to explore transmission in LTCF settings?

3.2 *Methods*

3.2.1 *Data collection and cleaning*

All confirmed COVID-19 cases reported among WA residents in the Washington Disease Reporting System (WDRS) as of December 19, 2022 were included, including reinfection cases.[51] Sequences uploaded to the GISAID EpiCoV database indicating WA in their geographic tag were linked to these cases using laboratory accession numbers or patient demographics.[26] For cases with multiple specimens sequenced, only the first specimen was used for analysis. Long-term care facilities were defined as: nursing homes, assisted living facilities, adult family homes, enhanced services facilities, and intermediate care facilities for individuals with intellectual disabilities. Cases in WDRS are categorized as LTCF-associated if association with a facility is noted in case interview, medical record, facility line list, address or telephone match to the facility or another measure indicated by the Local Health Jurisdiction. LTCF-associated cases therefore include residents, employees, and visitors if association is noted.

Enhanced data obtained on October 24, 2022 from Yakima Health District tracking additional details related to LTCF cases and outbreaks were linked to WDRS and GISAID data using name and date of birth and conducting probabilistic matching with manual review.

3.2.2 *Representativeness analysis*

All epidemiological data analysis was performed in R version 4.2.2.[52] Representativeness of LTCF-associated cases with sequencing performed was assessed by comparing to all LTCF-associated cases on: sex, age, race, ethnicity, language, outbreak association, symptom status, hospitalization, death, and facility type. Sampling for sequencing over time in the full population and in LTCFs was graphed.

3.2.3 *Definition of study time-periods*

Information available from the WA Governor's News Release Archive and WA DOH records was used to construct a timeline of key modifications to rules, regulations, or guidance for LTCFs. This timeline was used to divide the study period into six segments of approximately similar lengths, marked by key policy changes (Table 3.1). Events that impacted movement or visitation and sample selection for sequencing were prioritized in defining study time-periods.

Table 3.1: Dates and key events defining each study time-period.

Study Period	Event Date	Event Description
1 *(Jan 20,2020-Mar 9, 2020)*	Jan 20, 2020	First COVID-19 case confirmed in WA
2 *(Mar 10, 2020-Aug 11, 2020)*	Mar 10, 2020	Governor issues rules to restrict LTCF visitation, require visitor screening, and require isolation of residents testing positive for SARS-CoV-2
	Mar 23, 2020	Stay home, stay healthy order
	Jun 26, 2020	First statewide masking order takes effect
3 *(Aug 12, 2020-Mar 9, 2021)*	Aug 12, 2020	Updated LTCF visitation guidance allows for increased visitation
	Aug 25, 2020	Centers for Medicare & Medicaid Services (CMS) releases testing requirements for LTCF staff and residents
	Nov 15, 2020	LTCF visitation restrictions re-instituted
	Dec 20, 2020	LTCF vaccination campaign begins
	Mar 1, 2021	Sentinel sampling for genomic sequencing initiated
4 *(Mar 10, 2021-Aug 22, 2021)*	Mar 10, 2021	Masking and visitation restrictions lifted for fully vaccinated
	Mar 17, 2021	Second phase of vaccine roll-out begins
	Mar 19, 2021	Indoor LTCF visitation allowed if visitor or resident is fully vaccinated
	Apr 1, 2021	LTCF vaccination campaign complete
	Apr 15, 2021	Vaccines available for everyone aged 16+
	Jul 1, 2021	Implemented the 10/70 rule for visitation in LTCFs: indoor visitation restricted only for unvaccinated residents in facilities located in areas with >10% positivity and <70% of residents vaccinated
5 *(Aug 23, 2021-Mar 11, 2022)*	Aug 23, 2021	Statewide masking order takes effect
	Oct 18, 2021	State deadline for healthcare workers to be vaccinated or have exemption
6 *(Mar 12, 2022-Dec 19, 2022)*	Mar 12, 2022	Statewide masking order rescinded
	Sept 23, 2022	CMS removes recommendation for routine asymptomatic LTCF staff testing
	Oct 31, 2022	State of emergency ended

3.2.4 *Genomic subsampling*

Full global data, restricted to those samples with complete date information available, were downloaded from GISAID. Due to the challenges associated with the size of this dataset, we subsampled to include: all sequences from Washington State, 3,000 random sequences from North America, and 3,000 random sequences from regions outside North America to allow for

both spatiotemporal diversity and contextualization of LTCF-associated samples in WA. Contextual data included in the phylogenetic analyses were selected from this down-sampled dataset according to genetic proximity to the focal samples (LTCF-associated samples). We specified contextual data sampling to include up to 1,500 genomes per time-period from WA, sampled from all counties and months, ten genomes per month from other US states, and ten genomes per month from each of the global regions. Known duplicate samples were excluded from the contextual sampling.

3.2.5 *Phylogenetic tree generation*

Phylogenetic trees corresponding to the six study periods were constructed using Nextstrain SARS-CoV-2 workflow, which aligns sequences against the Wuhan Hu-1 reference using nextalign (https://github.com/nextstrain/nextclade), infers a maximum-likelihood phylogeny using IQ-TREE, and estimates molecular clock branch lengths using TreeTime. We specified the use of discrete trait analysis (DTA) within TreeTime.[9,13]

Data from Yakima LTCFs were separated into two time periods: January-August 2020 and August 2021-December 2022; phylogenetic trees corresponding to each of these time periods were constructed in Nextstrain as described above. These trees were used to select three facilities for further analysis.

3.2.6 *Discrete trait analysis*

Migration history was inferred for each of the time-periods using a LTCF-associated binary variable. We defined a migration event into a LTCF as occurring if a parent node had >50% probability to be assigned the "non-LTCF discrete trait", and the child node had >50% probability to be assigned as "LTCF." The Python library Baltic was used for parsing phylogenetic trees and estimating post-introduction clade sizes (version downloaded from:

https://github.com/alliblk/ncov-humboldt/blob/main/baltic.py).[53] The introduction rate was calculated as the number of unique introduction events over time.

3.2.7 *Genomic epidemiologic analysis*

Agreement between clade designation and "outbreak-association" status in the metadata was analyzed for clade sizes >1. Statewide data were not available for type of association (staff/resident/visitor); age group was evaluated as a proxy to understand possible staff versus visitor introductions. Microreact was used to visualize multiple data elements overlaid on the state-wide phylogenetic trees.[12] Sub-trees for each of the Yakima-specific facilities selected for further analysis were imported into MicrobeTrace for visualization and network analysis.[11]

3.2.8 *Transmission tree inference*

Time trees from the January-August 2020 analysis for the three Yakima facilities were input into TransPhylo version 1.3.2 to infer transmission trees and describe the role of staff versus resident introduction and transmission events.[15,54] Previous analyses of SARS-CoV-2 genomic data using TransPhylo were used as reference.[55-57] For this analysis, minimum branch distance was set to one day and viral generation times 1-14 days with a median of 5.5 days and equal sampling time were assumed,[55] along with a gamma distribution. Markov chain Monte Carlo (MCMC) analysis was performed with 500,000 iterations. Convergence was visually inspected.

3.3 ***Results***

3.3.1 *Sequencing coverage*

Among 58,086 LTCF-associated COVID-19 cases, 4,550 (7.8%) had sequencing performed on at least one specimen. This compares to an average of 9.6% of all reported WA cases with genomic data available. The proportion of cases with sequencing data available varies over time (Figure 3.1), ranging from 5% to 30% across study periods. LTCF-associated cases were

sequenced at higher frequencies than general-population cases prior to November 2021. During and after November 2021, LTCF-associated cases were sequenced at similar or lower frequency than all cases, with a notable drop-off in sampling beginning in May 2022. Sequencing rates vary at the facility- and outbreak-level.

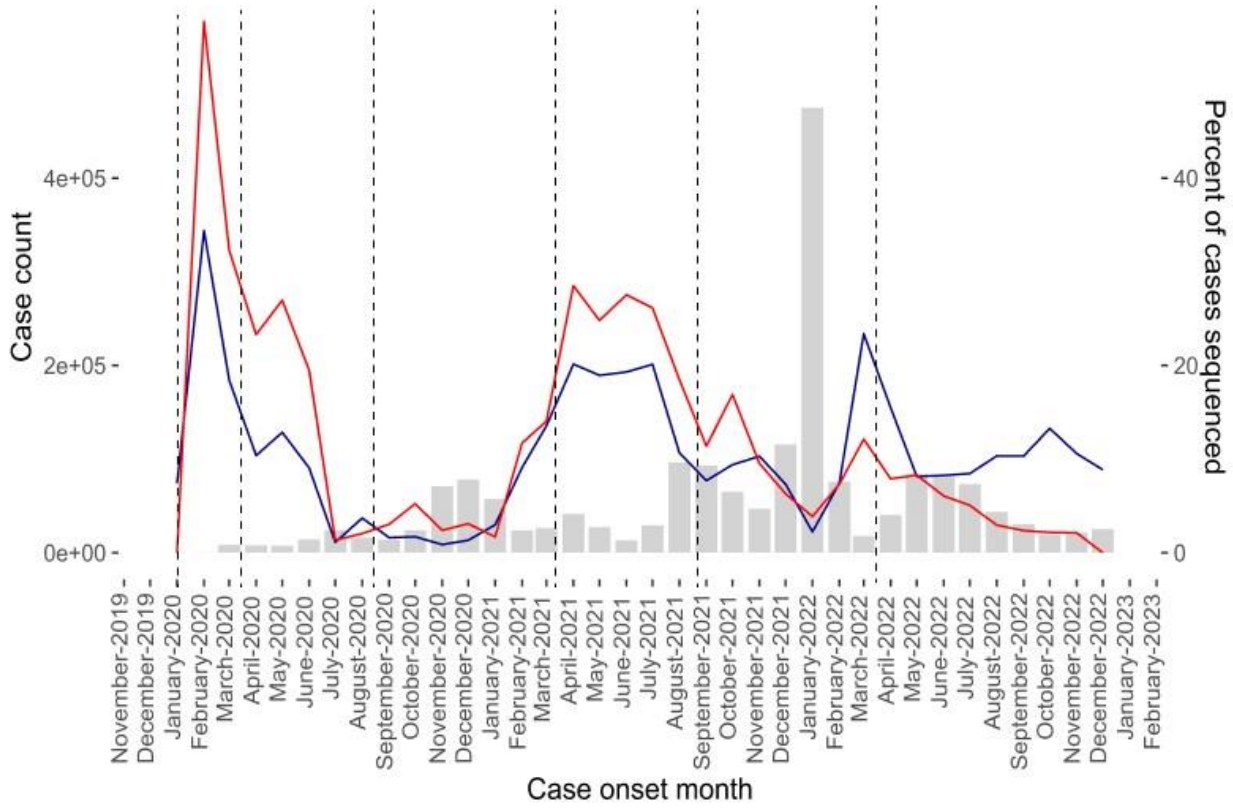


Figure 3.1: Number of reported cases and percent of cases sequenced.

Number of reported cases (gray bars), percent of all cases (blue line) and LTCF-associated cases (red line) sequenced by month, Jan 2020-Dec 2022. The dashed vertical lines indicate the start of each study-time period.

3.3.2 *Representativeness*

Table 3.2 compares LTCF-associated cases with sequences available to all LTCF-associated cases. Cases with sequences available were generally demographically representative of all cases by age group, sex, race/ethnicity, language, and facility type but were more likely fatal or hospitalized and were more likely to have symptom information available.

Table 3.2: Comparison of the demographic characteristics between all reported LTCF-associated cases and the subset of those cases with genomic data available (sequenced cases).*

Variable	All reported cases	Sequenced cases
Total no.	58,086	4,550
Sex		
Female	37705 (64.9)	2970 (65.3)
Male	17679 (30.4)	1431 (31.5)
Other	39 (0.1)	NA¶
Missing	2663 (4.6)	146 (3.2)
Age Group, y		
0–4	105 (0.2)	10 (0.2)
5–17	443 (0.8)	26 (0.6)
18–44	15274 (26.3)	1062 (23.3)
45–64	11177 (19.2)	836 (18.4)
65–79	12174 (21.0)	1068 (23.5)
≥80	18850 (32.5)	1548 (34.0)
Unknown	61 (0.1)	NA¶
COVID-19 deaths	4465 (7.7)	508 (11.2)
Hospitalized for COVID-19	7564 (13.0)	693 (15.2)
Outbreak-associated	37480 (64.5)	2781 (61.1)
Symptoms		
Yes	17014 (29.3)	1763 (38.7)
No	7415 (12.8)	518 (11.4)
Unknown	33655 (57.9)	2269 (49.9)
Race/Ethnicity		
Hispanic	3310 (5.7)	363 (8.0)
Non-Hispanic American Indian or Alaska Native	490 (0.8)	63 (1.4)
Non-Hispanic Asian	2265 (3.9)	191 (4.2)
Non-Hispanic Black	2494 (4.3)	166 (3.6)
Non-Hispanic multiracial	471 (0.8)	43 (0.9)
Non-Hispanic Native Hawaiian or other Pacific Islander	372 (0.6)	33 (0.7)
Non-Hispanic White	29429 (50.7)	2153 (47.3)
Non-Hispanic, other race	319 (0.5)	25 (0.5)
Unknown	1513 (33.3)	18934 (32.6)
Language		
English	13579 (23.4)	1256 (27.6)
Spanish	294 (0.5)	29 (0.6)
Other	295 (0.5)	37 (0.8)
Unknown	1298 (2.2)	88 (1.9)
Missing	42620 (73.4)	3140 (69.0)
Facility Type		
Adult Family Home	3764 (6.5)	282 (6.2)
Assisted Living Facility	26076 (44.9)	1888 (41.5)
Facility for Individuals with Intellectual disability	34 (0.1)	NA¶
Nursing Home	28212 (48.6)	2379 (52.3)

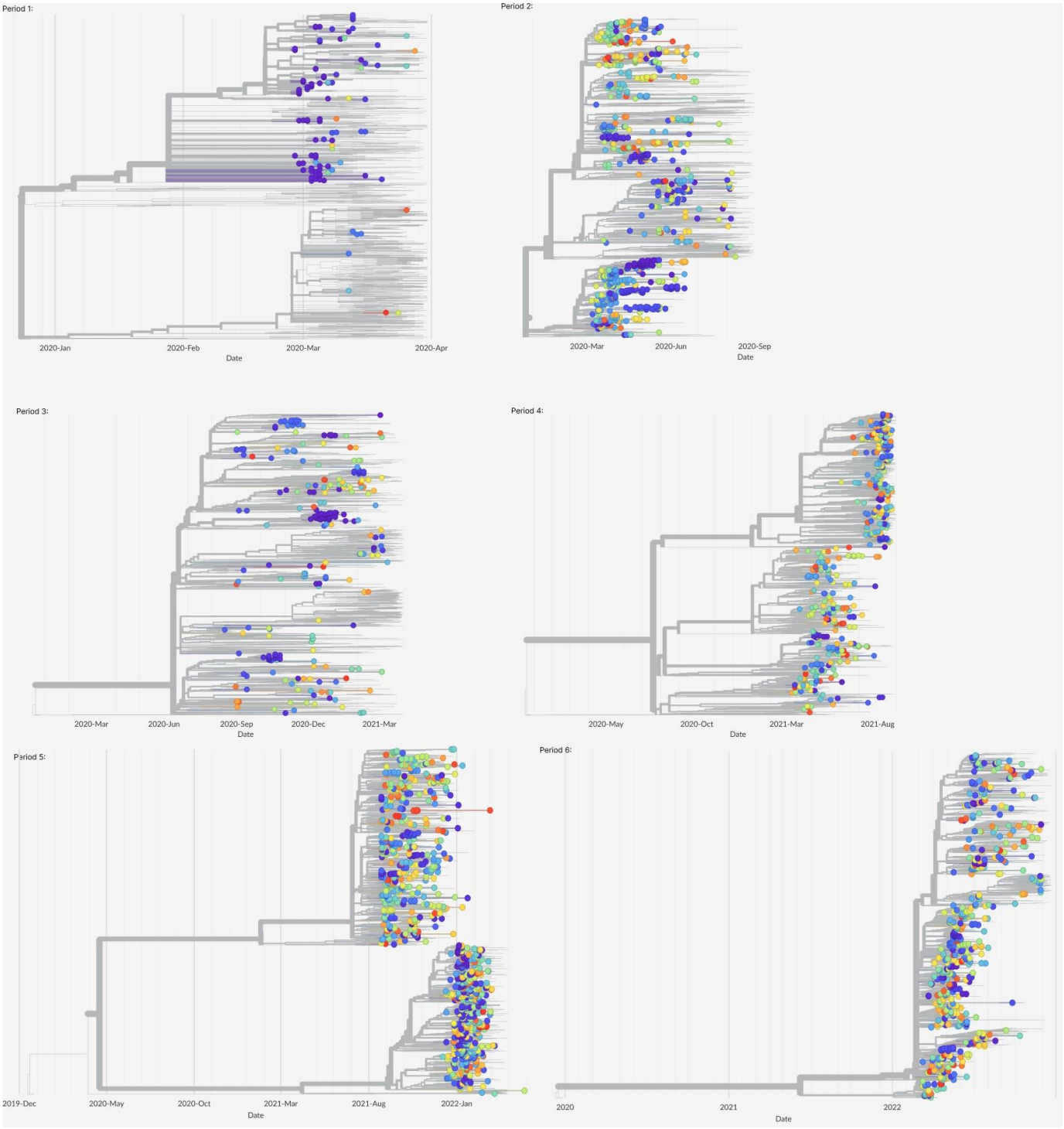
*Values are no. or no. (%).
¶Counts <10 are censored.

3.3.3 *Phylogenetic and genomic epidemiologic analyses*

Figure 3.2 shows time-scaled (A) and divergence-scaled (B) phylogenetic trees of sequenced LTCF cases across all time periods outlined in Table 1. LTCF-associated cases are dispersed and intermixed with both LTCF-associated and non-LTCF cases; across each time-period the dominant lineages match across these groups (Figure 3.3). Multiple epidemiological clusters within unique facilities are visualized, as well as linked cases from different facilities. Many visualized clusters reveal phylogenetic diversity with long branch lengths, indicating missing samples in the transmission chains consistent with known sampling patterns.

Age-group was evaluated as a proxy for resident status using supplemental data from Yakima County. The oldest age groups, consisting of persons aged 65 and older were >90% residents. Persons in the 45-64 age group were 43.3% residents; 95.5% of persons 18-44 were staff. Across all time periods, sequences from different age groups are interspersed.

A


Period 1:
2020-Jan
2020-Feb
2020-Mar
2020-Apr
Date
Period 2:
2020-Mar
2020-Jun
2020-Sep
Date
Period 3:
2020-Mar
2020-Jun
2020-Sep
2020-Dec
2021-Mar
Date
Period 4:
2020-May
2020-Oct
2021-Mar
2021-Aug
Date
Period 5:
2019-Dec
2020-May
2020-Oct
2021-Mar
2021-Aug
2022-Jan
Date
Period 6:
2020
2021
2022
Date

B

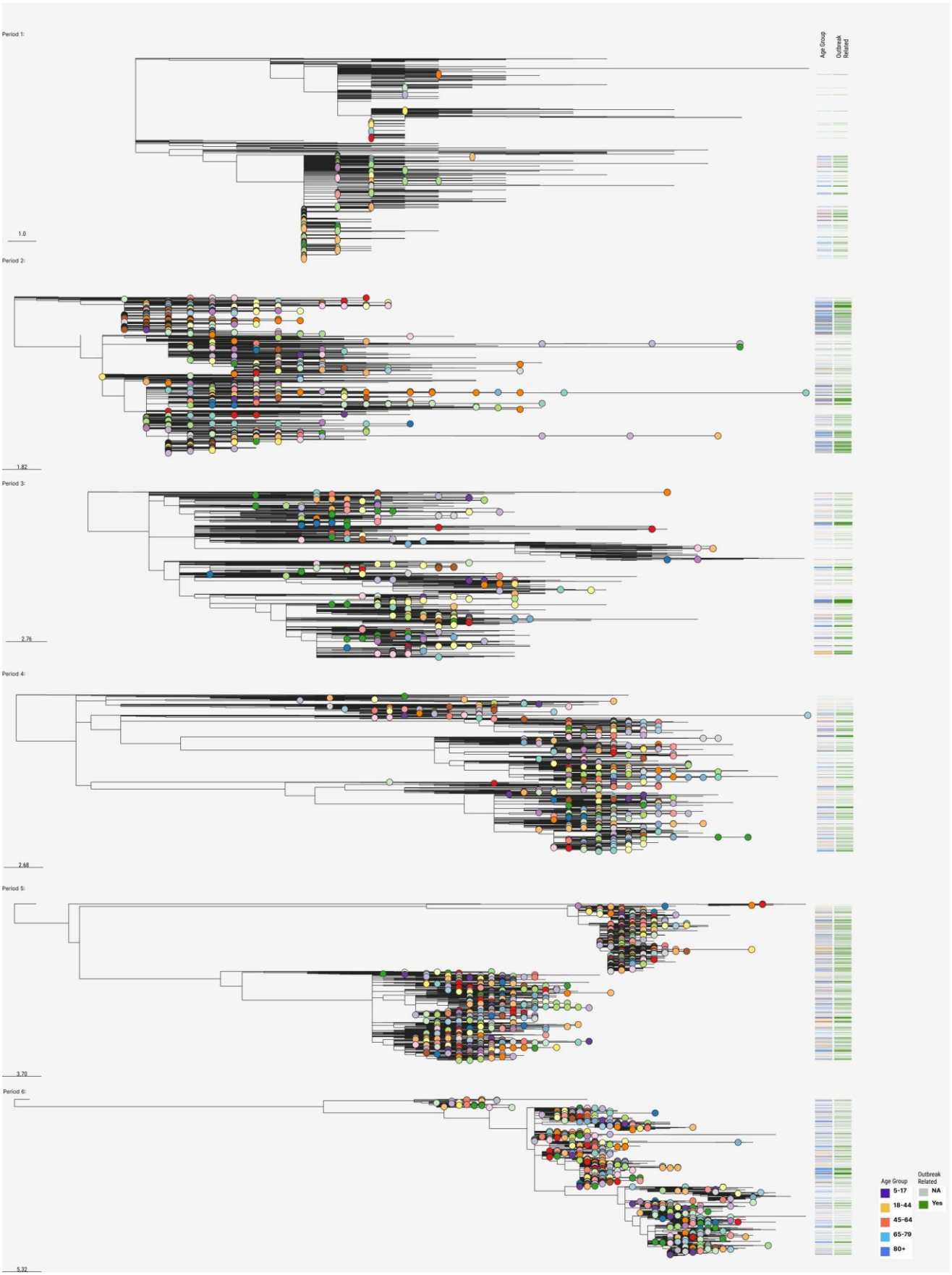

Period 1:
Age Group
Outbreak Related
1.0
Period 2:
1.82
Period 3:
2.76
Period 4:
2.68
Period 5:
3.70
Period 6:
5.32
Age Group
5-17
18-44
45-64
65-79
80+
Outbreak Related
NA
Yes

Figure 3.2: Maximum Likelihood phylogenetic trees from each study time-period.

Time-scaled (A) and divergence-scaled (B) Maximum Likelihood phylogenetic trees from each study time period. Divergence-scaled trees include indication of age group and outbreak status for LTCF-associated cases. Nodes are colored by individual facility; colored nodes are LTCF-associated cases, gray nodes are contextual samples.

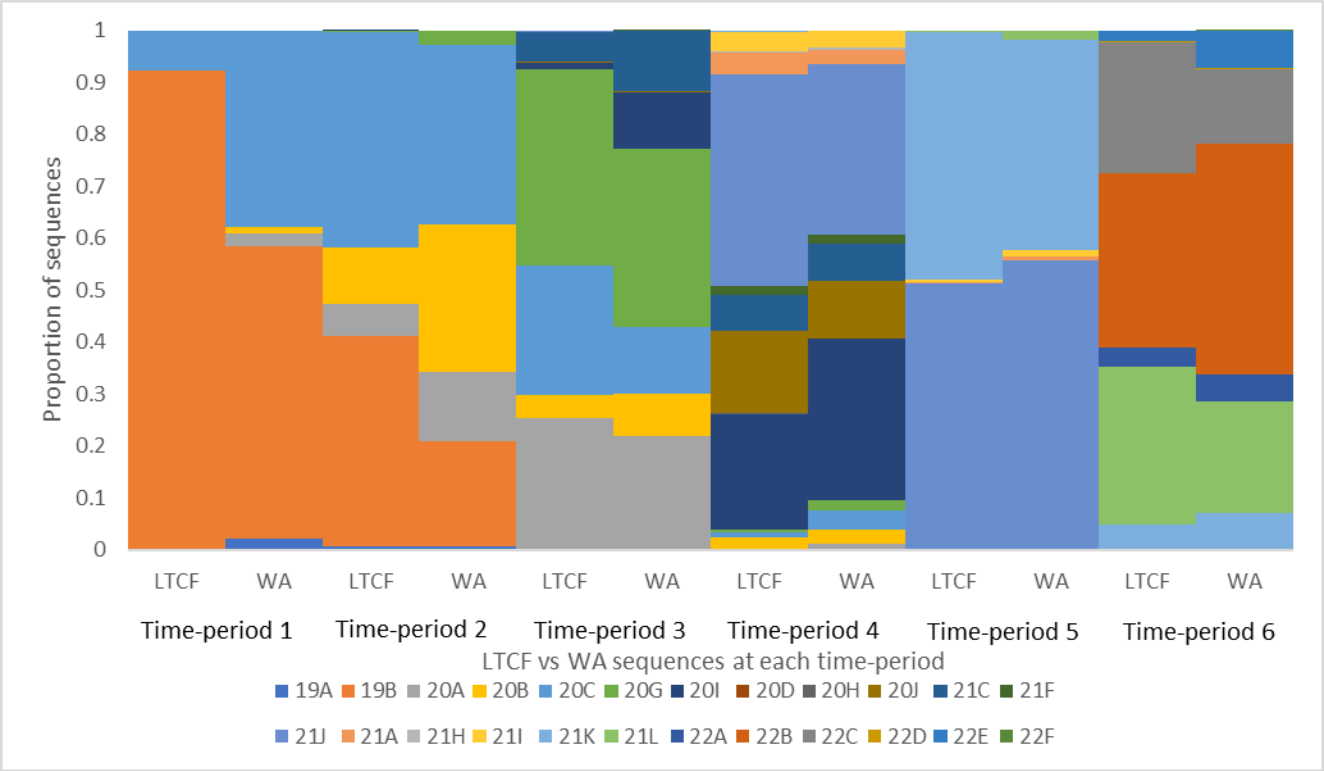


Figure 3.3: Proportion of Nextstrain clades among LTCF-associated vs non-LTCF Washington sequences, by time-period.

Figure 3.4 shows the post-introduction clade sizes among LTCFs in each time-period. Most clusters are single introductions across all time-periods, with large outbreaks (>10 sequences) becoming increasingly rare. The average number of introductions per day varied from 1.6 during time-period 4 to 0.7 during time-period 3. Additional detail regarding post-introduction clade sizes, introductions per day, and information regarding sampling during each time-period is provided in Table 3.3.

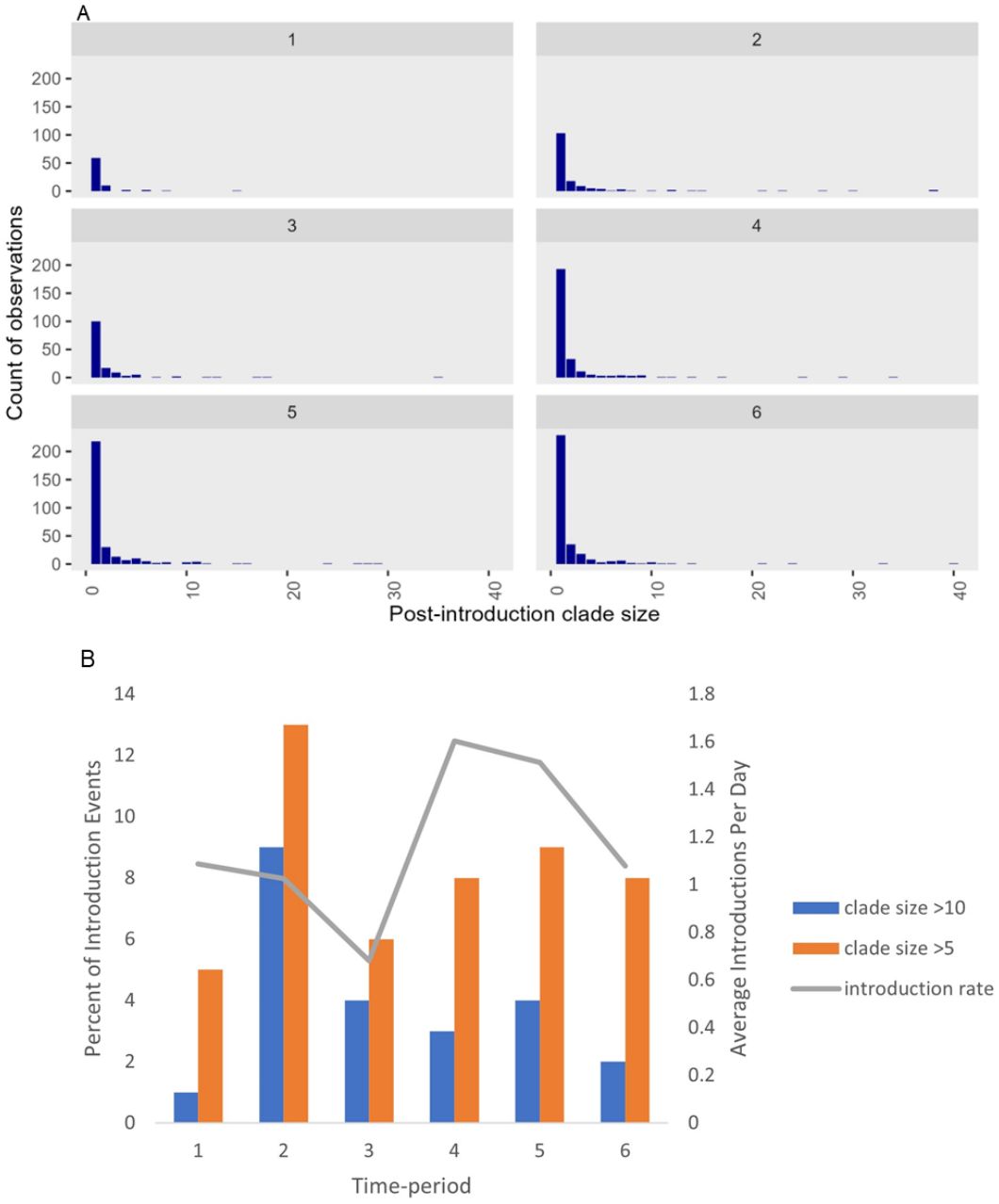


Figure 3.4: Post-introduction clade sizes and introduction rates across time-periods.

A) Post-introduction clade sizes among LTCFs in each time period, 1-6*.
*Footnote: Additional single observations outside of the figure scale were observed for the following time-periods
Time-period 2: 52, 64, 253, 303
Time-period 3: 51
Time-period 5: 57, 405
B) Introduction rate (average number of introduction events per day) and percent of introduction events leading to large clade sizes, time-periods 1-6

Table 3.3: Percent of introduction events leading to large clades, average introduction events per-day, and sampling proportion and intensity during each time-period.

Time-period	Percent of introduction events leading to clades >5	Percent of introduction events leading to clades >10	Average introductions per day	LTCF sampling deviation from community sampling (percentage points)	Percent of all LTCF cases sequenced
1	5%	1%	1.1	+9.6	30%
2	13%	9%	1.0	+8.3	18%
3	6%	4%	0.7	+1.0	5%
4	8%	3%	1.6	+6.1	23%
5	9%	4%	1.5	+0.9	11%
6	8%	2%	1.1	-6.8	5%

Among cases inferred to be associated with introduction clades sized >1, varying proportions were labeled as outbreak-associated in the epidemiologic dataset over time, ranging from 49.2%-97.4% (Table 3.4).

Table 3.4: Agreement of genomic and epidemiologic datasets.

Time-period	Proportion of cases inferred in LTCF post-introduction clades >1 and marked as outbreak-associated in epidemiologic datasets
1	56/63 (88.9%)
2	590/1050 (56.2%)
3	262/269 (97.4%)
4	323/382 (84.6%)
5	610/932 (65.5%)
6	223/453 (49.2%)

3.3.4 *Yakima County long-term care facility-associated transmission*

Yakima Health District reported supplemental data on 1,725 cases associated with ten facilities; 1,452 (84%) of these case records were linked to WDRS data by probabilistic matching. Genomic data were available for 667 cases. Sequenced cases from Yakima were highly representative based on age, sex, and race. Sequenced cases were more likely to be fatalities (11.1% of sequenced cases vs 8.1% of all facility cases).

Phylogenetic visualization spanned two time periods, which covered 98% of sequences: January-August 2020 and August 2021-December 2022. Several large facility-associated outbreaks were visualized; three facilities were selected for additional analyses (Figure 3.5).

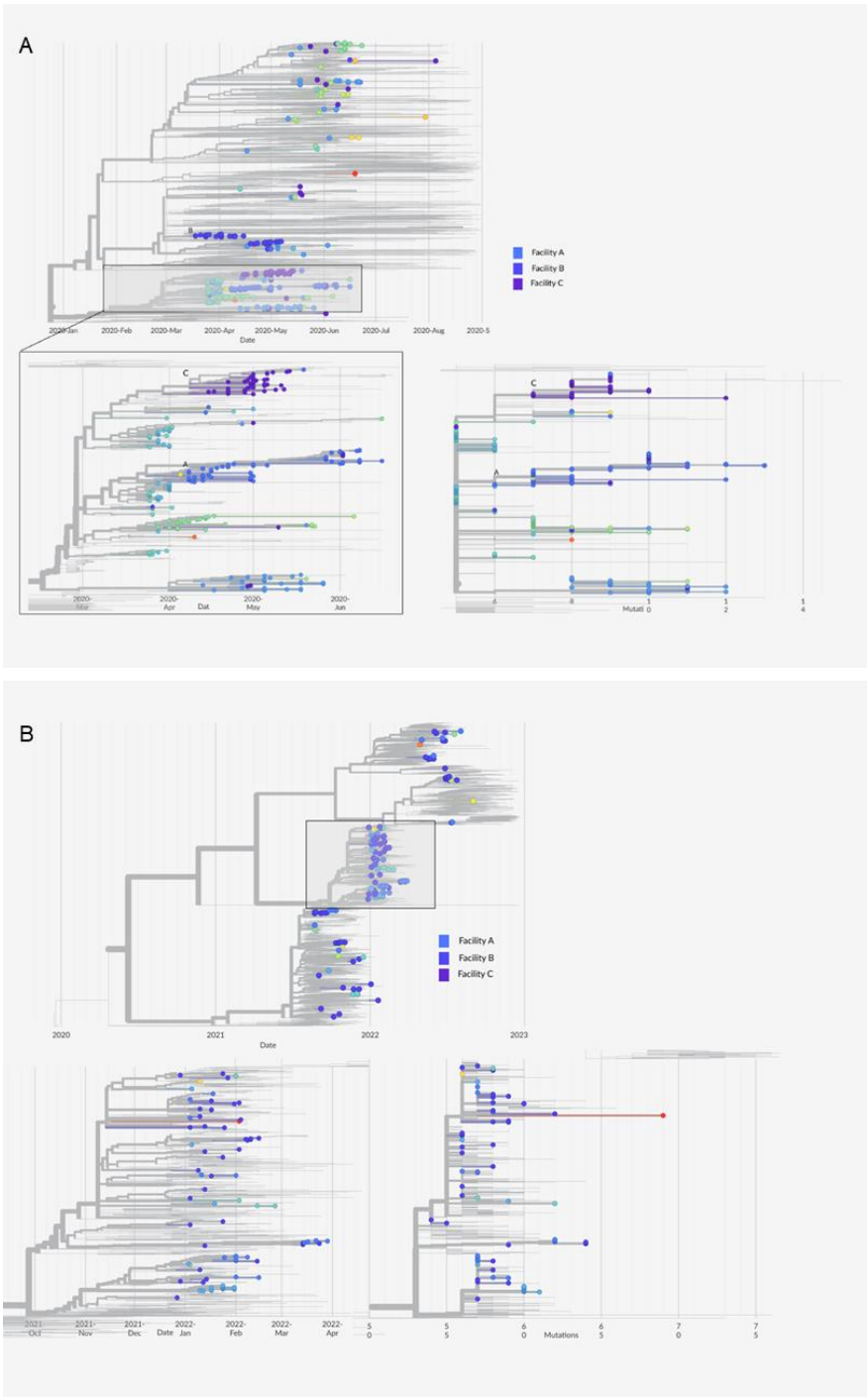


Figure 3.5: Time-scaled phylogenetic trees and divergence scaled phylogenetic trees, Yakima County.

Time-scaled phylogenetic tree and divergence scaled phylogenetic tree of sequence data from LTCF-associated cases, Yakima County, January-August 2020 (A), and time-scaled phylogenetic tree from LTCF-associated cases, Yakima County, August 2021-December 2022 (B). Nodes are colored by individual facility; colored nodes are LTCF-associated cases, gray nodes are contextual samples.

Facility A was selected due to identification of one prolonged cluster spanning April-June 2020; a divergence tree of each selected outbreak is shown in Figure 3.6. Facility B was selected due to two large overlapping outbreaks early in the pandemic with multiple introductions later in the pandemic. Facility C was selected due to apparent multiple introduction events over the course of the pandemic, including early in the pandemic. Resident and staff infections were interspersed across the tree and network visualizations.

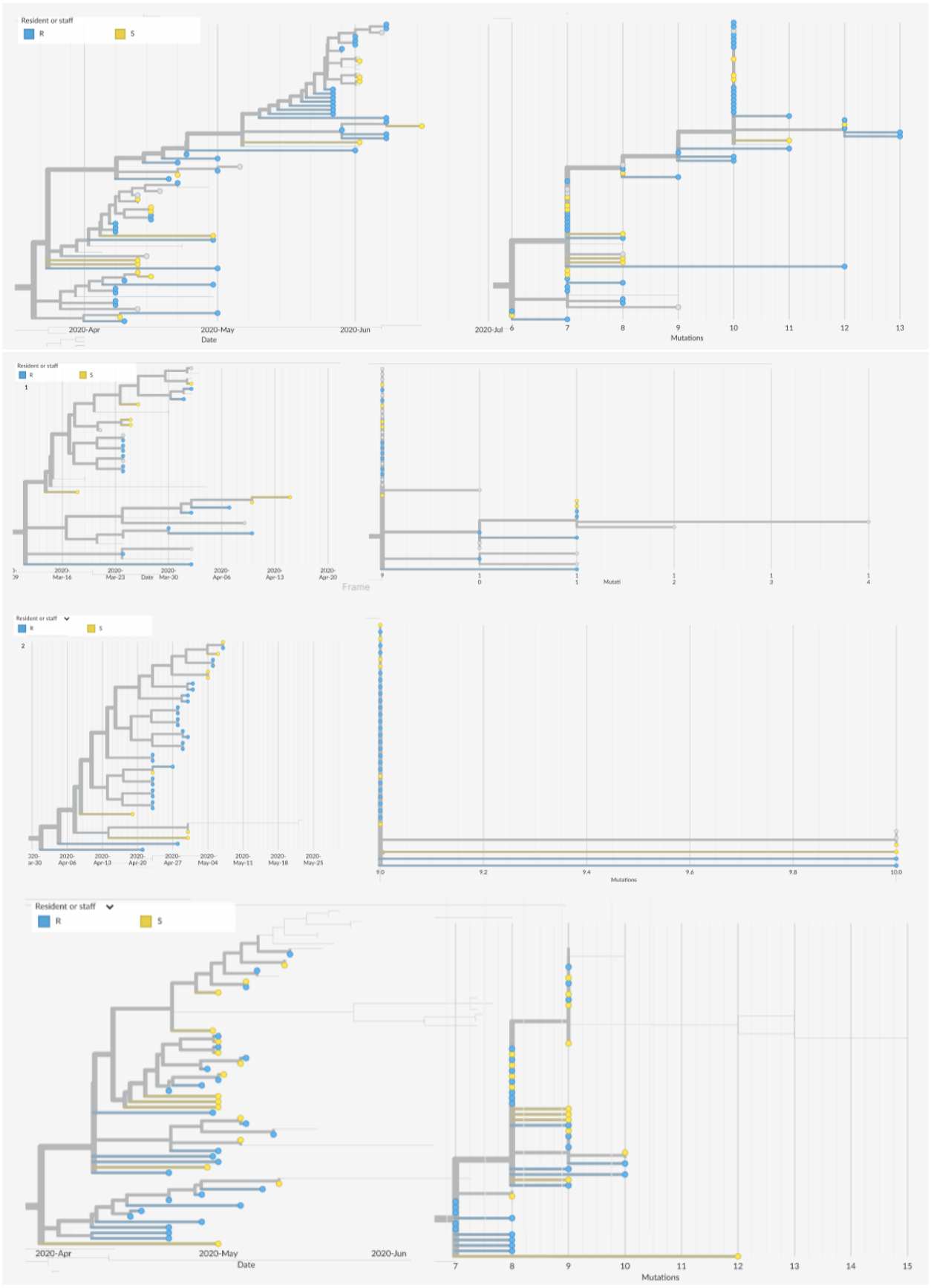

A
Resident or staff
R
S
2020-Apr
2020-May
2020-Jun
Date
2020-Jul
6
7
8
9
10
11
12
13
Mutations
B
Resident or staff
R
S
1
2020-
Mar-16
2020-
Mar-23
Date
2020-
Mar-30
2020-
Apr-06
2020-
Apr-13
2020-
Apr-20
Frame
Mutati
C
Resident or staff
R
S
2
2020-
Apr-06
2020-
Apr-13
2020-
Apr-20
2020-
Apr-27
2020-
May-04
2020-
May-11
2020-
May-18
2020-
May-25
9.0
9.2
9.4
9.6
9.8
10.0
Mutations
D
Resident or staff
R
S
2020-Apr
2020-May
2020-Jun
Date
7
8
9
10
11
12
13
14
15
Mutations

Figure 3.6: Time and divergences trees, Yakima County Facilities A-C.

Trees are colored by resident (blue) versus staff (yellow) classification.

A. Facility A, April-June 2020
B. Facility B-1, March-June 2020
C. Facility B-2, March-June 2020
D. Facility C, April-Aug 2020

Trace diagrams resulting from the TransPhylo analysis revealed uncertainty in the parameter values, likely due to preponderance of identical consensus genomes, impacting Transphylo's ability to resolve within- and between-case genetic diversity, as has been described previously for SARS-CoV-2 transmission reconstruction.[56]

The Facility A transmission reconstruction inferred 12% of cases as unsampled sources and inferred a resident as source. During this period, 56% of known cases from Facility A were sequenced (Table 3.5).

An outbreak spanning March 18, 2020 to April 15, 2020 included 27 Facility B sequences; during this period, 58% of known Facility B cases were sequenced. Another 33 sequences from this facility were associated with a separate outbreak spanning April 19, 2020 to May 7 2020.

From April-August 2020, 69% of reported cases from Facility C were sequenced and at least 18 separate introduction events were documented, only one of which apparently led to an outbreak of >5 cases as visualized in the genomic data. This outbreak included 62 sequences and spanned April 15-May 14, 2020.

The proportion of staff amongst all cases was consistent across these four outbreaks, ranging from 17%-22%. The ratio of observed to expected inferred transmission events attributed to staff ranged from 0.66-1.17, providing evidence that both staff and residents are driving transmission in these outbreaks (Table 3.5).

Table 3.5: Sampling and estimated staff contribution to analyzed outbreaks, Yakima

Facility/Outbreak	Sequencing proportion during period of interest	Transphylo inferred sampling	Proportion of staff among cases	Observed/expected transmission events attributed to staff
A	56%	88%	0.18	0.94
B-1	58%	95%	0.17	1.17
B-2	58%	95%	0.21	0.90
C	69%	79%	0.22	0.66

3.4 *Discussion*

Here, we analyzed epidemiologic and genomic data associated with LTCFs in WA to characterize transmission dynamics and inform ongoing data utilization. Transmission dynamics in LTCFs changed over the course of the COVID-19 pandemic, with variable introduction rates into LTCFs, but decreasing amplification within LTCFs. Particularly during March-August 2020, a period marked by little population immunity and initiation of non-pharmaceutical interventions, COVID-19 spread in LTCFs via high introduction rates and intra-facility transmission. The number of introduction events and intra-facility clade sizes decreased during August 2020-March 2021; vaccination campaigns began in December 2020. Additionally, CMS released testing requirements for staff and residents in August 2020. Although the introduction rate more than doubled between this time-period and the subsequent two study periods, the percentage of introduction events leading to large clade sizes remained stable. This indicates that despite more frequent introductions during these time periods, post-introduction within-LTCF transmission was curbed, possibly due to vaccination and improved IPC. These study periods were marked by transmission of Delta and Omicron variants, with high levels of community transmission likely contributing to introduction rates. While case counts were high, the genomic data show that incidence was largely driven by repeated introduction events rather than intensive within-LTCF spread.

Over the course of the pandemic, LTCF-associated cases are dispersed throughout the trees and intermixed with both LTCF-associated and non-LTCF cases, indicating that SARS-CoV-2 lineages circulating in LTCFs matched those circulating in surrounding communities. Dominant lineages in each time-period matched when comparing LTCF-associated cases to Washington cases included in the tree. This finding is consistent with a similar study performed in the UK.[50] Similarly, sequences from different age groups are interspersed, indicating likely bi-directional transmission between staff and residents. This observation was validated for a small number of outbreaks, demonstrating proportional inferred transmission from staff and residents. Interpretation of these findings is limited by variable sequencing over time. For much of the pandemic, testing and sequencing from LTCFs occurred at higher proportions than for the general population of COVID-19 cases. This over-sampling inflates the number of introductions and clade sizes when contextualized among other WA sequences. Changes in the relative proportion of LTCF cases sequenced and in sampling intensity are expected to impact findings of the DTA analysis and comparison across timepoints. However, when considering the direction of expected change, we anticipate the results identified herein are generally a conservative estimate. This conclusion was drawn after comparing the relative direction of change considering sampling proportion and sampling intensity across time-periods to the number of large clades identified. Overall, sequenced LTCF cases were found to be representative of COVID-19 cases in LTCFs.

The potential contribution of genomic data in defining outbreak-related cases was quantified. In the absence of genomic data, outbreak-association is determined using the current Council for State and Territorial Epidemiologists (CSTE) case definition. However, this definition cannot differentiate between concurrent but independent introduction events or outbreaks and relies on

epidemiologic data capture. Analysis of the agreement between outbreak-tagged cases in the epidemiological data and cases identified in post-introduction clades sized >1 revealed that epidemiologic data is growing more disparate from genomic data over time. Specifically, during periods 4-6, cases inferred within LTCF post-introduction clades were less likely to be recorded as outbreak-associated in the epidemiologic datasets compared to during study periods 1-3. This finding suggests that genomic data could greatly inform outbreak definitions, especially in settings of decreased epidemiologic data capture. In the absence of genomic data, outbreaks may also be over-estimated as multiple introduction events are not considered.

Although we attempted transmission reconstruction of four outbreaks in Yakima County, uncertainty in the parameter values limits interpretation of results. Indeed, based on known sequencing rates, TransPhylo estimated fewer missing links than expected and epidemiological data including onset dates provided conflicting results. Methods that utilize additional epidemiological data in reconstruction, such as extension of the outbreaker2 model, may be more useful in this setting[14,58].

Visualization of this large genomic dataset over time provides insight into useful bioinformatic tools and methods for application in public health practice. Early in the pandemic, many clusters of cases with long persistence were observed. Genomic epidemiology tools often rely on distance thresholds for defining clusters. These tools are difficult to apply in settings of prolonged transmission, as evolution over time is expected. Application of tools requiring thresholds may result in inference of independent clusters in situations of prolonged transmission. This was observed when attempting to use one such tool, MicrobeTrace, in the analysis of outbreaks in Yakima County. In this study, the utilization of DTA analysis with paired epidemiologic data allowed observation of prolonged outbreaks without the need for thresholds.

This study faced several important limitations. First, genomic data captured for LTCF-associated cases were associated with more severe cases. The majority of LTCF-associated outbreaks had no sequences available; this requires an assumption that the sampled LTCFs are representative of the unsampled facilities. Based on our case-level representativeness assessment, including proportional sampling by facility type, we believe this assumption is reasonable. The DTA analysis was performed using a binary variable for LTCF-association; analysis at the facility level may reveal additional introduction events and patterns of inter-facility spread. Demonstrating the relative rarity of large outbreaks caused by a single introduction late in the pandemic is an important finding; however, many guidance, policy, regulation, practice, immunity, and prevention method (including new availability of vaccines) changes occurred over the study period, prohibiting a causal analysis of which component changes led to this impact and limiting our study to observational findings.

This study had several notable strengths. First, we assessed genomic sampling representativeness at the case-level, enabling DTA analysis and interpretation. Second, paired epidemiologic and pathogen genomic data were available with additional detail available for Yakima County cases, facilitating in-depth analysis of transmission. In particular, the ability to de-duplicate sequences early in the pandemic impacted study findings; during the first time-period there were an average of three (triplicative) genomes available among sequenced cases. Analysis in the absence of epidemiologic data will over-represent these cases, inflating genomically-defined clusters. Finally, genomic studies to understand a single or a few outbreaks are commonly performed and reported in the literature. By looking at data over time, we add important context regarding the changing transmission dynamics associated with LTCFs.

Paired genomic and epidemiologic data enable phylogenetic analysis to understand transmission patterns, identify apparent clusters, and form hypotheses regarding transmission networks. However, metadata is not consistently available on some key variables, including type of LTCF association (staff/resident/visitor), dates of association, or travel history. Given currently available data, methods for tree building for hypotheses generation on a routine basis are recommended. Cluster detection tools for outbreak identification are likely of limited use, as most facilities do not have sequencing performed and data is not timely. However, cluster detection on available genomic data may help to identify temporal patterns of intra-facility spread versus repeated introduction. The current data types and quality captured by routine surveillance data collection is inadequate for applying methods to infer transmission or identify introduction sources with certainty. Although this data may be available through enhanced investigations in some counties, as with Yakima County, the general absence of this data limits broader analysis. Importantly, we noted a decrease in data capture from LTCFs over time. Depending on goals for use of genomic data, sentinel surveillance should be increased or targeted surveillance implemented to ensure available data for analysis; likewise, if cluster detection is a desired outcome, data timeliness should be improved.

These findings reflect challenges facing many SARS-CoV-2 genomic data capture systems presently. Antigen-based testing is common but is not compatible with available specimen retrieval practices and sequencing capacity; advances compatible with ongoing genomic data capture are needed. With present patterns of sequencing, LTCFs are underrepresented; expansion to sentinel facilities or during outbreak investigation is recommended. Additionally, genomic epidemiologic workforce capacity embedded within the teams that surveil for outbreaks in healthcare settings is required.

3.5 *Conclusions*

In conclusion, this analysis identified changing transmission dynamics in LTCFs over the course of the COVID-19 pandemic, with smaller post-introduction clades noted later in the study period despite periods of high introduction rates. This finding is encouraging for the many control efforts that have been put in place in these facilities over time, including vaccination, infection prevention, and testing and reporting to public health jurisdictions, although causal theories could not be tested and natural immunity was also accumulating during this time. LTCFs are likely to remain vulnerable institutions in which ongoing respiratory pathogen monitoring and outbreak control is warranted. Genomic data have the potential to increase the specificity of outbreak detection and resulting public health actions. Ongoing genomic epidemiologic analysis of LTCF-associated data is encouraged to facilitate situational awareness, potential cluster detection, and hypothesis-generation for further targeted analysis.

Chapter 4. A ONE HEALTH DATA INTEGRATION FRAMEWORK FOR REAL-TIME SURVEILLANCE AND APPLIED GENOMIC EPIDEMIOLOGY

4.1 *Introduction*

Traditional health surveillance systems in humans and animals, and of associated ecosystem events or impacts rely on sectorized and independent data systems, analysis platforms, and visualizations. The concept and approach outlined by One Health promotes the health and wellbeing of the planet and all living things, pushing our current systems beyond their siloed approaches to coordination and integration across the surveillance pathway. While surveillance systems may vary in their scope, objective, methods, and platforms, there are general commonalities across surveillance for health events, which include ongoing: 1) sample or data collection, 2) data storage and collation, 3) data analysis and interpretation, and 4) dissemination or outcome communication[1,59]. Moving from a single-sector surveillance system to a One Health surveillance system requires multi-sector coordination at a point or points along this surveillance pathway.

While many organizations, agencies, and authors have called for integration of surveillance systems across One Health, few examples of systems developed at the sub-national level, where response to health events generally occurs, are available. Those that exist are largely focused on a single condition or health hazard (e.g. West Nile virus, antimicrobial resistance) and are created in response to a known problem[60]. Systems that exist at the national or global scale suffer from reduced timeliness, completeness, and granularity, impeding response at the local level. Challenges that currently obstruct systems integration include data dispersion across many fields,

heterogeneous data collection methods, semantic interoperability, and complex data governance[19]. Jurisdiction and mandates differ across sectors, particularly public health, animal health, plant health, and environmental health and food safety[60]. Additionally, informatics capacity varies widely across systems, varying from paper data collection to complex systems with standardized reporting[19]. State and local governments often have aging data infrastructure, and an ongoing need for data modernization. Funding is vertically allocated with limited or no resources available for cross-sector work and scarce resources even within-sector[60].

Despite logistical, governance, and financial barriers, the development of integrated One Health data systems has potential to improve prevention and control efforts[18]. The ability to coordinate and integrate along the surveillance pathway, from data collection to dissemination, across knowledge domains could provide new insights into existing challenges, identify novel hypotheses related to health events, provide early warning for impending health disasters, and allow for integrated One Health solutions[61,62].

A similarly promising development for One Health surveillance is the use of pathogen genomic sequencing and analyses in support of infectious disease surveillance[63]. Pathogen genomic sequencing is host-agnostic, and phylogenetic analysis of resulting data allows for assessment of transmission dynamics at the human-animal-environment interface. This technology can be applied across bacterial, viral, fungal, and parasitic pathogens. A One Health implementation allows for early outbreak detection and improved understanding of pathogen reservoirs, evolution, and vehicles of transmission, enabling proactive prevention of One health threats[17]. However, challenges in the development and application of pathogen genomics surveillance systems remain, particularly in government institutions[3,17].

Use of an integrated approach to genomic epidemiology has been most commonly applied in the area of food-borne disease, with the collection of genomic and epidemiological data from human, veterinary, food, and environmental domains in systems such as PulseNet, GenomeTrakr, and NCBI in the United States and the EFSA One Health WGS System in the European Union[20,21]. The nature of food distribution systems require national or international coordination; however, data integration alone without joint analysis, interpretation, investigation, or intervention to improve health is not exemplary of a One Health approach – instead representing integration at a single level[60]. Implementing co-analysis requires building capacity in animal and environmental sectors for producing, sharing, and analyzing sequence data – including hiring bioinformaticians[2]. Expansion of a One Health approach to integrated genomic surveillance has not yet been widely extended to zoonotic or vector-borne disease pathogens, despite clear risk for impact of these pathogens on a global scale and clear benefits for understanding transmission at the human-animal-environmental interface.

Effective genomic-epidemiologic and traditional epidemiological analyses require compilation of data and metadata in a systematic way. Integration across domains to develop One Health surveillance systems, including for pathogen genomic data, requires implementation of emerging technologies and database infrastructures, including application programming interfaces (APIs), artificial intelligence (AI), machine learning (ML), and alternative data systems[3,19]. Application of these technologies could allow automated data collection from diverse sources, improved cross-domain analytics, and open access.

In Washington state government, One Health has been operationalized as a cross-agency collaborative – a group of agencies that meet quarterly to ensure ongoing collaborative relationships and communication[64]. Additionally, a One Health Surveillance and Data Systems

workgroup meets monthly to improve data sharing, integration, and visualization in support of One Health prevention and response. In 2022, this workgroup began discussing development of an integrated One Health surveillance system for Washington state. However, we lacked a framework for operationalizing data integration at the state level. To better understand the current landscape of One Health frameworks to guide surveillance system integration and co-analysis, we undertook a study of the existing literature. The overall objective of this work was to develop a conceptual framework, including concepts from both One Health and informatics and focusing on pathogen surveillance and genomic data integration, for One Health practitioners to utilize while implementing One Health data integration at the response-level.

4.2 *Methods*

This study used a mixed methods approach, combining a systematic literature review and structured interviews with purposively selected key informants representing One Health, informatics, and genomic epidemiology to understand existing frameworks and examples of cross-sectoral data integration in these domains. The framework development process consisted of four stages: 1) a review of the existing literature to draft an initial framework, 2) key informant interviews including review of the draft framework, 3) synthesis of information and design of the final framework, 4) key informant review of the final framework and revisions.

4.2.1 *Literature review search strategy and data extraction:*

We defined the search criteria to focus results related to “One Health” and “Data Systems” or “Data Integration” or “Genomic Data” or “Informatics” or “Digital Health” or “Surveillance System” and searched articles in PubMED and Web of Science using Medical Subject Headings (MeSH) terms. Following the structure provided in the PRISMA Statement, we organized the selection process in three phases: identification, screening, and inclusion[65]. One reviewer

conducted an initial screening on articles for relevance based on title and abstracts. Articles were screened-in if they included mention of One Health, and at least one intervention or outcome as outlined in Table 4.1. We excluded articles that describe an existing surveillance system or process, for example:

- Articles about application of genomic analyses to a One health problem
- Articles predominantly focused on describing or evaluating existing surveillance systems, without description of system development
- Articles predominantly focused on technological or research advances relevant to One Health (e.g. metagenomic sequencing or viral discovery)

Table 4.1: Scope of the literature review.

Population/Problem	Multi-sector (human/animal/environmental health), One Health
Intervention	Data systems development, data systems integration, framework for data integration
Outcome	Integrated surveillance, integrated data system, genomic data system, One health informatics, One Health surveillance system, framework

Following this, potentially relevant articles were downloaded and reviewed in full text by one reviewer. Reference lists of primary articles were further searched for additional studies. Additionally, we searched for relevant gray literature on the following agency websites: CDC, USAID, USDA, FDA, USGS, USDFW, EPA, ECDC, EFSA, One Health Commission, WHO, FAO, and WOAH. All relevant documents were downloaded in full text and underwent an eligibility assessment for inclusion by one reviewer.

We analyzed all included articles by collecting the following data elements: Title, authors, journal, year, article reports or discusses data system development (y/n), article reports or discusses data system integration (y/n), framework proposed for data integration (y/n), framework utilized for data integration (y/n). Articles were included in this study if yes to any of

the above questions. In addition, we captured whether articles included discussion of genomic data integration.

4.2.2 *Framework development*

Included articles were further reviewed for existence of frameworks related to data integration, lessons learned during data system development or integration, or for descriptions of framework implementation. We extracted these components and synthesized them into the design of a new framework for One Health practitioners to implement One Health data integration, with a focus on pathogen surveillance. A framework implementation guide was also developed to outline specific considerations for each framework step, as well as existing tools or references for each step.

Articles that included discussion of genomic data integration were identified and categorized as those that discussed genomic data generally versus those that gave examples or potential approaches for genomic data integration. Approaches were summarized, and a figure depicting summarized approaches was developed.

4.2.3 *Key informant interviews*

We developed two semi-structured tools to guide discussion for the key informant interviews. One tool focused on identification of existing One Health frameworks, examples of integrated data systems that exemplify this work, and a detailed review of the drafted framework and framework implementation guide. The second tool focused on identification of frameworks from informatics that may be applied to inform development of a novel framework and considerations for database structures. Purposive sampling was used to select the participants, based on expertise and involvement in One Health work or data systems and informatics work. We used these key informant interviews to supplement the literature review described above, to identify

additional examples of One Health data systems development, One Health data integration, or framework development or application for One Health data integration. Interviews were conducted during October-November 2023. A total of 19 individuals were invited via email to participate, 17 of whom agreed and were interviewed. All interviews were conducted one-on-one in English over video call. Feedback from these discussions was integrated into framework design, and an updated version of the framework was shared back to key informants for finalization.

4.3 ***Results***

4.3.1 *Literature Review*

The literature search identified a total of 1,515 records. Following screening, deduplication, and assessment of inclusion criteria, as well as review of references and gray literature review, 57 records were included in the final study (Figure 4.1).

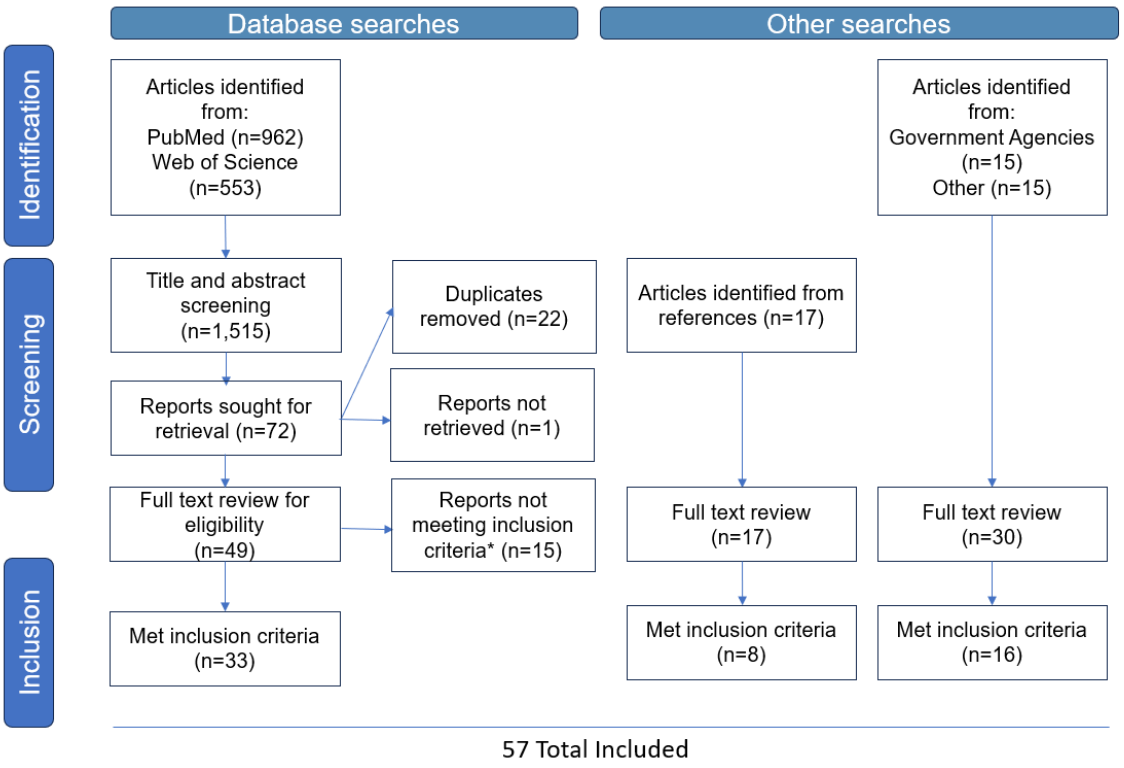


Figure 4.1: Identification, screening, and inclusion of articles from literature and gray literature searches.

From these 57 documents, we identified 2 broad health systems frameworks[66,67], 3 general One Health frameworks[68-70], and 8 One Health data integration-specific frameworks[18,59,60,71-75] (Table 4.2). The remaining 44 articles meeting inclusion criteria described integrated data system development[17,19,21,76-94], described general data system development/improvement[3,62,95-97], or discussed data system integration[20,61,98-112]. Although these articles did not define specific frameworks, the lessons learned during data system development or integration, or best practices in data integration were included into our final framework and framework implementation guide, which outlines specific questions and considerations for each framework step, as well as existing tools or references that may be useful (Supplementary Material).

Table 4.2: Health systems, general One Health, and One Health data integration-specific frameworks identified during literature review.

Name/Title	Reference	Description	Factors implemented into study framework
Health Systems Frameworks			
WHO Ten steps to systems thinking in the health system	[67]	Tool for applying systems thinking to health systems	Considered steps in development of final framework, notably including funding and a consideration of unexpected outcomes, as well as baseline and post-evaluation.
UNECE Generic Statistical Business Process Model	[66]	Generic process model for production of official statistics	Integrated steps from “specify needs” into system scoping.
General One Health Frameworks			
GOHF (Generalizable One Health Framework – CDC)	[68]	Generalizable One Health framework for the control of zoonotic disease	Ensured activities represented in the generalized framework were covered in developed framework.
Overarching One Health conceptual framework	[69]	Implementation cycle to inform One Health tool use	Integrated steps of cycle into overall framework structure.
OH-SMART	[70]	Multi-sectoral health system analysis and process improvement toolkit	Integrated OH-SMART steps into the planning/pre-funding steps of the final framework.
One Health Data Integration-Specific Frameworks			
Matrix Integrate-OHSS	[59]; https://ejp-matrix.eu/	A framework to develop a One Health surveillance System from an existing system	Integrated steps into overall framework structure, utilized multiple resources to develop implementation guide.
One Digital Health	[71,113]	A framework for future health ecosystems, joining the concepts of Digital Health and One Health	Integrated aspects of data standardization and interoperability, consideration of novel data sources
Socio-technical framework to develop common stakeholder vision for surveillance	[72]	A framework to help stakeholders develop a common vision of their desired surveillance system and forge the innovation pathway toward it	Consideration of strengthening existing surveillance capacities as part of the framework for integration. Framework steps 1-4 integrated in the final framework.
Standardized framework for data integration	[73]	Outlines essential data elements and a consistent reporting template, including mapping to SNOMED codes	Integrated into step 6 "Outline data design and user requirements."
A conceptual framework for organization of collaboration in a One Health surveillance system	[60]	Outlines organization factors conducive to sustainable collaboration, as well as aspects of collaboration supporting One Health surveillance systems	Integrated in the partner identification (1) and funding plan (5) steps of the framework.
A Tripartite Guide to Addressing Zoonotic Diseases in Countries Section 5.2.2	[18]	Outlines elements for establishing a comprehensive, coordinated system for surveillance and information sharing	Integrated outlined elements into overall framework.
Tripartite Surveillance and Information Sharing Operational Tool	[74]	Tool for establishing or strengthening a One Health multi-sectoral coordinated surveillance and information sharing (SIS) system for zoonotic diseases	Integrated outlined elements into overall framework, utilized multiple resources to develop implementation guide.
OHHLEP One Health surveillance system development framework	[75]	Outlines 6 steps to overcome barriers and optimize an integrated One Health Surveillance system	Integrated outlined elements into overall framework.

4.3.2 *One Health Data Systems Framework*

Common and unique elements of the above-identified frameworks were outlined; combined with the authors' own experience, these were integrated into a draft data integration framework, as shown in Table 4.2. One Health expert interviewees included 12 representatives of state agencies of public health, agriculture, and fish and wildlife, federal agencies of public health and agriculture, and university representatives. Five informatics expert interviewees represented state and federal public health agencies and university representatives. The final framework and framework implementation guide were developed following feedback from these interviews and final revision by the interviewees (Figure 4.2).

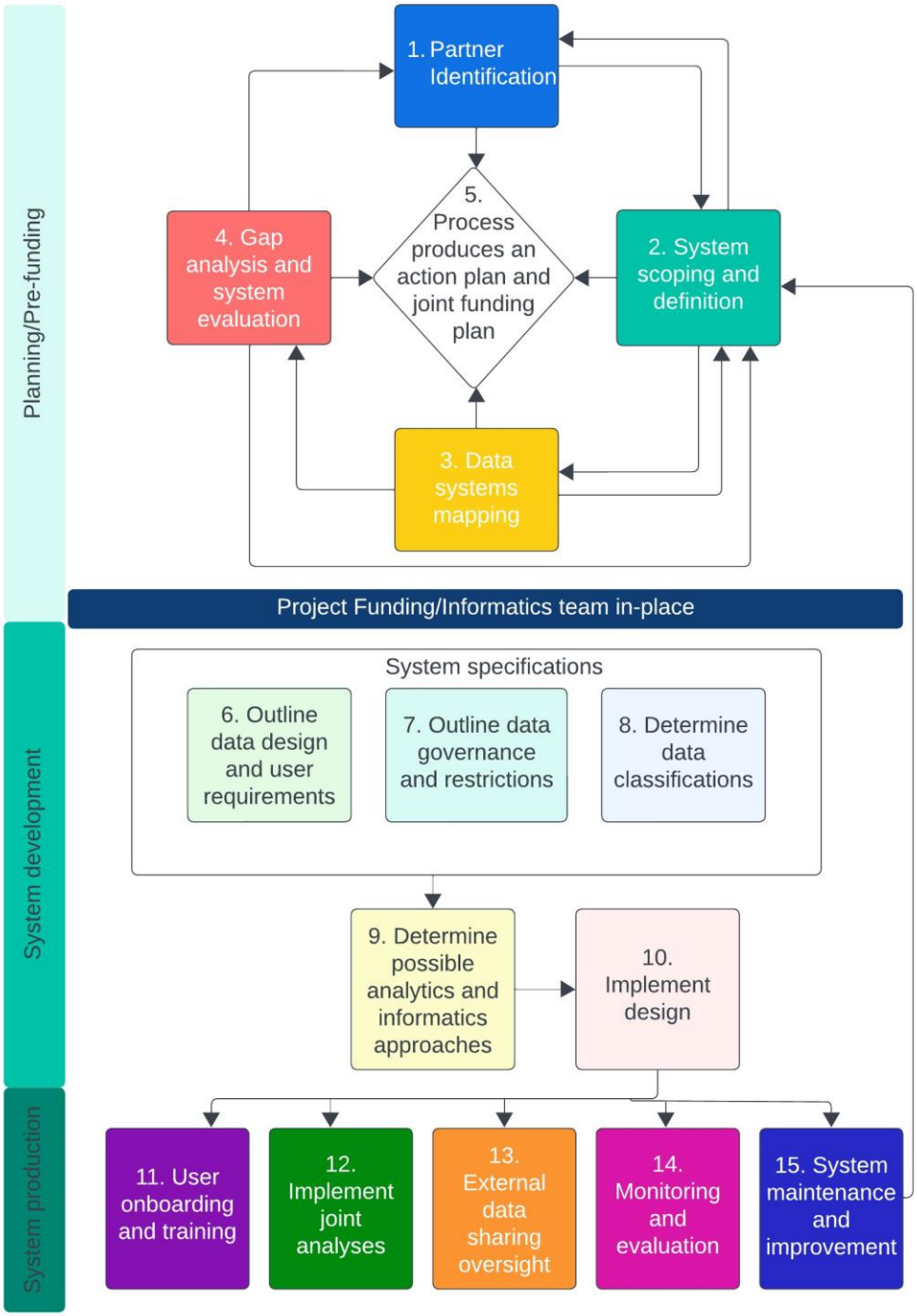


Figure 4.2: A One Health Systems Framework for Data Integration.

First, a workgroup is formed, considering participants from sectors that collect, analyze, and have governance over relevant data, as well as from different disciplines, the research community, public-private partnership, or community partners. This workgroup scopes the system to clearly define the purpose and outputs. Based on the specified scope, a data mapping process is performed

to understand what relevant data is available and whether the data timeliness, completeness, granularity, and quality support the scope. Current data structures, access, and connections are defined, and system capabilities (such as data export formats or capacity to send or receive standardized messaging) are outlined. Partner inclusion may be assessed throughout the process of data mapping, and system scope may need alterations dependent on available data. Based on the differences in the desired system and the mapped current system, gaps in data, data linkage, and data access are identified. Recommendations for improved data capture or sector-specific data systems to support a future integrated data system may be needed. An action and funding plan should be jointly developed throughout this process. A One Health data integration informatics team should be in-place prior to system development. In the system development phase, system specifications are outlined, including the data design, user requirements, data governance, and data security levels. Based on these system specifications, database structure options are considered, prioritizing modern data connections, limiting manual data manipulation, and future system needs for flexibility. The selected approach is used to identify potential system options. Once a system is in-place, the production phase should include user onboarding and training, system documentation, implementation of joint analyses, external data sharing oversight, ongoing monitoring and evaluation, and a plan for system maintenance and improvement. This last step may include revisiting system scope to add future capacity. See the framework implementation guide for additional detail.

Unique perspectives captured in the One Health expert interviews that altered the framework included the cyclical nature of the planning stage, with system scoping, data mapping, and gap analysis re-informing partner identification, and each step informing an action and funding plan. Consideration of data governance and external data sharing were highlighted as critical and additional considerations and resources on these topics were added to the implementation guide. A step for monitoring and evaluation was added to ensure attention to this important topic. A requirement for system flexibility and future re-scoping and improvements was highlighted. Additionally, a framework not previously identified through the literature or gray literature reviews was identified through expert interview: the Integrated Disease Surveillance and Response (IDSR) Section 9: Electronic Integrated Disease Surveillance and Response[114]. This framework was reviewed and incorporated into the framework implementation guide.

Informatics expert interviews pointed out the topic-agnostic common problems of data integration and interoperability, and commonalities of the planning stage to the work normally performed during business process mapping efforts. However, these efforts are generally occurring within the funded scope of an existing project at a single agency/institution. Within this One Health framework, funds are generally unavailable in the planning stages, often prohibiting informatics support, and the scope is not clearly defined at the outset. To ensure that adequate business process mapping is performed, working groups will likely need to revisit the scope, data mapping, and gap analysis after onboarding informaticians, prior to system specification. Importance was placed on clear identification of the problem to be solved and outlining the specific outputs of an integrated system. One key example of system scoping pointed to an approach whereby partners work through an exercise to agree on 20 questions the system should be able to answer, including enough specifics (time period, location), to inform data design[115]. The requirements inherently outlined in these questions can then be used to specify the system scope and work toward system specifications. Emphasis was placed on data collection and integration even in advance of standardization within this large of an integration effort to avoid getting stuck in system specification, as well as starting from raw data and including data transformation code within the system (e.g. using SQL). Finally, when considering flexibility for future system changes, consideration of changing infrastructure and technology should be included in addition to changing surveillance priorities.

4.3.3 *One Health Genomic Data Integration*

Genomic data integration was a topic considered by 10/57 articles (18%). Of these, 3 discussed the growth in genomic data production and analysis, the importance of considering genomic data when conducting integration, or the need for Standard Operating Procedures for laboratory and

bioinformatics efforts across One Health sectors[19,59,75]. Seven articles outlined examples or potential approaches for genomic data integration[3,17,21,62,77,84,87] (Table 4.3). Across these articles, common themes included pairing sequence data with a critical set of metadata according to a set of standards, standardization in quality checks and bioinformatics pipelines, and analyses across One Health sectors. Additional best practices included controlled data access to allow restricted and public views (data sharing and access principles), a process for data updates and corrections, connections of distributed databases through APIs, reproducible analyses, expanding the technical workforce across sectors, and open data sharing. In particular, the World Health Organization's "Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022-2023" calls for both leveraging genomics across One Health sectors and making the use of genomics routine in surveillance practice and disease prevention, preparedness, readiness, and response[62].

Table 4.3: Potential approaches for One Health genomic data integration.

Reference	Overview	Approach
[77]	Describes a shared secure surveillance platform between human and veterinary medicine in Switzerland	• Includes human, animal, environmental, and food isolates • Includes sequence data and associated metadata • Features controlled data access • Allow complex dynamic queries • Features dashboards • Automated data sharing with international repositories • Incoming data are quality- checked, curated, standardized where needed and processed/annotated with dedicated bioinformatics pipelines
[17]	Describes a shared platform for multisectoral data collection and bioinformatic analysis in Italy	• Includes isolates from food, environment, human and non-human and associated metadata • Bioinformatics tools sharing a common workflow system • Process for quality control
[21]	Outlines best practices for One Health contributions to open access databases	• Store the sequence with metadata • Inclusion of specimens from human, animals, food, and environment • Thresholds for QC • Contact information for submitters • Process for updating data, responding to requests, and correcting submissions
[84]	Describes a federated ecosystem as a possible solution for sharing genomic data	• Encourages use of federated databases • Connections of distributed databases through APIs
[87]	Describes a One Health system for hepatitis E virus surveillance in Europe	• Inclusion of human, animal, food, and environmental samples • Secure online environment • All sequence data with a restricted set of associated metadata become publicly available at a time specified by the data provider
[3]	Describes 10 recommendations for an informatic ecosystem to support pathogen genomic analysis in public health agencies	• Consistent data model (pairing sequence data with metadata) • Strengthen APIs to automate querying and analysis • Data management and stewardship • Bioinformatics pipelines open-source and accessible • Develop modular pipelines for data visualization and exploration • Improve the reproducibility of bioinformatics analysis • Utilize cloud computing • Expand the technical workforce • Improve the integration of genomic epidemiology with traditional epidemiology • Best practices to support open data sharing
[62]	Describes 5 objectives for genomic surveillance strengthening	• Improve access to tools for better geographic representation • Strengthen the workforce to deliver at speed, scale, and quality • Enhance data sharing and utility for public health decision-making and action • Maximize connectivity • Maintain a readiness posture for emergencies

Overall, there were two proposed structures that could be applied to One Health genomic epidemiologic data storage and analysis: the first in which the integrated database is configured to store full sequence data (either raw data or raw and assembled data) alongside metadata, with a process for standardized upload to public repositories; the second in which the integrated database stores only the public repository sequence identifiers alongside the metadata (Figure 4.3). In the former scenario, platforms allow for local analysis and visualization with controlled access but require substantial infrastructure and support[17,77,87]. In the latter scenario, sequence data from public repositories would need to be extracted and combined with metadata prior to analyses using tools external to the common platform[3]. In either scenario, ideal surveillance integration would include standardized assembly pipelines and quality checks across hosts, bioinformatics pipelines that process integrated One Health data, and joint visualization and interpretation across sectors.

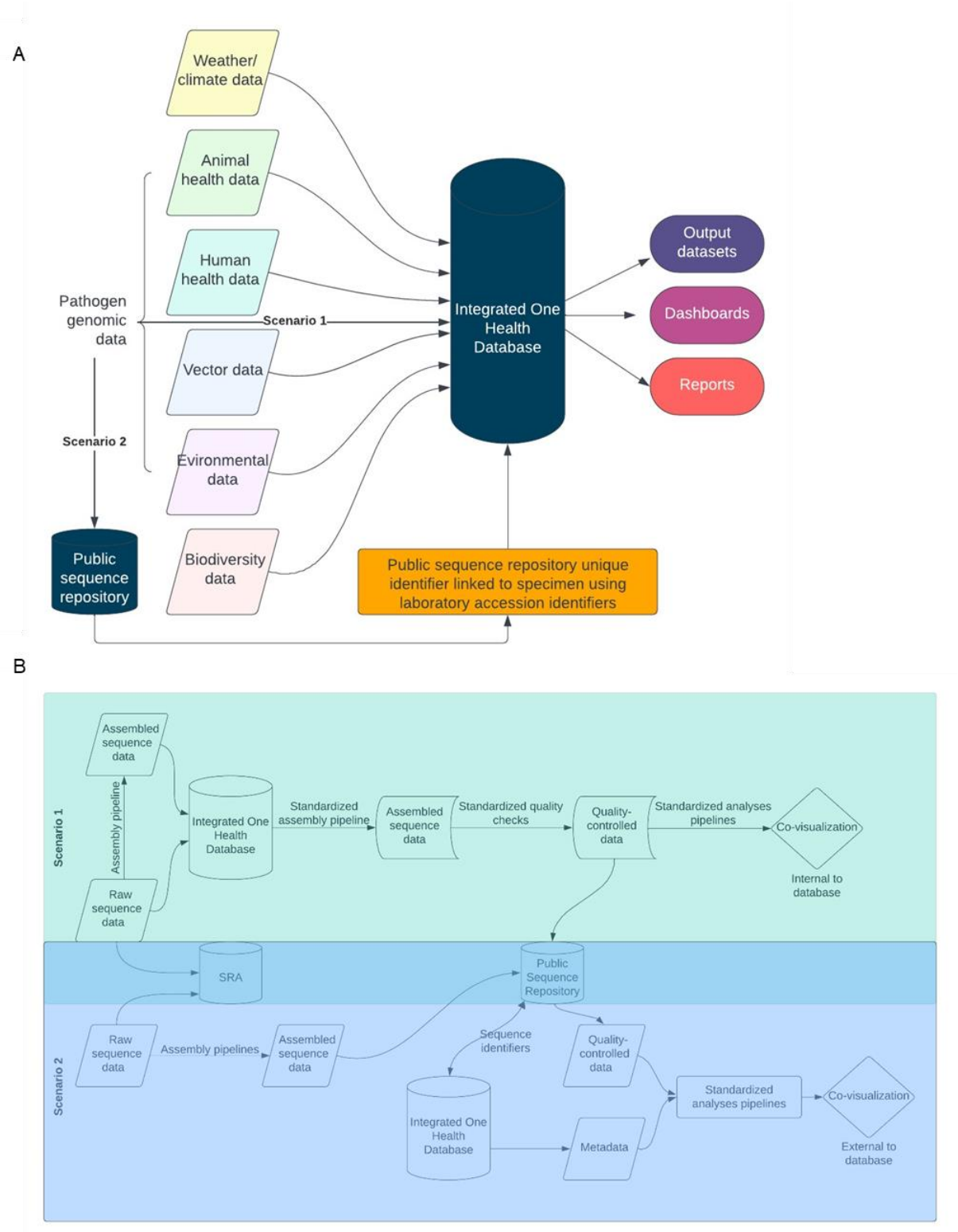


Figure 4.3: Example One Health database and scenarios for genomic data storage.

A) Example One Health Integrated Database, B) Elaboration of scenarios for genomic data storage and analysis. Scenario 1: sequence data is stored within the integrated One Health database, which

contains standardized assembly pipelines and quality checks as well as analyses pipelines and visualization tools. Scenario 2: sequence data is stored external to the integrated One Health database and linked through public sequence repository identifiers. To conduct analyses, sequence data is extracted from a public repository, combined with extracted data from the integrated One Health database, and analyzed using standardized pipelines. Analysis and visualization occurs external to the integrated One Health database.

4.4 *Discussion*

We undertook a search of the literature to inform development of a conceptual framework for One Health pathogen surveillance and genomic data integration. Although existing frameworks for One Health data integration were identified, none reflected the full scope of undertaking this work at the local or state level; similarly, none were identified as implemented for integration at this level. Combining the results of the literature review with the author's own experience in Washington's One Health Surveillance and Data Systems workgroup and expert interviews that included informaticians and One Health experts from inside and outside this workgroup enabled development of a One Health data integration framework that better captures elements required for start-to-finish implementation of this work. In particular, the overlay of key One Health considerations with generalized informatics frameworks represents a novel approach to framework development in this topic area. We believe this framework can be utilized by other jurisdictions seeking to undertake data integration at the response level.

We identified eight previously published One Health data integration-specific frameworks. Of these, two provided resources to inform the process and were the most informative to the proposed framework: the Matrix Integrate One Health Surveillance System (OHSS), which provided a step-by-step guide to creating a One Health Surveillance System from existing surveillance programs (https://ejp-matrix.eu/overview/)[59], and the Tripartite Zoonoses Guide Surveillance and Information Sharing Operational Tool (SIS OT)[74]. In particular, the data

mapping tool from OHSS provided useful considerations to assist jurisdictions in conducting this step of the process. This tool and other relevant tools are referenced throughout our framework implementation guide (Supplementary Material). Of these eight frameworks, only 3 considered the identification of resources as part of the framework[18,60,74]. None except SIS OT adopted feedback loops, instead proceeding in a stepwise fashion from start to finish. Most considered the steps reflected in our planning/pre-funding phase but did not consider steps following resource allocation. Developing these continuing steps of the framework gives jurisdictions a path beyond envisioning this system to move toward implementation. In addition to working toward data integration, One Health partners will need to consider how surveillance efforts supported by the data integration platform will support prevention and response. In parallel to data integration processes, jurisdictions should develop an integrated plan for conditions under surveillance, including how integrated data visualization and analyses will be utilized in support of prevention and response activities. This plan can help to inform the implementation of framework steps, including data design, analytics approaches, and implementation of joint analyses.

During the expert interviews, we discussed how this framework differed from one that would be commonly applied to general issues of data interoperability and systems design, such as a general data management framework or system development lifecycle model (SDLC). A general data management framework may consider elements such as data governance, data quality, data integration, data security, data privacy, data retention, data architecture, and data analytics. An SDLC encompasses planning, analysis, design, development, integration and testing, implementation, and maintenance. Although these generalized frameworks were reviewed to ensure consideration of all aspects across our framework, placing this framework within the One Health context requires several important adaptations. First, partner identification is

complexified by consideration of various sectors across human, animal, and environmental health that may be collectors, users, or interpreters of relevant data. The system scoping and definition is not an obvious or clearly defined problem as may often be the case within a simplified data integration problem, instead requiring engagement and co-design across sectors. Data systems mapping spans agencies and institutions and connections between sectors may not be apparent. Each of these steps may result in feedback loops to the previous steps. Although the planning stage may reflect the work commonly performed during business process analysis, this step is performed in advance of funding allocation and likely does not include a trained informatician or analyst to support the effort. Data governance conversations will be critical and will likely define parts of the system scope. Government agency IT requirements will limit possible informatics approaches. And finally, the requirement for planning for joint analysis, reporting, and interpretation across sectors is unique to the One Health model.

Genomic data integration was considered in one previous framework: the OHHLEP One Health surveillance system development framework acknowledges that design aspects should accommodate technological advances, such as whole genome sequencing. When discussing specific disciplines for implementation, laboratories are identified as an area where integration can occur both at the testing and analyses levels, including sequencing and bioinformatic, genomic, phylogenetic, and phenotypic analyses[75]. In addition, the One Health Surveillance Codex, which includes Integrate-OHSS, also includes a "Sequencing for Surveillance Handbook"[59,116]. This handbook is not referenced within Integrate-OHSS but is a separate tool. Consideration of different data types, including genomic data, is vital to system design and successful implementation. Failure to consider and include genomic data or data identifiers in a One Health system application may result in rapid obsolescence, as infectious disease

surveillance is increasingly reliant on genomic epidemiology to support surveillance and investigation.

We outline two scenarios identified in the literature for inclusion of genomic data in planning for an integrated One Health data system. In one, either raw or raw and assembled sequence data is included internal to the system, in the other, sequence identifiers are included as linkages without storage of the full sequence data. All implemented examples in the literature represented scenario 1. The potential benefits of this scenario were identified as internal standardization of assembly and quality checks, visualization of integrated data within a controlled-access system, and improved opportunities for joint analyses. However, substantial storage capacity to maintain full sequence data and infrastructure development for storage and computation are required. Computing resources available to the developed infrastructure may limit the types of analyses that could be performed. A potential modification of this scenario is the storage of assembled sequences only and the use of the Sequence Read Archive (SRA) for raw data storage; however, this removes the possibility of standardized assembly pipelines within the system which was outlined as a major benefit.

Scenario 2 potentially allows for more flexibility in selection of analyses and visualization tools and removes the storage burden of sequence data. However, this scenario requires establishing a process for linking sample identifiers (such as laboratory accession numbers) to repository unique identifiers to ensure linkage is maintained. In Washington state, scenario 2 has been implemented with respect to human genomic and epidemiologic data that are linked, and we foresee this as the preferred option for most databases, especially those developed in low-resource settings. Standardized quality checks are already in place in public repositories, and this option provides flexibility in case the amount of data exceeds storage capacity, as well as

flexibility in the selection of analyses and visualization tools. In many cases, access to raw sequence data may not be available, and data within public repositories may represent a wider capture of sequence data. However, one limitation of this scenario is the inability to perform standardized assembly, requiring additional coordination at the laboratory and bioinformatics levels. Integration at the laboratory level provides opportunities for increased efficiency and cost-effectiveness, while removing barriers to data integration.

One key requirement of integrated genomic epidemiologic analysis of emergent surveillance data is the availability of a platform for joint analysis and visualization within a controlled-access system. Indeed, this requirement appeared to drive the selection of scenario 1 in most if not all instances identified. Alternatives to developing additional systems for visualization and analysis, while allowing for controlled access include shared analysis files for visualization with Nextstrain auspice (https://auspice.us/), developing a Nextstrain group (https://nextstrain.org/groups/), or using an alternative visualization platform with controlled access, such as Data Flow and MicroReact[9,12]. Each of these alternatives come with their own challenges for process development that must be considered.

This review is subject to several limitations. The search strategy for inclusion of articles focused on those containing reference to One Health; therefore, this review does not provide an exhaustive overview of cross-sector data integration frameworks. A single author performed article screening and review for inclusion; review by multiple authors may have led to identification, inclusion, or exclusion of additional articles that may be relevant to this study. The focus of this review targeted pathogen surveillance and genomic data integration – there are also many health conditions unrelated to pathogens that require a One Health approach. Although we did not specifically address these conditions, and the extended need for novel data sources such

as from the social or behavioral sciences, this framework may likely be extended for application to non-infectious sources. Likewise, our work focused on the development of an integrated system from existing siloed systems; an ideal surveillance system likely includes elements not yet in existence such as animal health syndromic surveillance, citizen science reporting across sectors, and robust veterinary laboratory data reporting. Development of new primary systems was not included outside of consideration of areas for improved primary data capture or data systems.

Many of the identified frameworks included overlapping elements; indeed, although we re-conceptualized the planning stage of our framework as cyclical instead of stepwise, these elements were reflected across other frameworks. In addition to the cyclical nature of the planning stage, the novelty of our framework is in the expansion to system development and production, providing a framework for the path forward to implementation, and in the overlay of informatics frameworks and concepts. We emphasize the importance of considering multi-level data, including plans for pathogen genomic data early in the process to ensure a holistic One Health surveillance approach that recognizes the increasing importance of genomic epidemiology in infectious disease surveillance methods. The development of this framework makes it clear that a technical workforce with expertise in informatics is required to support the development of One Health systems. The lack of such available expertise devoted to One Health work may partially explain the dearth of integrated systems at the response level. Similarly, government systems are often outdated and do not often make use of technologies like artificial intelligence and machine learning. These technologies hold great promise for overcoming challenges such as semantic interoperability, data mapping, and data integration[19]. Modern data architecture should be considered key in the sustainable development of integrated One Health

data systems. To ensure not only the development of integrated systems, but also their sustainability, this framework outlines developing common goals, strong governance, and routine coordination and communication[18]. In addition, policy or legislation change is considered at multiple steps, to improve the landscape for data collection, data sharing, and process support. In Washington State, this work was exemplified during recent efforts to improve data sharing between the Washington State Department of Agriculture and the Washington State Department of Health; in addition to co-creation of a new process for data collection and cross-reporting, the Washington Administrative Code was updated to require reporting of animal diseases of public health concern.

4.5 *Conclusions*

Conducting real-time surveillance and response using the One Health approach requires a range of expertise both across One Health sectors and across disciplines, such as epidemiology, veterinary medicine, genomics, bioinformatics, and laboratory sciences. Benefits are gained not just from combining data, but from conducting joint analyses, bringing together a sufficient range of expertise to improve early detection and response[117]. This One Health data systems framework will help jurisdictions to operationalize this work at the response-level, moving past envisioning a system to allow implementation of systems development, leading to joint analyses and response. The framework's focus on real-time pathogen surveillance and genomic data integration reinforces, modernizes, and expands our joint ability to prevent and control disease for the health of humans, animals, and the environment we share.

Chapter 5. CONCLUSION

Throughout these chapters, I have explored initiation and evaluation of the first sentinel surveillance system for genomic data in Washington State, application of genomic-epidemiologic methods to a dataset produced by this system for public health action, and expansion of data integration to encompass One Health domains. In Chapter 2, I highlight the pressing need for public health surveillance systems to modernize to include integration of pathogen genomic data in real-time to inform public health response. As this work is initiated and developed across the country, attention should be paid to sampling methods and representativeness. Common approaches involve convenience sampling or collection of data primarily produced for research purposes. However, to ensure generalizability and equity when using paired genomic and epidemiologic data for public health purposes, the methods for capturing this data must ensure representativeness of the underlying population of interest. Beyond these concerns, there are limitations to the interpretation of results from maximum likelihood-based methods for analysis when random sampling is assumed and not tested. Genomic epidemiology training is not yet widespread in the public health workforce and public health practitioners should be cognizant of data and interpretation limitations. To our knowledge, this paper represents the first formal surveillance evaluation of a SARS-CoV-2 genomic epidemiological surveillance system and introduces methods both for implementation and evaluation of such a system. These methods could be modified and expanded by other public health jurisdictions to similarly assess genomic surveillance, with a goal of improved data collection and reliability of resulting findings across the country.

In Chapter 3, I apply genomic epidemiologic methods and tools to understand the spread of SARS-CoV-2 in long-term care facilities (LTCF) over the course of the pandemic. I identify

changing transmission dynamics in LTCFs, with smaller post-introduction clades noted later in the study period despite periods of high introduction rates. This finding is encouraging for the many control efforts that have been put in place in these facilities over time, including vaccination, infection prevention, and testing and reporting to public health jurisdictions. Importantly, I identify gaps in outbreak detection and highlight the potential contribution of combining genomic and epidemiologic data in determining whether cases are outbreak-associated. Genomic data thereby have the potential to increase specificity of public health actions. Ongoing use of genomic epidemiologic methods at the population-level can facilitate situational awareness, cluster detection and differentiation, and hypothesis-generation of further targeted analysis. At this time, tools for conducting such analyses have not been well-incorporated across communicable disease programs; by clearly demonstrating their potential use, this work supports wide-scale adoption in public health practice.

Finally, in Chapter 4, I consider expansion of data integration and genomic epidemiologic analysis concepts to One Health domains. For many emerging and endemic pathogens, risk and transmission dynamics are dependent not just on human host factors, but on distribution of animal reservoirs or vectors, environmental conditions, and interactions at the human-animal-environment interface. Only through joint analysis across One Health sectors can the full picture of pathogen evolution, transmission dynamics, and possible points of intervention be understood. A One Health implementation of genomic epidemiology allows for early warning of impending One Health events, promotes identification of novel hypotheses and insights, and supports integrated One Health solutions. In this chapter, I propose a novel framework, including concepts from both One Health and informatics, for One Health practitioners to utilize while implementing One Health data integration. The framework's focus on real-time pathogen

surveillance and genomic data integration reinforces, modernizes, and expands our joint ability to prevent and control disease for the health of humans, animals, and the environment we share. The work to develop the infrastructure and workforce required for comprehensive integration of genomic data into public health practice, and even beyond to encompass all One Health domains, requires significant investment. Demonstrating the value of genomic data to public health action can support future funding and policy changes for this work. Herein, I outline a possible future state for this field, through the development of an integrated surveillance system, ongoing use of integrated data for public health action, and envisioning expansion to One Health domains. I implement genomic-epidemiologic methods in public health practice and demonstrate possible impacts to public health action. Integration of pathogen genomic data into infectious disease surveillance and response promises to improve our understanding of health risks, disease distribution, and application of targeted interventions. Use of paired genomic and epidemiologic data is the future of the field; this work pushes public health and other One Health domains forward toward this expanded ability to prevent and control disease for the health of humans, animals, and the environment we share.

BIBLIOGRAPHY

1. Armstrong, G.*, et al.* Updated Guidelines for Evaluating Public Health Surveillance Systems Recommendations from the Guidelines Working Group. 1-35 (2001).
2. Armstrong, G.L.*, et al.* Pathogen Genomics in Public Health. (2019).
3. Black, A., MacCannell, D.R., Sibley, T.R. & Bedford, T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature Medicine* **26**, 832-841 (2020).
4. Ferdinand, A.S.*, et al.* An implementation science approach to evaluating pathogen whole genome sequencing in public health. *Genome Medicine* **13**(2021).
5. De Maio, N., Wu, C.H., O'Reilly, K.M. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics* **11**(2015).
6. Black, A.D., G. An applied genomic epidemiological handbook. (2023).
7. UCSC. Ultrafast Sample Placement on Existing Tree. (2020).
8. Aksamentov, I., Roemer, C., Hodcroft, E. B., & Neher, R. A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software* **6**, 3773 (2021).
9. Hadfield, J.*, et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).
10. Bouckaert, R.*, et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* **15**, e1006650 (2019).
11. Campbell, E.M.*, et al.* MicrobeTrace: Retooling molecular epidemiology for rapid public health response. *PLOS Computational Biology* **17**, e1009300 (2021).
12. Argimón, S.*, et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics* **2**(2016).
13. Sagulenko, P., Puller, V. & Neher, R.A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042 (2018).
14. Campbell, F.*, et al.* outbreaker2: a modular platform for outbreak reconstruction. *BMC Bioinformatics* **19**, 363 (2018).
15. Didelot, X., Fraser, C., Gardy, J., Colijn, C. & Malik, H. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution* **34**, 997-1007 (2017).
16. WHO. One Health. (2017).
17. Knijn, A.*, et al.* IRIDA-ARIES Genomics, a key player in the One Health surveillance of diseases caused by infectious agents in Italy. *Front Public Health* **11**, 1151568 (2023).
18. FAO, O., WHO. Taking a Multisectoral, One Health Approach: A Tripartite Guide to Addressing Zoonotic Diseases in Countries. (2019).
19. Ho, C.W.L. Operationalizing "One Health" as "One Digital Health" Through a Global Framework That Emphasizes Fair and Equitable Sharing of Benefits From the Use of Artificial Intelligence and Related Digital Technologies. *Frontiers in Public Health* **10**(2022).
20. Aarestrup, F.M., Bonten, M. & Koopmans, M. Pandemics- One Health preparedness for the next. *Lancet Reg Health Eur* **9**, 100210 (2021).
21. Timme, R.E.*, et al.* Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. *One Health Outlook* **2**, 20 (2020).
22. Klaucke, D.N.B.J.W.T.S.B.P.R.G.T.F.L.B.R.L. Guidelines for Evaluating Surveillance Systems. *MMWR* **37**, 1-18 (1988).
23. Paul, P.*, et al.* Genomic Surveillance for SARS-CoV-2 Variants Circulating in the United States, December 2020-May 2021. *MMWR Recommendations and Reports* **70**, 846-850 (2021).
24. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 2-4 (2017).

25. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33-46 (2017).
26. Khare, S.*, et al.* GISAID's Role in Pandemic Response. *China CDC Weekly* **3**, 1049-1051 (2021).
27. Bedford, T.*, et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science (New York, N.Y.)* **370**, 571-575 (2020).
28. Jorden, M.A.*, et al.* Evidence for Limited Early Spread of COVID-19 Within the United States, January–February 2020. *MMWR. Morbidity and Mortality Weekly Report* **69**, 680-684 (2020).
29. Fauver, J.R.*, et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990-996.e995 (2020).
30. Tordoff, D.M.*, et al.* Phylogenetic estimates of SARS-CoV-2 introductions into Washington State. *The Lancet Regional Health - Americas* **1**, 100018-100018 (2021).
31. Müller, N.F.*, et al.* Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. *Science Translational Medicine* **13**, 1-12 (2021).
32. Magee, D. & Scotch, M. The effects of random taxa sampling schemes in Bayesian virus phylogeography. *Infection, Genetics and Evolution* **64**, 225-230 (2018).
33. Lemey, P.*, et al.* Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathogens* **10**(2014).
34. State, W. SARS-CoV-2 Sequencing and Variants in Washington State. (2022).
35. Team, R.C. R: A language and environment for statistical computing. (Vienna, Austria, 2020).
36. Influenza Virologic Surveillance Right Size Roadmap.
37. Wohl, S., Lee, E.C., DiPrete, B.L. & Lessler, J. Sample Size Calculations for Variant Surveillance in the Presence of Biological and Systematic Biases. *medRxiv*, 2021.2012.2030.21268453-21262021.21268412.21268430.21268453 (2022).
38. Washington State Department of, H. COVID-19 Long Term Care Monthly Report. (2023).
39. McMichael, T.M.*, et al.* Epidemiology of Covid-19 in a Long-Term Care Facility in King County, Washington. *New England Journal of Medicine* **382**, 2005-2011 (2020).
40. McMichael, T.M.*, et al.* COVID-19 in a Long-Term Care Facility — King County, Washington, February 27–March 9, 2020. *MMWR. Morbidity and Mortality Weekly Report* **69**, 339-342 (2020).
41. Kimball, A.*, et al.* Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility — King County, Washington, March 2020. *MMWR. Morbidity and Mortality Weekly Report* **69**, 377-381 (2020).
42. Cdc. Preparing for COVID-19 in Nursing Homes | CDC.
43. Washington State Department of, H. COVID-19 Infection Prevention in Health Care Settings. (2023).
44. Inslee, J. New Release Archive. (2023).
45. Quality, C.f.C.S.a. Interim Final Rule (IFC), CMS-3401-IFC, Additional Policy and Regulatory Revisions in Response to the COVID-19 Public Health Emergency related to Long-Term Care (LTC) Facility Testing Requirements. (ed. Group, S.C.) (2020).
46. Oltean, H.N.*, et al.* Sentinel Surveillance System Implementation and Evaluation for SARS-CoV-2 Genomic Data, Washington, USA, 2020–2021. *Emerging Infectious Diseases* **29**, 242-251 (2023).
47. Douglas, P.*, et al.* 1374. Utilizing Genomic Epidemiology to Explore SARS CoV-2 Transmission Patterns and Support Outbreak Investigations in Long Term Care Facilities, Washington State, April-October 2021. *Open Forum Infectious Diseases* **9**, ofac492.1203 (2022).
48. Arons, M.M.*, et al.* Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. *The New England Journal of Medicine* **382**, 2081-2090 (2020).
49. Aggarwal Mrcp, D.*, et al.* The role of viral genomics in understanding COVID-19 outbreaks in long-term care facilities. *Review Lancet Microbe* **3**, 151-158 (2022).

50. Hamilton, W.L.*, et al.* Genomic epidemiology of COVID-19 in care homes in the east of England. *eLife* **10**(2021).
51. Kathryn Turner, S.L.D., Jim Collins, Caitlin S. Pedati, Ruth Lynfield, Sharon M. Watkins, Bernadette Albanese, Zack Moore, Audrey Kunkes, Lisa McHugh. Update to the standardized surveillance case definition and national notification for SARS-CoV-2 infection (the virus that causes COVID-19). (2022).
52. Team, R.C. R: A language and environment for statistical computing. R Foundation for Statistical Computing. (Vienna, Austria, 2022).
53. Dudas, G. Backronymed adaptable lightweight tree import code for molecular phylogeny manipulation, analysis, and visualization (baltic). (2016).
54. Didelot, X., Kendall, M., Xu, Y., White, P.J. & McCarthy, N. Genomic Epidemiology Analysis of Infectious Disease Outbreaks Using TransPhylo. *Curr Protoc* **1**, e60 (2021).
55. Perera, D*., et al.* Reconstructing SARS-CoV-2 infection dynamics through the phylogenetic inference of unsampled sources of infection. *PLoS One* **16**, e0261422 (2021).
56. Gallego-Garcia, P*., et al.* Limited genomic reconstruction of SARS-CoV-2 transmission history within local epidemiological clusters. *Virus Evol* **8**, veac008 (2022).
57. Wang, L*., et al.* Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nat Commun* **11**, 5006 (2020).
58. Hjorleifsson, K.E*., et al.* Reconstruction of a large-scale outbreak of SARS-CoV-2 infection in Iceland informs vaccination strategies. *Clin Microbiol Infect* **28**, 852-858 (2022).
59. Filter, M*., et al.* One Health Surveillance Codex: promoting the adoption of One Health solutions within and across European countries. *One Health* **12**, 100233 (2021).
60. Marion Bordier, T.U.-A., Aurélie Binot, Pascal Hendrikx, Flavie L. Goutard. Characteristics of One Health surveillance systems: a systematic literature review. *Prev. Vet. Med.* **181**, 104560 (2020).
61. Aenishaenslin, C*., et al.* Evaluating the Integration of One Health in Surveillance Systems for Antimicrobial Use and Resistance: A Conceptual Framework. *Front Vet Sci* **8**, 611931 (2021).
62. WHO. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022–2032. (2022).
63. Urban, L*., et al.* Real-time genomics for One Health. *Mol Syst Biol*, e11686 (2023).
64. Washington State Department of, H. One Health.
65. Page, M.J*., et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
66. Europe, U.N.E.C.f. Generic Statistical Business Process Model. (2009).
67. WHO. SYSTEMS THINKING for Health Systems Strengthening. (2009).
68. Ghai, R.R*., et al.* A generalizable one health framework for the control of zoonotic diseases. *Scientific Reports* **12**(2022).
69. Pelican, K*., et al.* Synergising tools for capacity assessment and One Health operationalisation. *Rev Sci Tech* **38**, 71-89 (2019).
70. Vesterinen, H.M*., et al.* Strengthening multi-sectoral collaboration on critical health issues: One Health Systems Mapping and Analysis Resource Toolkit (OH-SMART) for operationalizing One Health. *PLoS One* **14**, e0219197 (2019).
71. Benis, A., Tamburis, O., Chronaki, C. & Moen, A. One Digital Health: A Unified Framework for Future Health Ecosystems. *Journal of Medical Internet Research* **23**, e22189-e22189 (2021).
72. Bordier, M*., et al.* Engaging Stakeholders in the Design of One Health Surveillance Systems: A Participatory Approach. *Front Vet Sci* **8**, 646458 (2021).
73. Shanbehzadeh, M., Nopour, R. & Kazemi-Arpanahi, H. Designing a standardized framework for data integration between zoonotic diseases systems: Towards one health surveillance. *Informatics in Medicine Unlocked* **30**(2022).

74. Tripartite, O.H. Surveillance and Information Sharing Operational Tool. (2022).
75. Hayman, D.T.S*., et al.* Developing One Health surveillance systems. *One Health* **17**(2023).
76. Raymond, K*., et al.* Informatics progress of the Global Burden of Animal Diseases programme towards data for One Health. *Rev Sci Tech* **42**, 218-229 (2023).
77. Neves, A*., et al.* The Swiss Pathogen Surveillance Platform - towards a nation-wide One Health data exchange platform for bacterial, viral and fungal genomics and associated metadata. *Microb Genom* **9**(2023).
78. Karimuribo, E.D*., et al.* A Smartphone App (AfyaData) for Innovative One Health Disease Surveillance from Community to National Levels in Africa: Intervention in Disease Surveillance. *JMIR Public Health Surveill* **3**, e94 (2017).
79. Uchtmann, N., Herrmann, J.A., Hahn, E.C., 3rd & Beasley, V.R. Barriers to, Efforts in, and Optimization of Integrated One Health Surveillance: A Review and Synthesis. *Ecohealth* **12**, 368-384 (2015).
80. Mremi, I.R., Rumisha, S.F., Sindato, C., Kimera, S.I. & Mboera, L.E.G. Comparative assessment of the human and animal health surveillance systems in Tanzania: Opportunities for an integrated one health surveillance platform. *Glob Public Health*, 1-17 (2022).
81. Bordier, M*., et al.* Antibiotic resistance in Vietnam: moving towards a One Health surveillance system. *BMC Public Health* **18**, 1136 (2018).
82. Jato-Espino, D., Mayor-Vitoria, F., Moscardo, V., Capra-Ribeiro, F. & Bartolome Del Pino, L.E. Toward One Health: a spatial indicator system to model the facilitation of the spread of zoonotic diseases. *Front Public Health* **11**, 1215574 (2023).
83. Pley, C., Evans, M., Lowe, R., Montgomery, H. & Yacoub, S. Digital and technological innovation in vector-borne disease surveillance to predict, detect, and control climate-driven outbreaks. *Lancet Planet Health* **5**, e739-e745 (2021).
84. Health, T.G.A.f.G.a. A federated ecosystem for sharing genomic, clinical data. *Science* **352**, 1278-1280 (2016).
85. Ope, M*., et al.* Regional initiatives in support of surveillance in East Africa: The East Africa Integrated Disease Surveillance Network (EAIDSNet) Experience. *Emerg Health Threats J* **6**(2013).
86. Wendt, A., Kreienbrock, L. & Campe, A. Zoonotic disease surveillance--inventory of systems integrating human and animal disease information. *Zoonoses Public Health* **62**, 61-74 (2015).
87. Mulder, A.C*., et al.* HEVnet: a One Health, collaborative, interdisciplinary network and sequence data repository for enhanced hepatitis E virus molecular typing, characterisation and epidemiological investigations. *Euro Surveill* **24**(2019).
88. Pandit, N. & Vanak, A.T. Artificial Intelligence and One Health: Knowledge Bases for Causal Modeling. *J Indian Inst Sci* **100**, 717-723 (2020).
89. Ecowas. Promoting One Health Approaches through Integrated Data Systems. (2018).
90. European Food Safety, A*., et al.* Coordinated surveillance system under the One Health approach for cross-border pathogens that threaten the Union - options for sustainable surveillance strategies for priority pathogens. *EFSA J* **21**, e07882 (2023).
91. Kaur, J*., et al.* ICMR's Antimicrobial Resistance Surveillance system (i-AMRSS): a promising tool for global antimicrobial resistance surveillance. *JAC Antimicrob Resist* **3**, dlab023 (2021).
92. Leandro, A.S*., et al.* The adoption of the One Health approach to improve surveillance of venomous animal injury, vector-borne and zoonotic diseases in Foz do Iguacu, Brazil. *PLoS Negl Trop Dis* **15**, e0009109 (2021).
93. Falzon, L.C*., et al.* One Health in Action: Operational Aspects of an Integrated Surveillance System for Zoonoses in Western Kenya. *Front Vet Sci* **6**, 252 (2019).

94. McIntyre, K.M.*, et al.* A Fully Integrated Real-Time Detection, Diagnosis, and Control of Community Diarrheal Disease Clusters and Outbreaks (the INTEGRATE Project): Protocol for an Enhanced Surveillance System. *JMIR Res Protoc* **8**, e13941 (2019).
95. Bracken, J. Roadmap to the Digital Transformation of Animal Health Data. *Front Vet Sci* **4**, 123 (2017).
96. VanderWaal, K., Morrison, R.B., Neuhauser, C., Vilalta, C. & Perez, A.M. Translating Big Data into Smart Data for Veterinary Epidemiology. *Front Vet Sci* **4**, 110 (2017).
97. Forum, W.E. Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data. (2019).
98. Zhang, R.*, et al.* From concept to action: a united, holistic and One Health approach to respond to the climate change crisis. *Infect Dis Poverty* **11**, 17 (2022).
99. Tamburis, O. & Benis, A. One Digital Health for more FAIRness. *Methods Inf Med* **61**, e116-e124 (2022).
100. Jin, L.*, et al.* Integrating Environmental Dimensions of "One Health" to Combat Antimicrobial Resistance: Essential Research Needs. *Environ Sci Technol* **56**, 14871-14874 (2022).
101. Gulfidan, G., Beklen, H. & Arga, K.Y. Artificial Intelligence as Accelerator for Genomic Medicine and Planetary Health. *Omics (Larchmont, N.Y.)* **25**, 745-749 (2021).
102. Lustgarten, J.L., Zehnder, A., Shipman, W., Gancher, E. & Webb, T.L. Veterinary informatics: forging the future between veterinary medicine, human medicine, and One Health initiatives-a joint paper by the Association for Veterinary Informatics (AVI) and the CTSA One Health Alliance (COHA). *JAMIA Open* **3**, 306-317 (2020).
103. K, B.Y.*, et al.* Assessing Climate Change Impact on Ecosystems and Infectious Disease: Important Roles for Genomic Sequencing and a One Health Perspective. *Trop Med Infect Dis* **5**(2020).
104. Gardy, J.L. & Loman, N.J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet* **19**, 9-20 (2018).
105. Wendt, A., Kreienbrock, L. & Campe, A. Joint use of Disparate Data for the Surveillance of Zoonoses: A Feasibility Study for a One Health Approach in Germany. *Zoonoses Public Health* **63**, 503-514 (2016).
106. Shaikh, A.T., Ferland, L., Hood-Cree, R., Shaffer, L. & McNabb, S.J. Disruptive Innovation Can Prevent the Next Pandemic. *Front Public Health* **3**, 215 (2015).
107. Peter Rabinowitz, M.M., Matthew Scotch, PhD MPH, and Lisa Conti, DVM, MPH. Human and Animal Sentinels for Shared Health Risks. *Vet. Ital.* **45**, 23-24 (2009).
108. Duane A. Steward dvm, m., phd Rosina C. Krecek phd, mba Harry M. Chaddock dvm, eml Julie M. Green dvm, ms Lisa A. Conti dvm, mph. The contribution of biomedical informatics to one health. *JAVMA* **248**(2016).
109. George, J.*, et al.* A systematic review on integration mechanisms in human and animal health surveillance systems with a view to addressing global health security threats. *One Health Outlook* **2**, 11 (2020).
110. Mirzaei, A., Aslani, P. & Schneider, C.R. Healthcare data integration using machine learning: A case study evaluation with health information-seeking behavior databases. *Res Social Adm Pharm* **18**, 4144-4149 (2022).
111. FAO, U., WHO, WOAH. ONE HEALTH JOINT PLAN OF ACTION (2022-2026). (2022).
112. Zanet, S.*, et al.* Literature review on worldwide surveillance systems targeting transboundary zoonotic and emerging diseases within the holistic One-Health perspective. *EFSA Supporting Publications* **19**(2022).
113. Benis, A. & Tamburis, O. One Digital Health Is FAIR. *Stud Health Technol Inform* **287**, 57-58 (2021).
114. WHO. Integrated Disease Surveillance and Response Technical Guidelines. (2017).

115. A, G.J.S. Where the rubber meets the sky: Bridging the gap between databases and science. (Microsoft Corporation, 2004).
116. OHEJP, O.a.B. One Health Sequencing for Surveillance HandBook.
117. Dushoff, C*., et al.* First Nations Health: The need for linked genomic surveillance of SARS-CoV-2. *Canada Communicable Disease Report* **48**, 131-131 (2022).