Genomic epidemiology on the frontline: Inferring disease dynamics from pathogen genomes and supporting genomic analysis in applied public health settings.

Allison Black

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Trevor Bedford, Chair

M. Elizabeth Halloran

Frederick Matsen IV

Program Authorized to Offer Degree:
Epidemiology

University of Washington

Abstract

Genomic epidemiology on the frontline: Inferring disease dynamics from pathogen genomes and supporting genomic analysis in applied public health settings.

Allison Black

Chair of the Supervisory Committee:
Affiliate Assistant Professor Trevor Bedford
Department of Epidemiology

Within infectious disease epidemiology, genomic epidemiology is a field that seeks to describe pathogen transmission dynamics using evolutionary analysis of pathogen genome sequences and associated metadata. Genomic data have a wealth of information; we can use them to group related cases of disease, detect cryptic disease transmission, differentiate source and sink populations, and describe how introductions and sustained transmission contribute to an epidemic. In this dissertation I describe two genomic epidemiological studies of Zika virus, one in Colombia and the other in the United States Virgin Islands. I describe how each country's outbreak was shaped by regional seeding events and endemic transmission after introduction. These studies indicate differences in introduction frequency between the two countries, possibly related to the timing of their outbreaks and the number of other countries having concurrent outbreaks. In the last chapter, I describe recommendations for supporting open pathogen genomic analysis in public health agencies. These recommendations were developed from long-form interviews with public health agencies, and are designed to facilitate the development of genomic epidemiology outside of academia. Taken together, this body of work describes the application of genomic epidemiologic techniques and demonstrates possible strategies for operationalizing genomic surveillance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Trevor Bedford, for giving me the freedom to explore my ideas and grow my independence balanced with the support to help those endeavours succeed. Your optimism about projects and faith in me has carried me through times when I lacked both. To my committee — thank you for your insightful questions, perspectives, and advice throughout my doctoral work. Thank you to my co-lead author and friend, Dr. Louise Moncla. Working alongside you has helped me grow as a scientist and made long wet lab hours simply more fun. While I hope to never do another gel extraction, if I had to, I'd want to do it with you. To my labmates — you have helped me mull over data, troubleshoot code, and design more effective visualizations. I am so lucky to have been surrounded by all of you and I'm deeply grateful for your scientific, technical, and moral support. I sincerely thank the Bloom and Geballe labs at Fred Hutch, where I performed the wet lab portions of my research. They provided space, equipment, and knowledge that were critical to my studies of Zika virus. I would like to acknowledge the US Virgin Islands Department of Health, the Colombian National Institute of Health, and the CDC Office of Advanced Molecular Detection for the collaborations that made my dissertation work possible. I am grateful for financial support from the NSF Graduate Research Fellowship Program (DGE-1256082). On a personal note, thank you to my fiancé, my family, and my friends. This journey would have been so much harder without your patience, love, and support.

Chapter 1

INTRODUCTION

1.1 Public health and public health surveillance

What is public health? This question engenders varied responses, likely due to the inherent complexity and breadth of topics studied within public health. In 1988, an Institute of Medicine committee gave this definition [8]: "The mission of public health [is] the fulfillment of society's interest in assuring the conditions in which people can be healthy. The substance of public health is the organized community efforts aimed at the prevention of disease and promotion of health."

The first step in preventing disease is to detect it. Finding cases of disease requires a definition of what we consider a case and a mechanism by which to report it. Those two components allow us to build up a dataset; with that datatset we can describe how the disease is distributed in a population and determine which factors shape that distribution. Notably, we can use these data to detect a problem and, later on, to evaluate our solutions to it. This process of consistent data collection, management, analysis, interpretation, and policy development is public health surveillance [61]".

In infectious disease epidemiology, the data that we collect and analyze are usually counts of cases of a disease. We detect outbreaks by looking for elevated counts that we cannot attribute to changes in a case definition or reporting practices. We describe the growth of an outbreak from the numbers of new cases occurring over time, and we determine which populations are most at risk by analyzing the spatial, temporal, and demographic characteristics of case counts [64]. The goals of genomic epidemiology are the same. We want to know the distribution of states of disease over space, time, and demography, and we want to know the determinants of those patterns. The

difference is simply the data we use to explore those questions. Rather than treating case counts as the record of transmission history, we query the genome of the infecting pathogen.

Increasingly, infectious disease epidemiologists have access to molecular data, and expanded public health sequencing programs have ushered in the era of "precision public health" [1], so named for the ability of genomic data to distinguish related and unrelated cases of disease with high sensitivity. Beyond grouping related cases of disease, we can analyze genomic data to detect cryptic disease transmission[30, 20, 48], to differentiate sources and sinks of disease [4], to describe patterns of introductions [15], to monitor for the emergence of adaptive variants that may transmit more effectively [10, 71], and to evaluate whether an infection is susceptible or resistant to antimicrobials [12].

However, a single sequence on its own is not sufficient. Akin to traditional surveillance, where baseline incidence rates provide necessary context to detect the elevated incidence of an outbreak, the genomic epidemiologist needs broad datasets, sampled over space and time, which describe the circulating diversity of a pathogen at baseline. In this dissertation I refer to the systematic collection of pathogen genomic sequence data to build these broad datasets as 'genomic surveillance'.

To investigate pathogen dynamics, the genomic epidemiologist performs comparative genomic analyses on the broad dataset. Frequently, we reconstruct phylogenetic trees, which cluster genome sequences based on their similarity. At their most basic, these trees provide graphical summaries of genetic divergence, showing shared and unique changes to the genome sequence. However, we can perform more sophisticated evolutionary analysis as well. For example, inferring phylogenetic trees under the coalescent [39] enables us to infer changes in the size of the viral population. Given an estimate of the serial interval of the pathogen, we can estimate prevalence and incidence from these demographic patterns [72, 3, 25]. When we join them with sample collection date and geographic information, we can estimate when an outbreak

started and spatial movements over time. Given their fundamental importance to genomic epidemiologic analysis, I describe these techniques in greater detail in the next section of this introduction. While genomic epidemiologists perform these analyses on various types of pathogens, including parasites and bacteria, I describe the analysis of RNA viruses, which are the pathogens I have studied during my doctoral work.

1.2 Viral phylodynamics

With their fast mutation rates, large population sizes, and the strong selective and stochastic forces working upon them, the genomes of RNA viruses evolve on the same time scale as disease transmission [29]. Understanding how epidemiological processes, host immunity, and evolutionary processes shape those evolutionary trajectories is the objective of 'viral phylodynamics' [72]. In this section, I briefly describe the evolutionary process of RNA viruses and why evolution is detectable on the same time scale as transmission. I then describe two frequently used analytic procedures for understanding epidemiologic processes from phylogenies.

1.2.1 Mutation rates, evolutionary rates, and recording the transmission process

Viral RNA-dependent polymerases have low fidelity and lack proofreading activity. As they replicate the viral genome for packaging into progeny virions, the polymerase makes errors; on average, one mutation occurs somewhere in the genome every replicative cycle [13, 47]. While sometimes a mutation confers a benefit to the virus that selection can act upon, the vast majority of these mutations have no functional impact or are detrimental to the virus [31]. Thus we should not regard the mutational process as evolution towards fitness; rather, most mutations are simply markers that replication occurred.

The occurrence of mutations during replication means that 'an infection' is in fact characterized by a diverse collection of viral variants. While many of these variants will only ever exist at low frequencies [75], some will rise in frequency to become the

main variant in the infection. The sequence of this primary variant is summarized by a 'consensus sequence'. Many genomic epidemiological studies, and all of the studies I have conducted for my doctoral work, analyze differences between consensus genome sequences. Through the processes of mutation, selection, stochastic drift, and epidemiological processes, different variants will rise to majority frequency in different infected individuals[36]. While this leads to an accrual of changes to the consensus sequence as transmission occurs, this rate of change is lower than the viral mutation rate, and sometimes the consensus sequence may not change at all between directly infected cases. In contrast to the viral mutation rate, we generally refer to the rate of change in consensus sequences during an outbreak as the 'evolutionary rate' of the virus[36]. Notably, evolutionary rates are time-dependent; they will vary by pathogen due to differences in the duration of the disease, different strengths of selective pressures exerted on the virus, and the time scale over which cases are sampled[36].

As markers of transmission, the nucleotide changes that accrue in the genome over the course of an outbreak allow grouping of similar sequences within phylogenies, essentially family trees of viruses. As described above, there are various models under which we can construct phylogenies, and different inferences than we can draw from these reconstructions, which I detail below.

1.2.2 Coalescent rates and their relationship with pathogen prevalence

Under the coalescent, we infer changes in the pathogen population size from the rate of coalescence observed in the phylogeny. In an idealized panmictic population without selection or population structure, when population sizes are large, genetic diversity is higher, and it takes longer for two randomly sampled lineages to trace back to their most recent common ancestor and coalesce. When population sizes are small, genetic diversity is low, and the time to the next coalescent event is short. Thus in haploid organisms, such as viruses, the rate of coalescence is inversely related to the

population size, such that the coalescent rate can be written as

$$\lambda = \frac{1}{N_e}$$

Flexible demographic models in coalescent phylogenetics estimate changes in population size over time given the changes in coalescent rates observed in the genealogy (Figure 1.1A) [53].

While the genealogy provides information about the pathogen population size based on changes in the rate of coalescence, the parameters estimated directly from coalescent models are always scaled, as the coalescent measures time in generations[53]. Thus, the topology of the geneaology will depend on both the effective population size (N_e) and on the generation time of the pathogen, τ[3]. This means that only the product of these two values, or $N_e\tau$ can be estimated directly from the sequence data. Thus, to estimate N_e, or pathogen prevalence, we must divide $N_e\tau$ by τ, where τ is denoted in years per generation[3]. The generation time, or τ, is the same as the transmission interval, defined as the amount of time between successive infections, regardless of whether the infection is sub-clinical or not. Notably, this is a slightly different quantity from the serial interval, which represents the duration of time that passes between clinical, or observed, infections[24]. While there will be variation in transmission intervals during an outbreak, τ can be approximated by the average transmission interval, which is the sum of the average duration of time between being infected and becoming infectious, and the average duration of infectiousness[24]. Thus, once one has an estimate of τ, the effective pathogen population size, N_e, can be estimated as

$$\frac{N_e\tau}{\tau}$$

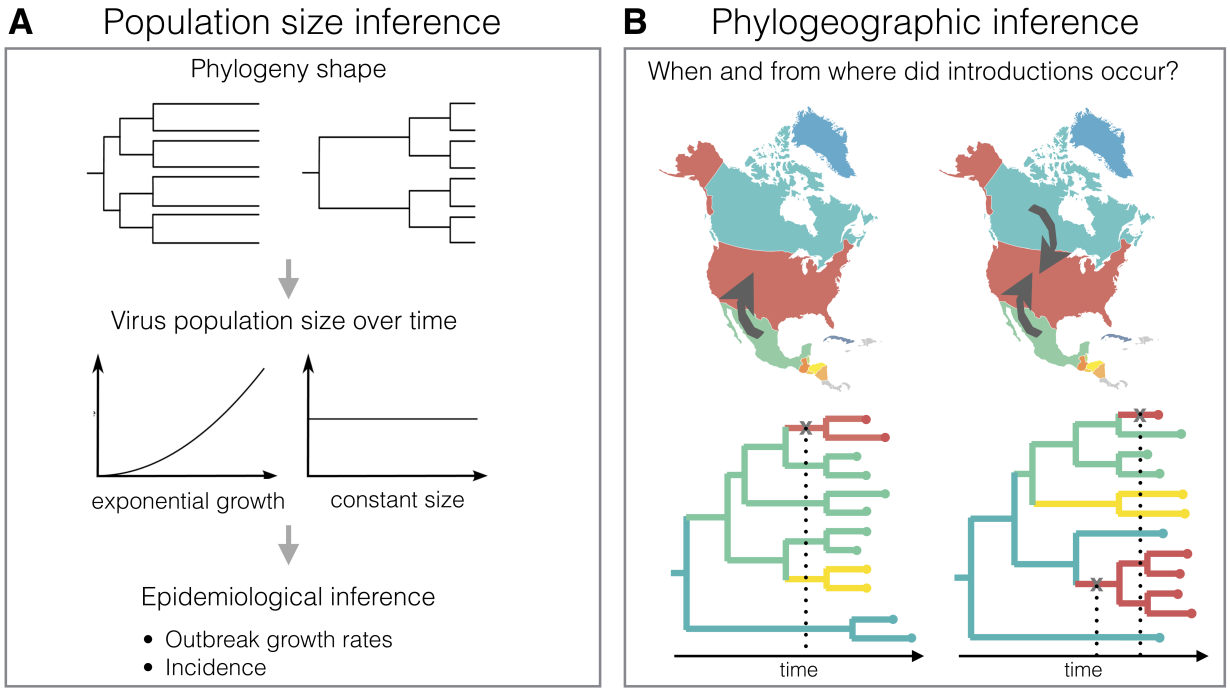

Figure 1.1. Pathogen dynamics that we can infer from phylogenies. A) We can infer changes in the pathogen population size through time from the shape of the phylogeny. The rate of coalescence is given as $\frac{1}{N_e}$, where N_e is the effective population size. When effective population sizes are large, the time to a coalescent event will be longer. With additional information, this relationship allows us to estimate prevalence and outbreak growth rates. B) When geographic sampling information is available for sequenced samples, we can probabilistically reconstruct where unsampled, ancestral infections, represented by internal nodes in the tree, were located. When this reconstruction is performed on temporally-resolved trees, we can infer both the patterns of geographic movement and the timing of movement events.

1.2.3 Reconstructing spatiotemporal patterns with phylogeography

In addition to estimating the frequency of a disease, epidemiologists seek to understand the spatiotemporal distributions of pathogens. Genomic epidemiologists infer these distributions using phylogeographic analyses. Broadly, phylogeographic meth-

ods model movement between demes, where a 'deme' is any bin defining a subpopulation; common demes are geographic divisions (such as countries, counties, or cities), different host species (for example, bats and humans), or even tissue divisions within a single host. One of the most frequently used phylogegraphic methods is Discrete Trait Analysis (DTA)[43]. We commonly use DTA to infer geographic locations of unsampled ancestral pathogens in temporally-resolved phylogenies, which allows us to reconstruct a pathogen's spatial movements over time (Figure 1.1B).

1.3 About this dissertation

My dissertation work focuses on two sides of genomic epidemiology. In Chapters 2 and 3 I present two genomic epidemiological studies of Zika virus, one in Colombia, and the other in the United States Virgin Islands. These studies illustrate how we can generate sequence data in the field and use those data to investigate epidemiological dynamics. Specifically, both studies describe how each country's outbreak was shaped by country-to-country transmission within the Americas and transmission within the country after introduction. It is my hope that, by providing greater information about Zika virus transmission, these studies will help us evaluate the degree to which outbreak response measures inhibited transmission. Additionally, by increasing our understanding of the scale of these outbreaks, I hope that we will be better equipped to model the risk of Zika virus re-emergence, and thus more effectively prepare for future outbreaks.

In Chapter 4 I explore how we can improve support for genomic epidemiology in public health settings. Despite the utility of genomic data for public health surveillance, genomic epidemiology studies have most frequently been conducted by academics. Now that public health institutions regularly generate large amounts of genomic data, the comparatively limited amount of genomic epidemiological analysis conducted by public health agencies belies a critical gap. The capacity to transform, analyze, and interpret genomic data within an epidemiological context has not kept

pace with the ability to develop and scale up laboratory protocols for large-scale sequencing programs. My third dissertation aim presents work I led in collaboration with the Office of Advanced Molecular Detection at the Centers for Disease Control and Prevention. Through extensive long-form interviews with bioinformaticians and epidemiologists working in public health, I sought to learn about the challenges that public health practitioners face in trying to develop their genomic epidemiology programs, and how, in some cases, they overcame those challenges. These interviews formed the basis of our recommendations for supporting open pathogen genomic analysis in public health settings. I am proud that this work is truly a starting point, and that support for this initial effort has been taken up and amplified by the Public Health Alliance for Genomic Epidemiology (PHA4GE).

What happens when everything comes together, and the capacity to generate genomic data quickly and consistently is matched by the capacity to also analyze and interpret those data on actionable time scales? In the concluding chapter of this dissertation, I illustrate what the integration of real-time genomic surveillance with efficient genomic analysis can yield in an outbreak scenario. This picture comes from collaborative work with the Institut Nationale de Recherche Biomédicale (INRB) in the Democratic Republic of the Congo, which has been performing genomic surveillance for Ebola virus disease during the Nord-Kivu outbreak.

Many PhD students finish their dissertations wondering how the field will value and build off of their work. I am incredibly grateful for the opportunity to have actually experienced achievements that have been made possible through the decades of work that has come before mine.

Chapter 2

GENOMIC EPIDEMIOLOGY SUPPORTS MULTIPLE INTRODUCTIONS AND CRYPTIC TRANSMISSION OF ZIKA VIRUS IN COLOMBIA

2.1 Background

In recent years, countries across the Americas have experienced the emergence and endemic circulation of various mosquito-borne viruses, making this a critical area for public health surveillance and epidemiologic research. Zika virus (ZIKV) caused a particularly widespread epidemic, with over 800,000 suspected or confirmed cases reported [55]. Given estimated seroprevalence rates between 36% and 76% [17, 76, 2, 51], the true number of ZIKV infections in the Americas is likely much higher. With neither a vaccine nor ZIKV-specific treatments available, understanding the epidemiology of ZIKV is our primary tool for controlling disease spread [45]. However, because many infections are asymptomatic [17], the analysis of surveillance data alone yields inaccurate estimates of when ZIKV arrived in a country [20, 30]. In such cases, introduction timing and transmission dynamics post-introduction are better inferred from genomic epidemiological studies, which use joint analysis of viral genome sequences and epidemiologic case data. Indeed, such studies have defined our understanding of when ZIKV arrived in Brazil [20], described general patterns of spread from Brazil into other countries in the Americas [20, 69, 48], and been used to investigate the extent of endemic transmission occurring post-introduction [30]. Genomic epidemiological studies of the spread of ZIKV in the Americas have aided our understanding of the epidemic [20, 30, 69, 48, 21, 5, 50, 41, 40, 19, 32, 58, 27], but generally, ZIKV pathogen sequencing has remained a challenge for the public health community [62].

Colombia has a population of approximately 48 million people. In addition, Colombia has *Aedes aegypti* and *Ae.albopictus* mosquitoes, which are commonly found at elevations below 2000m above sea level [54]. Public health surveillance for arboviral diseases, along with other notifiable conditions, is performed by the Instituto Nacional de Salud de Colombia (INS) [54]. While suspected cases from other municipalities were reported earlier [65], the INS first confirmed ZIKV circulation in mid-September 2015, in the Turbaco municipality on the Caribbean coast. ZIKV spread throughout the country, appearing in areas infested with *Ae.aegypti* that experience endemic dengue transmission and previous circulation of chikungunya virus [37]. Over the entire epidemic Colombia reported 109,265 cases of Zika virus disease [9], making it the second most ZIKV-affected country in the Americas after Brazil. The extent of the epidemic led the INS to start active surveillance for congenital Zika syndrome [49] as well as other neurological syndromes associated with ZIKV infection [57]. While the INS determined that epidemic ZIKV transmission ended in July 2016, they continue to perform surveillance for endemic transmission.

Despite numerous reported cases, only 12 whole ZIKV genomes from Colombian clinical samples were publicly available. These sequences included 1 sample from Barranquilla, Atlántico department, 4 samples from Santander department, and 7 sequences for which departmental or municipal information was unspecified. We sequenced an additional 8 samples from ZIKV-positive human clinical and *Ae.aegypti* specimens, sampled from previously unrepresented Colombian departments. We describe here the first detailed phylogeographic analysis of Colombian ZIKV to estimate when, and how frequently, ZIKV was introduced into Colombia.

2.2 Methods

2.2.1 INS sample selection and processing

The INS National Virological Reference Laboratory collected diagnostic specimens from over 32,000 suspected ZIKV cases over the course of the epidemic. Of these, roughly 800 serum specimens were positive for ZIKV by real time RT-PCR (rRT-PCR) and had cycle threshold (Ct) values less than 30. From this set we selected 176 serum specimens that were ZIKV-positive, but negative for dengue and chikungunya viruses, as per results from the Trioplex rRT-PCR assay [67]. Specimens were selected such that each Colombian department was represented over the entire time period in which specimens were submitted. We extracted RNA using the MagNA Pure 96 system (Roche Molecular Diagnostics, Pleasanton, CA, USA) according to manufacturer's instructions. An extraction negative was used for each plate; positive controls were eschewed given the risk of cross-contaminating low titer clinical samples [62]. We attempted reverse transcription and PCR-amplification of ZIKV using the two-step multiplex PCR protocol developed by Quick and colleagues on all 176 extracted samples [62]. Briefly, cDNA was generated using random hexamer priming and the Protoscript II First Strand cDNA Synthesis Kit (New England Biolabs, Ipswich, MA, USA). We amplified cDNA using the ZikaAsian V1 ZIKV-specific primer scheme [62], which amplifies 400bp long overlapping amplicons across the ZIKV genome, over 35 cycles of PCR. Amplicons were purified using 1x AMPure XP beads (Beckman Coulter, Brea, CA, USA) and quantified using with the Qubit dsDNA High Sensitivity assay on the Qubit 3.0 instrument (Life Technologies, Carlsbad, CA, USA). Due to long storage periods and variable storage temperature, the vast majority of samples were too degraded to amplify. Of the 176 processed samples, only 15 amplified sufficiently to perform sequencing.

2.2.2 UR sample selection and processing

Universidad del Rosario (UR) collected and performed diagnostic testing on 23 human clinical samples from different geographic regions, and 38 *Ae.aegypti* samples from the Cordoba department of Colombia. RNA was extracted using the RNeasy kit (Qiagen, Hilden, Germany) and a single TaqMan assay (Applied Biosystems, Foster City, CA, USA) directed to ZIKV was employed [22] to confirm ZIKV presence. Approximately 60% of samples (14 clinical samples and 23 *Ae.aegypti* samples) were found to be ZIKV-positive by rRT-PCR. From these, we attempted amplification on 8 samples with sufficiently high viral copy numbers that they were likely to amplify. Amplification, purification, and quantification of ZIKV amplicons from UR samples were performed as described above. Of the eight samples that we performed PCR on, four samples amplified sufficiently to conduct sequencing; three samples were from human clinical specimens and one was from an *Ae.aegypti* sample.

2.2.3 Sequencing protocol

We sequenced amplicons from 4 UR and 15 INS samples using the Oxford Nanopore MinION (Oxford Nanopore Technologies, Oxford, UK) according to the protocol described in Quick et al [62]. Amplicons were barcoded using the Native Barcoding Kit EXP-NBD103 (Oxford Nanopore Technologies, Oxford, UK) and pooled in equimolar fashion. Sequencing libraries were prepared using the 1D Genomic DNA Sequencing kit SQK-LSK108 (Oxford Nanopore Technologies, Oxford, UK). We used AMPure XP beads (Beckman Coulter, Brea, CA, USA) for all purification steps performed as part of library preparation. Prepared libraries were sequenced on R9.4 flowcells (Oxford Nanopore Technologies, Oxford, UK) at the INS in Bogotá and at the Fred Hutchinson Cancer Research Center in Seattle.

2.2.4 Bioinformatic processing

Raw signal level data from the MinION were basecalled using Albacore version 2.0.2 (Oxford Nanopore Technologies, Oxford, UK) and demultiplexed using Porechop version 0.2.3_seqan2.1.1 (`https://github.com/rrwick/Porechop`github.com/rrwick/Porechop). Primer binding sites were trimmed from reads using custom scripts, and trimmed reads were mapped to Zika reference strain H/PF/2013 (GenBank Accession KJ776791) using BWA v0.7.17 [46]. We used Nanopolish version 0.9.0 (github.com/jts/nanopolish) to determine single nucleotide variants from the event-level data, and used custom scripts to extract consensus genomes given the variant calls and the reference sequence. Coverage depth of at least 20x was required to call a SNP; sites with insufficient coverage were masked with N, denoting that the exact nucleotide at that site is unknown. After bioinformatic assembly, 8 samples produced sufficiently complete genomes to be informative for phylogenetic analysis.

2.2.5 Dataset curation

All publicly available Asian lineage ZIKV genomes and their associated metadata were downloaded from ViPR [60] and NCBI GenBank. The full download contained both published and unpublished sequences; we sought written permission from submitting authors to include sequences that had not previously been published on. Any sequences for which we did not receive approval were removed. Additionally, we excluded sequences from the analysis if any of the following conditions were met: the sequence had ambiguous base calls at half or more sites in the alignment, the sequence was from a cultured clone for which a sequence from the original isolate was available, the sequence was sampled from countries outside the Americas or Oceania, or geographical sampling information was unknown. Finally, we also excluded viral sequences if they appeared to have too many or too few mutations given the average rate of evolution of ZIKV. This deviation usually occurs if the given sampling date

is incorrect, or if the sequence has been affected by contamination, lab adaptation, or sequencing error. After curation, the final dataset consisted of 360 sequences; 352 publicly available ZIKV full genomes from the the Americas (including Colombia) and Oceania, and the 8 Colombian genomes from the present study.

2.2.6 Phylogeographic analysis

Data were cleaned and canonicalized using Nextstrain Fauna github.com/nextstrain/fauna, a databasing tool that enforces a schema for organizing sequence data and sample metadata, thereby creating datasets compatible with the Nextstrain Augur analytic pipeline github.com/nextstrain/augur and the Nextstrain Auspice visualization platform github.com/nextstrain/auspice. A full description of the Nextstrain pipelines can be found in Hadfield et al [33]. Briefly, Nextstrain Augur performs a multiple sequence alignment with MAFFT [38], which is then trimmed to the reference sequence. A maximum likelihood phylogeny is inferred using IQ-TREE [52]. Augur then uses TreeTime [66] to estimate a molecular clock; rates inferred by TreeTime are comparable to BEAST [66], a program that infers temporally-resolved phylogenies in a Bayesian framework [68]. Given the inferred molecular clock, TreeTime then creates a temporally-resolved phylogeny, infers sequence states at internal nodes, and estimates the geographic migration history across the tree. These data are exported as JSON files that can be interactively visualized on the web using Nextstrain Auspice.

2.2.7 Rarefaction analysis

We generated a series of datasets in which ZIKV genomes from either Colombia or Mexico were subsampled. For each value from 1 to n, where n is the total number of available sequences for the country of interest, we generated 30 subsampled datasets in which n sequences were randomly sampled without replacement. Each subsampled set was then combined with all ZIKV genomes from other countries included in the main phylogeographic analysis described above, and the phylogeographic inference

was rerun. For the Colombian rarefaction analysis, we have $n = 20$ high quality whole genomes, so we inferred $20 \cdot 30 = 600$ phylogeographically labelled trees. For the Mexican rarefaction analysis, we have $n = 51$ high quality whole genomes in total, and inferred $51 \cdot 30 = 1530$ phylogeographically labelled trees. For each labelled tree, we used custom scripts to traverse the resulting phylogeny and count the number of ZIKV introductions into the country of interest. We then plotted the inferred introduction count as a function of the number of sequences sampled.

Table 2.1. Metadata and percent of genome that was able to be called unambiguously for 19 samples that amplified sufficiently to attempt sequencing. Not all samples had annotated Ct values; rather, in some cases, samples were only specified to be positive or negative for ZIKV.

Strain Name	Collection Date	Ct Value	Municipality	Department	Coverage
COL/FH01/2016	2016-02-24	Positive	Montería	Córdoba	4.1%
COL/FH02/2016	2016-02-29	Positive	Montería	Córdoba	94.3%
COL/FH03/2016	2016-03-09	Positive	Ibagué	Tolima	94.3%
Aedes_aegypti/COL/FH04/2016	2016-01-29	Positive	Montería	Córdoba	96.1%
COL/FH05/2016	2016-07-27	16	Belén de Umbría	Risaralda	96.1%
COL/FH06/2016	2016-05-03	22	Barranquilla	Atlántico	0.0%
COL/FH07/2016	2016-05-16	23	Pitalito	Huila	55.7%
COL/FH08/2016	2016-08-25	Positive	Cali	Valle del Cauca	0.0%
COL/FH09/2016	2016-11-02	Positive	Cali	Valle del Cauca	94.3%
COL/FH10/2015	2015-12-16	24.72	Girardot	Cundinamarca	0.0%
COL/FH11/2016	2016-02-10	Positive	Taminango	Nariño	0.0%
COL/FH12/2016	2016-03-22	24.46	Villagarzón	Putumayo	0.0%
COL/FH13/2016	2016-04-01	27.56	San Andrés	San Andrés	10.4%
COL/FH14/2016	2016-05-26	21	Villavicencio	Meta	35.5%
COL/FH15/2016	2016-05-12	23.85	Imported Case	Imported Case	21.0%
COL/FH16/2016	2016-06-22	24	Cali	Valle del Cauca	81.8%
COL/FH17/2016	2016-06-27	Positive	Barranquilla	Atlántico	0.0%
COL/FH18/2016	2016-08-08	Positive	Barranquilla	Atlántico	0.0%
COL/FH19/2016	2016-08-30	Positive	Cali	Valle del Cauca	81.8%

2.3 Results

2.3.1 Sequencing and sampling characteristics of reported ZIKV genomes

In total, we attempted to amplify ZIKV nucleic acid from 184 samples collected by the Instituto Nacional de Salud de Colombia (INS) and Universidad del Rosario (UR). Given the low viral titers associated with most ZIKV infections, as well as long storage times, most samples did not amplify well. We attempted sequencing on 19 samples that amplified sufficiently to generate sequencing libraries (Table 1). Sequencing efforts yielded eight ZIKV sequences with at least 50% coverage across the genome with unambiguous base calls (Table 1). Seven of these viruses came from humans; one virus came from an *Ae.aegypti* pool. Three sequences came from samples collected from infected individuals in Cali, department of Valle del Cauca, two sequences came from Montería, department of Córdoba, and one sequence each came from Ibagué, department of Tolima, Belén de Umbría, department of Risaralda, and Pitalito, department of Huila (Figure 2.1). Colombian viruses are sampled across the period of peak ZIKV incidence in Colombia (Figure 2.2).

Figure 2.1. Geographic sampling locations for the eight Colombian genomes with at least 50% unambiguous genome coverage. Colombia is highlighted with departmental boundaries shown. Three genomes come from Cali and and two genomes come from Montería. All other cities have one genome each. The land-sea mask, coastline, lake, river and political boundary data are extracted from datasets provided by Generic Mapping Tools (GMT) licensed under GNU General Public License.

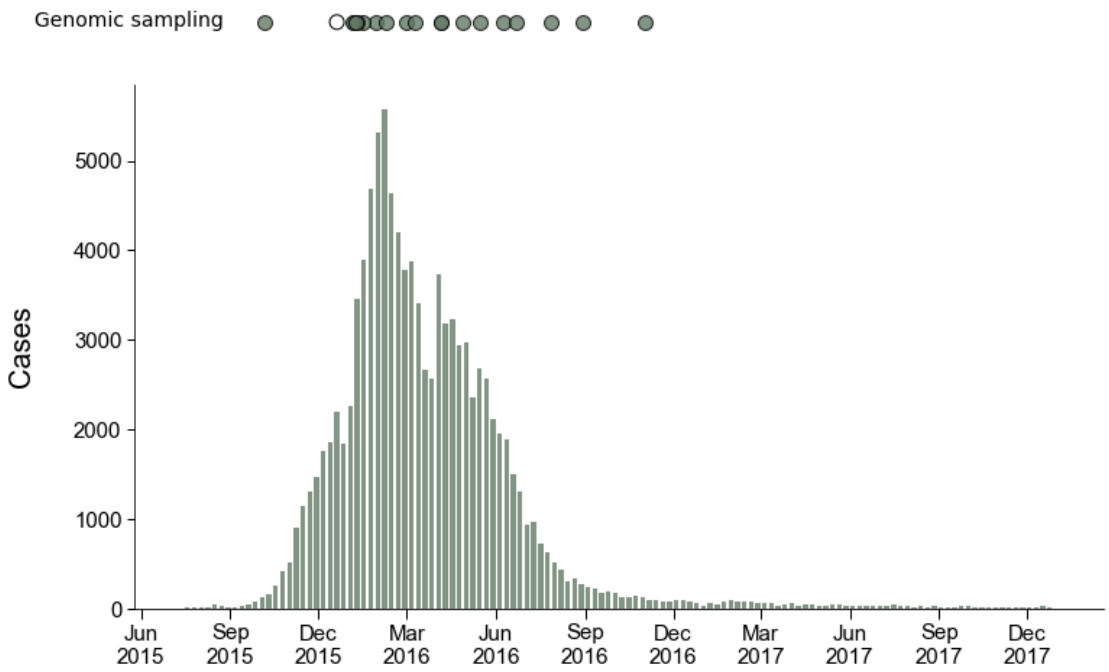

Figure 2.2. Numbers of recorded cases weekly from the beginning of the Colombian epidemic to the end of 2017. Sampling dates for 20 Colombian genomes, including the eight sequenced as part of this study, are indicated by circles above the epidemiologic curve. One sample lacked information about sampling date, and sampling date was inferred. This sample is indicated as an unfilled circle.

2.3.2 General patterns of ZIKV transmission in the Americas

We conducted a maximum likelihood phylogeographic analysis of 360 Asian lineage ZIKV genomes; 18 sequences are sampled from Oceania, and 342 sequences are sampled from the Americas. We estimate the ZIKV evolutionary rate to be 8.65×10^{-4} substitutions per site per year, in agreement with other estimates of the ZIKV molecular clock [20, 21, 48, 59] made with BEAST [68], a Bayesian method for

estimating rates of evolution and inferring temporally-resolved phylogenies. Consistent with other studies, we find that ZIKV moved from Oceania to the Americas, and that the American epidemic descends from a single introduction into Brazil (Figure 2.3A). We estimate that this introduction occurred in late September 2013 (95%CI: August 2013 - March 2014), inline with Faria et al′s [21] initial estimate of introduction to Brazil between May and December 2013, and updated estimate of introduction between October 2013 and April 2014 [20]. We also confirm findings from previous studies [20, 48] that ZIKV circulated in Brazil for approximately one year before moving into other South American countries to the north, including Colombia, Venezuela, Suriname, and French Guiana. Movement of ZIKV into Central America occurs around late-2014 while movement into the Caribbean occurs slightly later, around mid-2015 (Figure 2.3A).

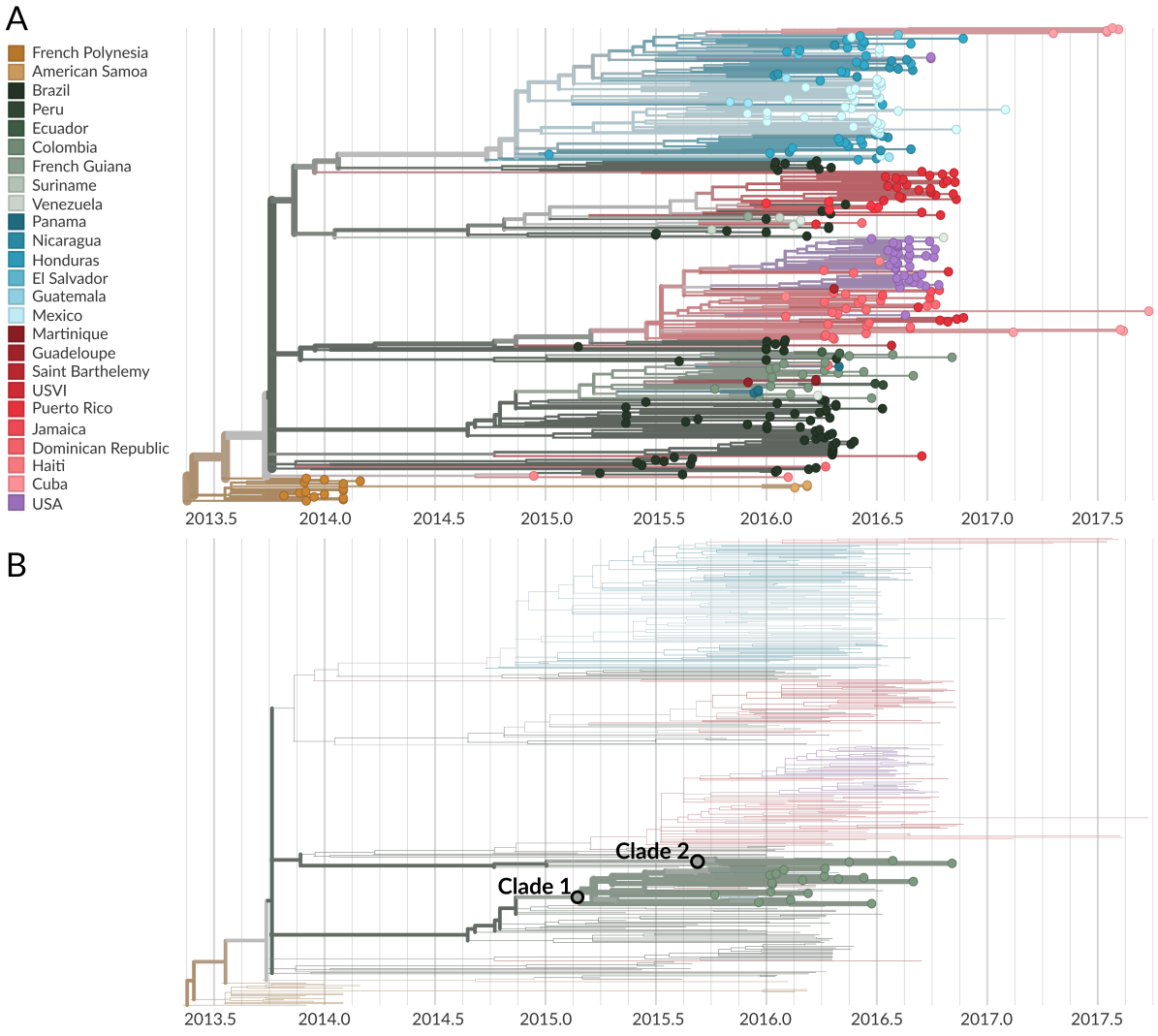

Figure 2.3. Phylogeographic analysis of 360 publicly available ZIKV genomes. A) Temporally-resolved maximum likelihood phylogeny of 360 ZIKV genomes from the Americas and Oceania. Tip colors indicate known country of sampling and branch colors indicate geographic migration history inferred under the phylogeographic model. The full tree can be explored interactively at nextstrain.org/community/blab/zika-colombia. B) The same phylogeny as in panel A, but filtered to highlight only genomes sampled from Colombia. The phylogeny shows two separate introductions of ZIKV into Colombia, resulting in varying degrees of sampled onward transmission.

2.3.3 Multiple introductions of ZIKV to Colombia

We infer patterns of ZIKV introduction and spread in Colombia from the phylogenetic placement of 20 Colombian sequences. Previously, 12 of these sequences were publicly available. We add an additional 8 Colombian sequences sampled over a broad geographic and temporal range. Colombian ZIKV sequences clustered into two distinct clades (Figure 2.3B). Both clades are descended from viruses inferred to be from Brazil (Figure 2.3A). Viruses immediately ancestral to both Colombian clades are estimated by the phylogeographic model to have 100% model support for a Brazilian origin. However, lack of genomic sampling from many ZIKV-affected countries in the Americas may limit our ability to infer direct introduction from Brazil or transmission through unsampled countries prior to arrival in Colombia.

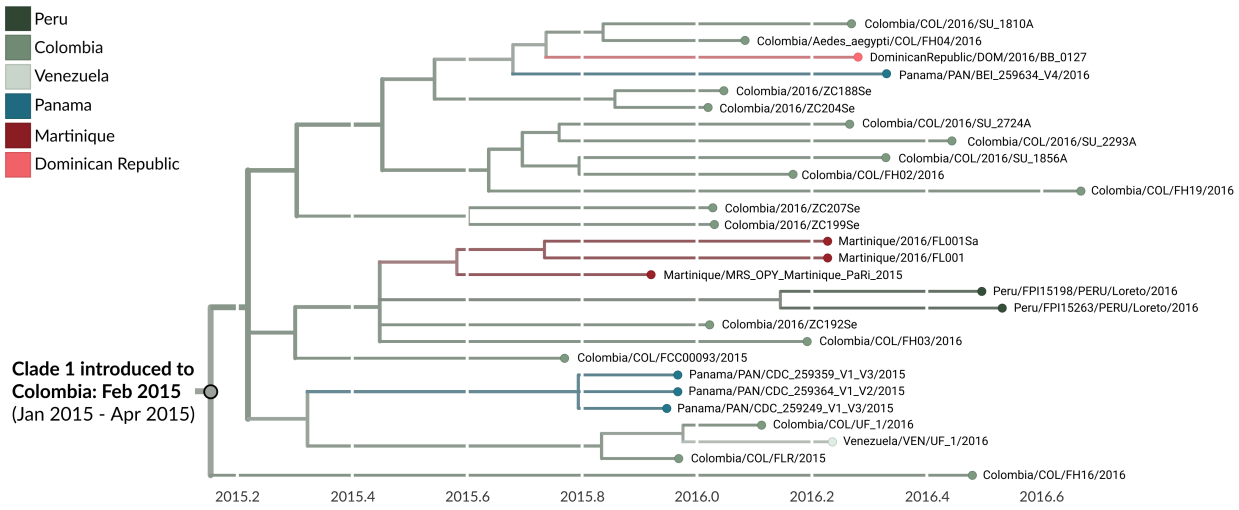

Figure 2.4. Phylogeny of 28 clade 1 viruses. The inferred geographic migration history indicates movement from Brazil into Colombia, with subsequent migration into other countries in South America, Central America, and the Caribbean. The estimate of introduction timing and the 95% confidence interval are given for the most basal node of the clade.

Clade 1 is comprised of 28 viruses, 17 of which are from Colombia, and is char-

acterized by nucleotide mutations T738C, C858T, G864T, C3442T, A3894G, C5991T, C9279T, and A10147G. This clade contains all previously reported Colombian genomes, as well as five of the genomes generated during this study (Figures 2.3 and 2.4). The phylogeographic model places the root of this clade in Colombia with 99% model support. The most parsimonious reading is that this clade of viruses resulted from a single introduction event from Brazil into Colombia.

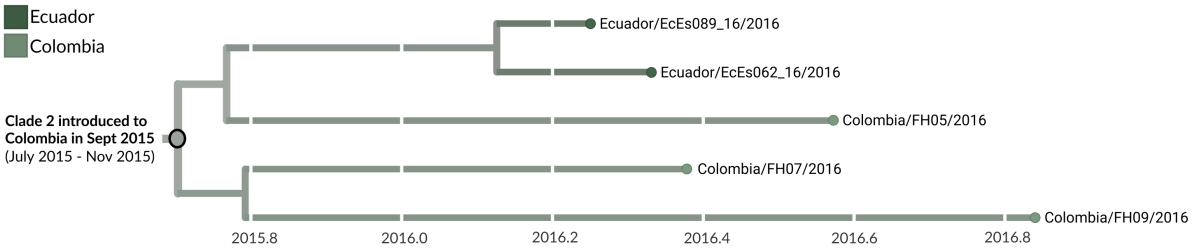

Figure 2.5. Phylogeny of 5 clade 2 viruses. The inferred geographic migration history indicates movement from Brazil into Colombia, with subsequent migration into Ecuador as evidenced by the nesting of Ecuadorian genomes EcEs062_16 and EcEs089_16. The estimate of introduction timing and the 95% confidence interval are given for the most basal node of the clade.

Clade 2 contains 5 viruses, 3 of which are from Colombia (Figures 2.3 and 2.5). All three were sequenced during this study, and thus this clade was not previously recognized in Colombia. This clade is characterized by mutations T1858C, A3780G, G4971T, C5532T, G5751A, A6873G, T8553C, and C10098T. The phylogeographic model places the root of this clade in Colombia with 98% model support and also suggests a Brazil to Colombia transmission route that is likely, but not certainly, direct. Two additional genomes with less that 50% genomic coverage also group within these clades; COL/FH14/2016 within clade 1 and COL/FH15/2016 within clade 2 (data not shown).

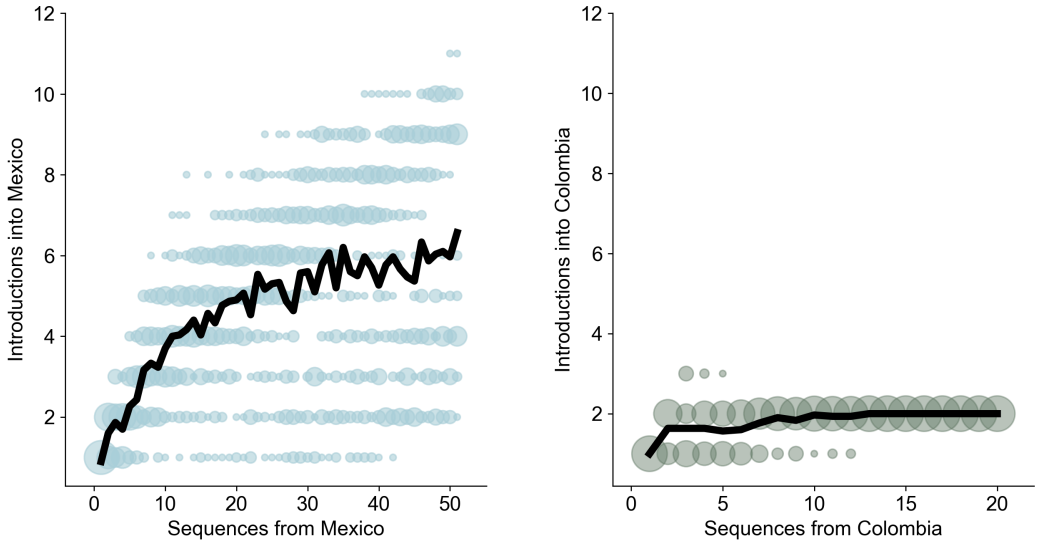

Figure 2.6. Rarefaction curves for Mexican ZIKV and Colombian ZIKV. For both Mexico and Colombia, the number of introductions into the country is plotted as a function of the number of genomes sampled from that country, where genomes are subsampled from available sequences. The colored circles show introduction counts for all of the phylogeographically labelled trees; they are sized according to the frequency with which a specific number of introductions was observed for a given level of subsampling. The black line shows the mean number of introductions observed for a given level of subsampling.

To examine how our estimate of two introductions to Colombia might be affected by the number of sequences available, we conducted rarefaction analyses. Under the assumption that sequenced viruses are sampled at random, these curves show how much genomic sampling is required to fully sample the circulating viral diversity. Figure 6 shows the number of introductions to Colombia or Mexico inferred under the phylogeographic model as a function of the number of sequences sampled from that country. For ZIKV in Mexico, we see that the rarefaction curve begins to flatten once roughly 25 to 30 Mexican viruses have been sampled (Figure 2.6). In contrast,

for Colombia we do not observe further ZIKV introductions once we have sampled around 4 genomes (Figure 2.6). Thus, while there are only 20 ZIKV whole genomes available from Colombia, we think it is unlikely that we would observe more ZIKV introductions given more sequence data.

2.3.4 Transmission within Colombia

We estimate that clade 1 was introduced to Colombia around late February of 2015 (95% CI: January 2015 to April 2015) (Figures 2.3B and 2.4), and that clade 2 was introduced in mid September of 2015 (95% CI: July 2015 to November 2015) (Figures 2.3B and 2.5). Our estimate of clade 1 introduction timing supports between five and eight months of cryptic ZIKV transmission within Colombia prior to initial case detection in September 2015, a finding that is consistent with other genomic epidemiological studies of ZIKV [20, 48, 30].

Genomes were sampled from 7 of 32 departments within Colombia (Figure 2.7A). Four genomes were sampled from Santander, three viruses were sampled from Valle del Cauca, and two viruses came from Córdoba. One virus each was sampled from Risaralda, Tolima, Huila, and Atlántico departments respectively (Figure 2.7B and C). Seven of the publicly-available Colombian ZIKV genomes lacked department-level information (Figure 2.7B). Given that many ZIKV-affected departments within Colombia have only minimal genomic sampling, or lack it all together, we have refrained from using phylogeographic methods to reconstruct the direction of transmission between Colombian departments. However, we do note some signals of geographic influence on transmission within the phylogenies. For example, two closely related clade 1 viruses (COL/2016/SU_1810A and COL/2016/SU_2724A) were both sampled from Santander, and may be linked cases.

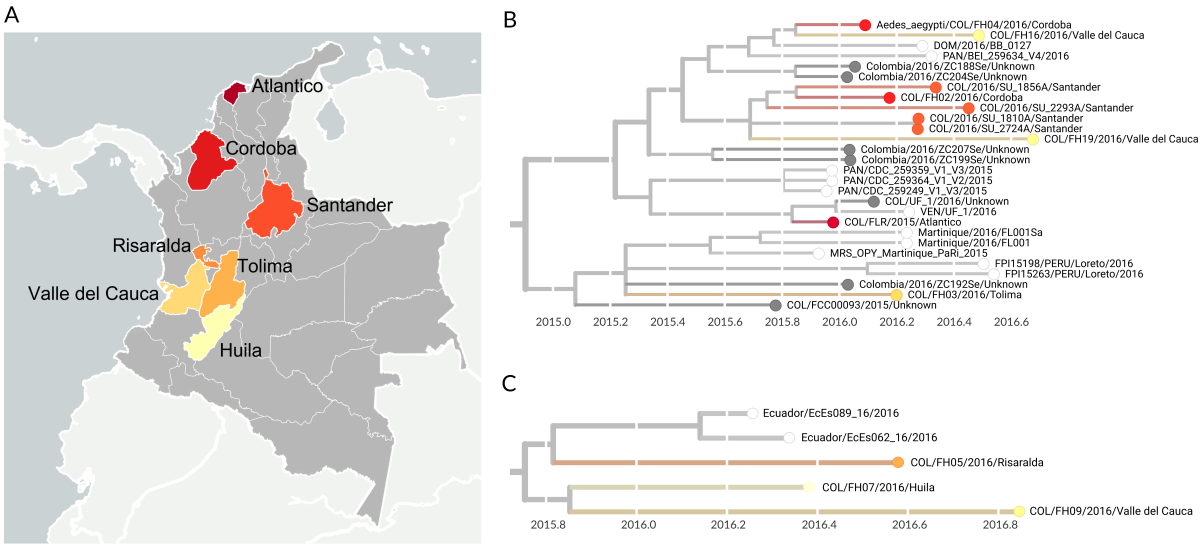

Figure 2.7. Department-level sampling information for 20 viruses sequenced from Colombia. A) Colombian departments from which ZIKV was sampled are highlighted. These departments included Santander, Valle del Cauca, Córdoba. Risaralda, Tolima, Huila, and Atlántico departments. Seven genomes lack public information about which department they were sampled from. The land-sea mask, coastline, lake, river and political boundary data are extracted from datasets provided by Generic Mapping Tools (GMT) licensed under GNU General Public License. B) Phylogeny of clade 1 viruses, colored by department of sampling. Viruses not sampled from Colombia are indicated with white tips, and Colombian viruses lacking department information are colored dark grey. C) Phylogeny of clade 2 viruses, colored by department of sampling. Viruses not sampled from Colombia are indicated with white tips.

2.3.5 Transmission from Colombia to other countries

We find evidence for onward transmission from Colombia into other countries in the Americas. Clade 1 shows movement of viruses into countries that share a border with Colombia, namely Panama, Venezuela, and Peru, as well as into the Dominican Republic and Martinique (Figure 2.4). Clade 2 indicates movement of Colombian ZIKV into neighboring Ecuador (Figure 2.5). Transmission from Colombia into bordering

countries seems reasonable, and these patterns agree with previously documented trends of ZIKV expansion in the Americas [20, 48, 69], but provide more detail due to the greater amounts of sequence data now available. For instance, analysis by Metsky et al [48] also supports movement of ZIKV from Colombia to Martinique and the Dominican Republic. However, without sequence data from Panama, Peru, or Venezuela, they were unable to capture spread from Colombia into these countries.

2.4 Discussion

Despite the scale of the Colombian epidemic, publicly available sequence data were limited, and no detailed genomic epidemiological analysis of ZIKV dynamics had been performed. We sought to improve genomic sampling for Colombia, and to perform a detailed genomic analysis of the Colombian epidemic. Only 12 Colombian genomes were available prior to this study. To these data we added 8 new sequences sampled broadly across Colombia, and performed a phylogeographic analysis of American ZIKV. We describe general transmission patterns across the Americas and present estimates of ZIKV introduction timing and frequency specific to Colombia. We find evidence of at least two introductions of ZIKV to Colombia, yet remarkably the majority of Colombian viruses cluster within a single clade, indicating that a single introduction event caused the majority of ZIKV cases in Colombia. Under the assumption that viruses sequenced from Colombia are random samples of Colombian ZIKV cases, we find that Colombian ZIKV diversity is well represented by 20 Colombian genomes. It is therefore unlikely that further genomic sampling would reveal more introductions of ZIKV into Colombia. ZIKV dispersal out of Colombia also appears widespread, with movement to bordering countries (Panama, Venezuela, Ecuador, and Peru) as well as more distal countries in the Caribbean.

While it may be tempting to read the inferred phylogeographic migration history as a complete record of transmission between countries, we caution against doing this for analyses of ZIKV. In contrast to other large outbreaks, such as the Ebola epidemic

in West Africa, genomic sampling of the American ZIKV epidemic is sparse. Many ZIKV-affected countries have minimal genomic sampling; others have none at all. Thus while the phylogeographic model will correctly infer the geographic location of internal nodes given the dataset at hand, adding sequences from previously unsampled countries may alter migration histories such that apparent direct transmission from country A to country C instead becomes transmission from country A to country B to country C.

Consistent with other studies, our estimates of when introductions to Colombia occurred support cryptic ZIKV transmission prior to initial case confirmation. Perhaps more surprisingly, our estimate of the age of clade 1 indicates that ZIKV likely spread to Colombia even before official confirmation of ZIKV circulation in Brazil [56]. These findings underscore the utility of genomic epidemiology to date introduction events and describe transmission patterns that are difficult to detect using traditional surveillance methods, thereby providing more accurate definitions of the population at risk and a better understanding of how importation and within-country transmission shape epidemics.

2.5 Conclusions

We found evidence for two separate introductions of ZIKV to Colombia, one of which occurred five to eight months prior to the official confirmation of ZIKV in Colombia. Refining our estimates of when ZIKV circulated in Colombia improves our definition of when individuals were at risk for ZIKV infection. Accurately defining this exposure period increases the ability of population-level observational studies to properly measure associations between ZIKV infection and outcomes of interest. In addition, we found that the majority of ZIKV diversity in Colombia descends from one of these two introductions, and rarefaction analyses suggest that we would not identify more introductions with greater genomic sampling. Taken together, these findings suggest that most cases of ZIKV infection were attributable to ZIKV transmission within

Colombia after a single introduction event, and that cases of ZIKV infection acquired in other countries and brought back to Colombia were rare. As the majority of Colombian ZIKV infections were locally-acquired, infection prevention and control measures targeting local spread might have limited the scale of the outbreak within Colombia.

Chapter 3

MULTIPLE INTRODUCTIONS OF ZIKA VIRUS TO THE UNITED STATES VIRGIN ISLANDS, AND THE CHALLENGES OF DATA SPARSITY IN PHYLOGEOGRAPHIC INFERENCE

3.1 Introduction

In 2007, the first reported outbreak of Zika virus occurred on Yap island, Federated states of Micronesia [17]. After an initial introduction, the outbreak appeared to sweep through the island's population; Duffy and colleagues [17] estimated ZIKV seroprevalence to be 73%. In 2013-2014, this pattern was apparently repeated when ZIKV was introduced to French Polynesia. ZIKV transmission within French Polynesia was extensive, and seroprevalence rates there were 49% [2]. We currently believe that ZIKV infection results in lifelong, sterilizing immunity. Given the absence of concurrent ZIKV outbreaks in other countries, and the scale of epidemics in these countries, these ZIKV epidemics indicate an island model of ZIKV transmission characterized by a single introduction followed by widespread within-island transmission.

Currently, we do not know whether this pattern of transmission also characterizes islands affected by ZIKV during the American ZIKV epidemic. Importantly, opportunities for disease control differ depending on the frequency of introduction and the degree of spread that occurs post-introduction. Epidemic transmission dynamics are generally driven by regional seeding events followed by local transmission. This means that we must understand two processes to understand the epidemiology of a pathogen: the process by which new infections are introduced into a population and the characteristics of sustained transmission post-introduction. Despite the importance of

disentangling these processes, inferring introduction frequency from case-surveillance data is challenging, especially once endemic transmission has become established. This is because distinct transmission chains are not easily discernible from each other within incidence data. In such cases, genomic data can be used to resolve large groups of reported infections into distinct clusters [1]. This allows us to detect distinct transmission chains resulting from discrete introduction events.

We sought to describe the patterns of ZIKV introduction, and endemic spread post-introduction, within the United States Virgin Islands (USVI) using genomic epidemiology. Within this chapter, the analysis of the data produces an estimate of the lower bound on the number of times ZIKV was introduced to the USVI, and shows variability in the degree of onward transmission that occurred post-introduction. This chapter also features discussion of particular challenges in accurately estimating introduction frequency when data from other countries are sparse. Specifically, we show the difference in inference that occurs when genomic data from Puerto Rico are included in the phylogeographic reconstruction.

3.2 Methods

3.2.1 Sample collection

The US Virgin Islands is made up of three islands, Saint Croix (population 50,000 people), Saint Thomas (population 51,000 people), and Saint John (population 4,100 people). Suspected ZIKV cases from all three islands were reported to the USVI Department of Health (DoH) between January 2016 and December 2018, although over 95% of cases were reported between January and December 2016. Blood and/or urine samples were collected by primary care physicians from individuals presenting with symptoms of arboviral infection (including ZIKV, Dengue virus, and Chikungunya virus), or from non-symptomatic pregnant women voluntarily receiving ZIKV screening. Samples from each of the three islands, along with questionnaires filled out by

primary care physicians, were collated and entered into a database at the DoH in Saint Croix. Clinical specimens (either serum or urine) were aliquoted at the DoH; one aliquot was shipped to the Centers for Disease Control and Prevention in Atlanta, Georgia, and was tested for ZIKV, Dengue virus, and Chikungunya virus using the Trioplex qPCR assay [67]. The second aliquot was stored at the DoH at -80 degrees Celsius.

In December 2016, we traveled to Saint Croix to sequence ZIKV viral genomes from clinical specimens archived at the DoH. Candidate samples for sequencing were positive for ZIKV by real time polymerase chain reaction (RT-PCR), but RT-PCR negative for Dengue and Chikungunya viruses. From ZIKV-positive samples, we preferentially selected samples with low cycle threshold (Ct) values as these samples were more likely to yield full length genomes. In cases where low Ct samples did not provide sufficient temporal or geographic breadth across the outbreak, we selected additional samples with higher Ct values. We sought to sample each of the three islands in rough proportion to the incidence observed on the island. Epidemiological metadata, including symptom onset dates, sample collection dates, and patient symptoms, were de-identified prior analysis.

3.2.2 Isolation and amplification of viral nucleic acid

Viral RNA was extracted from archival aliquots using the QIAamp Viral RNA Mini Kit (QIAGEN, Hilden, Germany). We extracted RNA from 200 uL of either serum or urine according to manufacturer's specifications. A single extraction negative was included in each extraction batch. We did not use positive controls given the high risk of cross-contaminating clinical samples which generally have low viral titers [62]. Eluted RNA was stored at -80 degrees Celsius. In preparation for sequencing, we generated ZIKV amplicons that tile across the whole genome using a two step approach described in [62]. Briefly, we generated cDNA from 7uL of RNA using the Protoscript II First Strand cDNA Synthesis Kit (New England Biolabs) with random

hexamer priming. We amplified cDNA over 35 to 40 cycles of PCR in two pools using Q5 High-Fidelity DNA Polymerase (New England Biolabs). For each pool we used the primers specified in the ZikaAsian V1 primer scheme, which generates 35 400nt long overlapping amplicons [62]. We used Agencourt AMPure XP beads (Beckman Coulter) to purify amplicons. Finally, we quantified the amplicons using the Qubit dsDNA High Sensitivity assay on the Qubit 3.0 instrument (Life Technologies).

3.2.3 Whole genome sequencing on the Oxford Nanopore MinION

The majority of sequencing for this study was conducted on the Oxford Nanopore MinION, which allowed us to sequence isolates in-country at the USVI DoH laboratory on Saint Croix. After quantification, samples yielding sufficient material were barcoded using the Native Barcoding Kit EXP-NBD103 (Oxford Nanopore Technologies) and pooled in an equimolar fashion. The first two MinION sequencing libraries were prepared using the Genomic DNA Sequencing Kit SQK-LSK208, the 2D sequencing protocol as described in [62]. Genomes were assembled from the 2D reads using the ZiBRA project bioinformatic pipeline [62, 20]. During the study Oxford Nanopore discontinued the 2D sequencing chemistry, and deprecated all software for analyzing 2D sequencing reads. Therefore, subsequent libraries were prepared using the Genomic DNA Sequencing Kit SQK-LSK108. All sequencing libraries were loaded and run on R9.4 flowcells for between 24 and 48 hours. Signal level data were basecalled with Albacore version 2.0.2 (Oxford Nanopore Technologies, Oxford, UK) and demultiplexed using Porechop version 0.2.3_seqan2.1.1 (`https://github.com/rrwick/Porechop`). Custom scripts were used to trim primer sequences from the reads. Afterwards, reads were mapped to the complete genome of Zika reference strain H/PF/2013 (GenBank Accession KJ776791) using BWA v0.7.17 [46]. We inspected all BAM files, including sequenced negative controls, in Geneious 11.1.2 to look for evidence of cross-contamination and to assess coverage patterns across the genome. Nanopolish 0.9.0 `https://github.com/jts/nanopolish` was

used to call single nucleotide variants from event-level data and a custom script was used to generate consensus genomes from VCF files. The entire pipeline was automated as a workflow in Snakemake. All code for the pipeline can be found at `https://github.com/blab/zika-seq/pipeline`.

3.2.4 Validation sequencing across sequencing chemistry and platforms

We validated our MinION sequencing approach by sequencing strains USVI/1/2016 and USVI/4/2016 on MinION 2D chemistry, MinION 1D chemistry, and on the Illumina MiSeq. Illumina MiSeq library preparation and bioinformatic analysis were performed as described in Grubaugh et al [30]. For both MiSeq and MinION data we required a minimum read depth of 20 reads at a site to make a base call. Areas with insufficient coverage were masked with N characters. While consensus sequences had slightly different genomic coverage between chemistries, the unambiguous basecalls were identical across all platforms and chemistries.

3.2.5 Collation of an Americas-wide ZIKV genomic dataset

We curated a dataset of ZIKV whole genome sequences from across the Americas according to the same procedure as described in Chapter 2. However that due to the duration of this project, new data became available over time, and thus updated datasets were curated repeatedly over the past three years. For each curation, we downloaded all ZIKV whole genomes publicly-available from ViPR [60], NCBI GenBank, and from GitHub repositories shared with us. Since sequences on GenBank and ViPR may or may not have already have been included in published analyses, We verified publication status of the genomes manually. We sought written permission to include any genomes that were not published on previously, and any sequences for which we did not receive approval were dropped. Additionally, we excluded any ZIKV sequences that had ambiguous base calls at half or more sites in the alignment, any sequences that were derived from a cultured isolate if a sequence from the origi-

nal clinical diagnostic specimen was available, and any sequences that were sampled from countries outside the Americas or Oceania. Finally, we excluded sequences that appeared to have too many or too few mutations given the date when they were sampled. The first curated dataset analyzed here includes 31 genome sequences from the US Virgin Islands, 225 genomes from other countries in the Americas, and 1 genome from French Polynesia. The second analyzed dataset, collated at the end of 2019, includes 30 genome sequences from the US Virgin Islands, 381 ZIKV genomes from other countries in the Americas, and 18 sequences from Oceania. We removed one of the USVI sequences from the second dataset because, when re-sequenced on the 1D chemistry, the consensus genome was less than 50% complete. This indicates that the initial sequence from this sample was likely an artifact of contamination during the original sequencing run.

3.2.6 Data analysis in Nextstrain

Both the 257 sequence and 429 sequence datasets were canonicalized using Nextstrain Fauna (`https://github.com/nextstrain/fauna`)[33] and analyzed using Nextstrain Augur (`https://github.com/nextstrain/augur`. Within Nextstrain Augur, the data were aligned with MAFFT [38], trimmed to the reference sequence, and then a maximum likelihood phylogeny was inferred from the alignment using IQ-TREE [52]. We used TreeTime [66] within Augur to estimate a molecular clock from the data and to create a temporally-resolved phylogeny. The tree structure and information inferred for ancestral nodes were exported as JSON files for interactive visualization on the web using Nextstrain Auspice (`https://github.com/nextstrain/auspice`).

3.2.7 Bayesian phylogenetic inference in BEAST

Using the aligned and trimmed dataset, I also conducted Bayesian phylogenetic analysis in BEAST (version 1.8.4)[68]. At this time, Bayesian analyses have only been conducted for the initial dataset of 257 ZIKV genomes. We used the single French

Polynesian sequence to root the phylogeny and enforced a monophyly for all sequences sampled from the Americas. This ensured proper rooting and faciliated estimation the root height of the American clade of ZIKV. We inferred independent substitution models and tree likelihoods across 5 genomic partitions: the 5' untranslated region (UTR), the 3' UTR, and 3 codon partitions across the coding sequence. For the site model, site specific rates were scaled by relative rates across the 5 partitions. Molecular evolution was modeled in each partition with an HKY substitution model [35] with Gamma-4 distributed rate variation. Evolutionary rates were modeled with either a strict clock with an uninformative continuous-time Markov Chain (CTMC) reference prior [23], which constrains evolutionary rates to be constant across all branches in the tree, or with an uncorrelated lognormal relaxed clock, where evolutionary rates can vary by branch and are drawn from a lognormal distribution [44]. Finally, we tested two non-parametric demographic models of viral population size. The first model was a Bayesian Skyline model [14] with 50 partitions. The second was a Skygrid model [26] with 100 grid partitions. We tested each of the 4 combinations of clock and demographic models, and each model was run with 4 separate chains to ensure convergence on the same stationary distribution.

3.2.8 Inference of geographic transmission history

To ensure that geographic signal in the data was not overpowered by the phylogenetic signal when constructing the tree topology, we performed phylogeographic inference separately for a posterior distribution of trees inferred using the process described above. After discarding the first 20% of trees from the posterior distribution as burn-in, we modeled the migration history of ZIKV into and across the Americas, treating country as a discrete evolutionary trait [43] and assuming a non-reversible transition matrix [18]. The prior distribution for transition rates between trait states was set as an exponential distribution with mean of 1. The geographically-labelled posterior trees, transition rates, and ancestral states were sampled via Markov chain Monte

Carlo over 20 million steps with trees and transition rates sampled every 5,000 steps after allowing the first 2 million steps for burn-in.

3.2.9 Protocol, data, and code availability

A full list of sequencing reagents and kits, as well as all experimental protocols and bioinformatics pipelines for sequence data analysis are publicly available at `https://github.com/blab/zika-seq`. All consensus ZIKV genomes from the USVI are available in fasta format from `https://github.com/blab/zika-usvi` and were released openly in draft form as they were generated. All scripts for data cleaning and figure generation are also available on `https://github.com/blab/zika-usvi`.

3.3 Results and discussion

3.3.1 Patterns of ZIKV infection from case surveillance data

Between January 3, 2016 and May 23, 2017, 4,385 individuals were tested for ZIKV infection. Of these individuals, 3,072 persons were negative for ZIKV infection, 20 persons were probable cases, 22 persons were suspected cases, and 1,271 persons were confirmed to have ZIKV infection by RT-PCR. Among the confirmed cases, 859 cases were from Saint Thomas, 310 were from Saint Croix, and 102 were from Saint John. Of the confirmed cases 913 were female and 355 were male. An even greater imbalance was seen among individuals testing negative for ZIKV; 2614 were female and 437 were male. This imbalance is likely attributable to the targeted screening of pregnant women for ZIKV regardless of symptom status. The vast majority of confirmed cases reported acquiring their infection within the USVI. Indeed, only three individuals were thought to have acquired their infection outside the USVI. Two of these three individuals reported acquiring their infection in Puerto Rico; the third did not report a country of acquisition.

3.3.2 Genomic sampling of the outbreak

We selected 93 samples upon which to attempt sequencing. The samples were selected to cover the temporal span of the outbreak, and we attempted to keep the number of samples selected from an island proportional to the incidence on that island. There were a variety of constraints on sample selection. Firstly, during the early portion of the outbreak many individuals who had laboratory-confirmed ZIKV infections had antibodies to ZIKV, but no longer had viral RNA in their specimens. Indeed only two PCR-positive samples were available from the first five months of the outbreak, and genomic sampling was thus shifted to the later portion of the outbreak. Additionally, many of the samples which were PCR-positive had very high cycle threshold (Ct) values, indicating a low concentration of virus in the sample. As samples with high Ct samples tend to degrade easily and rarely yield full length genomes, we enriched the sample selection for specimens with Ct values lower than 32.

Of the 93 samples that we extracted and attempted PCR amplification on, 46 samples amplified sufficiently to attempt sequencing. We sequenced all 46 of these samples; 24 samples yielded sequences with 80% or greater genomic coverage, 6 sequences had genomic coverage between 50% and 80%, and 16 samples yielded sequences that were less than 50% complete. For all phylogenetic analyses, we used only sequences that had unambiguous basecalls across more than 50% of the genome. Despite poor sequencing efficiency, we were still able to generate sequence data from each of the three islands covering the temporal span of the peak of transmission (Figure 3.1).

3.3.3 Introductions of ZIKV to the USVI

Using the initially curated dataset of 257 ZIKV genomes, we inferred temporally-resolved phylogenetic trees labelled with the most probable country of origin for unsampled ancestral nodes in the tree. We defined an introduction to the USVI as occurring any time a parent node was reconstructed to have circulated in a country

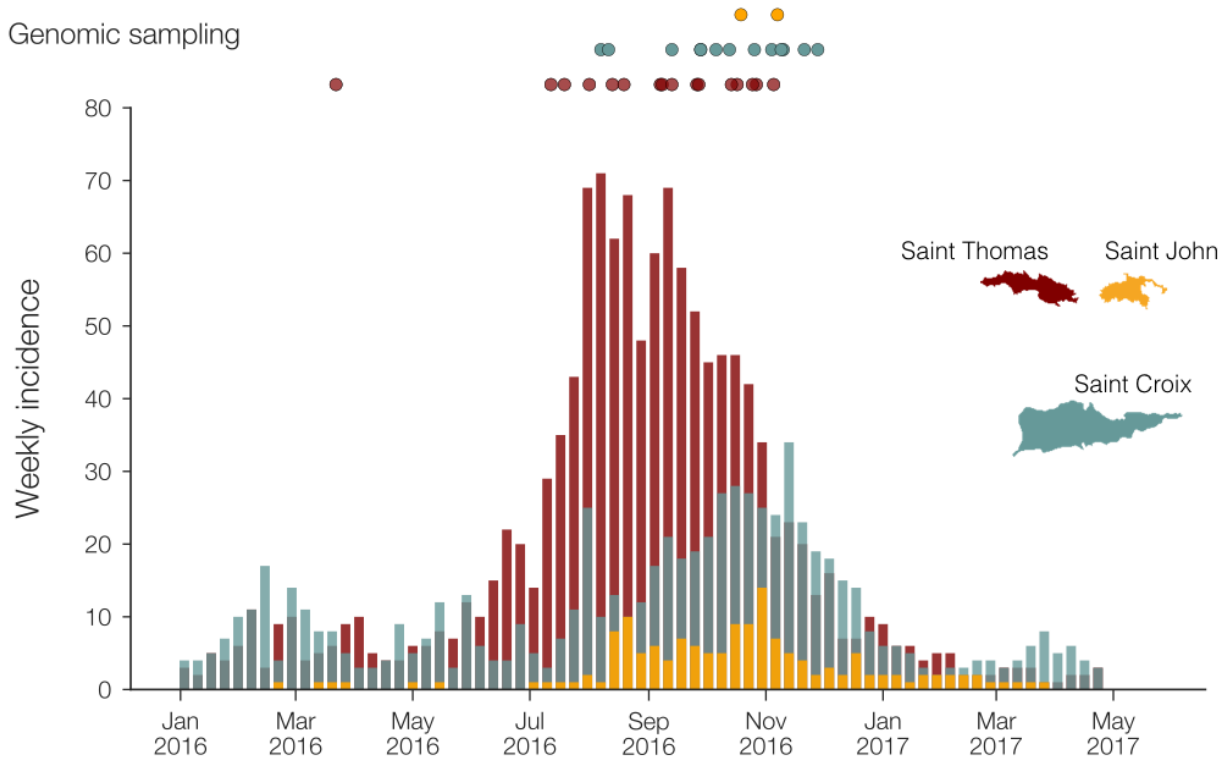

Figure 3.1. Incidence and genomic sampling intensity, by island. Incidence of ZIKV cases, by island. Saint Thomas is shown in maroon, Saint John is shown in yellow, and Saint Croix is shown in teal. Sampling dates for 34 genomes collected from the USVI are shown above, again colored by island.

other than the USVI, and the child node was inferred to circulate within the USVI, with all descendent nodes also inferred to circulate with the USVI.

Under this model, we inferred that ZIKV was introduced to the USVI 7 times (Figure 3.2). However, we hypothesize that this is an underestimate of the true number of introductions. Notably, at the time that this initial dataset was curated,

many ZIKV-affected countries had little or no genomic data available. This was true for Puerto Rico, which was represented by only 6 ZIKV genome sequences despite having over 40,000 confirmed cases [55].

Like the US Virgin Islands, Puerto Rico is a territory of the United States, and travel between the two islands is frequent. While we suspected that the Puerto Rican and USVI outbreaks would likely be linked due to mobility between the islands and concurrent outbreaks, we sought greater detail regarding the source-sink dynamics of ZIKV between the islands. Namely, were introductions into the USVI seeded from Puerto Rico, or was transmission in Puerto Rico seeded by introductions from the USVI?

The paucity of genomic data from Puerto Rico initially made this question challenging to investigate. Discrete Trait Analysis (DTA) [43], the phylogeographic model we initially used, models trait evolution the same way that sequence evolution is modeled. Given a distribution of tree topologies and branch lengths, as well as geographic locations for each tip in the tree, DTA computes the probability that an internal node was sampled from a certain deme, moving from root to tips. Additionally, DTA assumes that evolution across sites is independent, and that evolution along lineages is independent. For this independence assumption to hold true, the observed sequences at the tips must be a random sample of the whole pathogen population. Under this assumption, DTA treats the number of samples that are observed to come from a particular deme as representing the true number of pathogens in that deme at that time.

This means that DTA can infer erroneous migration histories when sampling is heavily biased to a particular deme. For example, in Zhang et al's [77] analysis of MERS Coronavirus (MERS-CoV) evolution, they infer that MERS-CoV migrated from humans to camels 5 times, but that MERS-CoV moved from camels into humans only once. In contrast, appropriate analysis of MERS-CoV genomic data under a structured coalescent model finds the exact opposite relationship

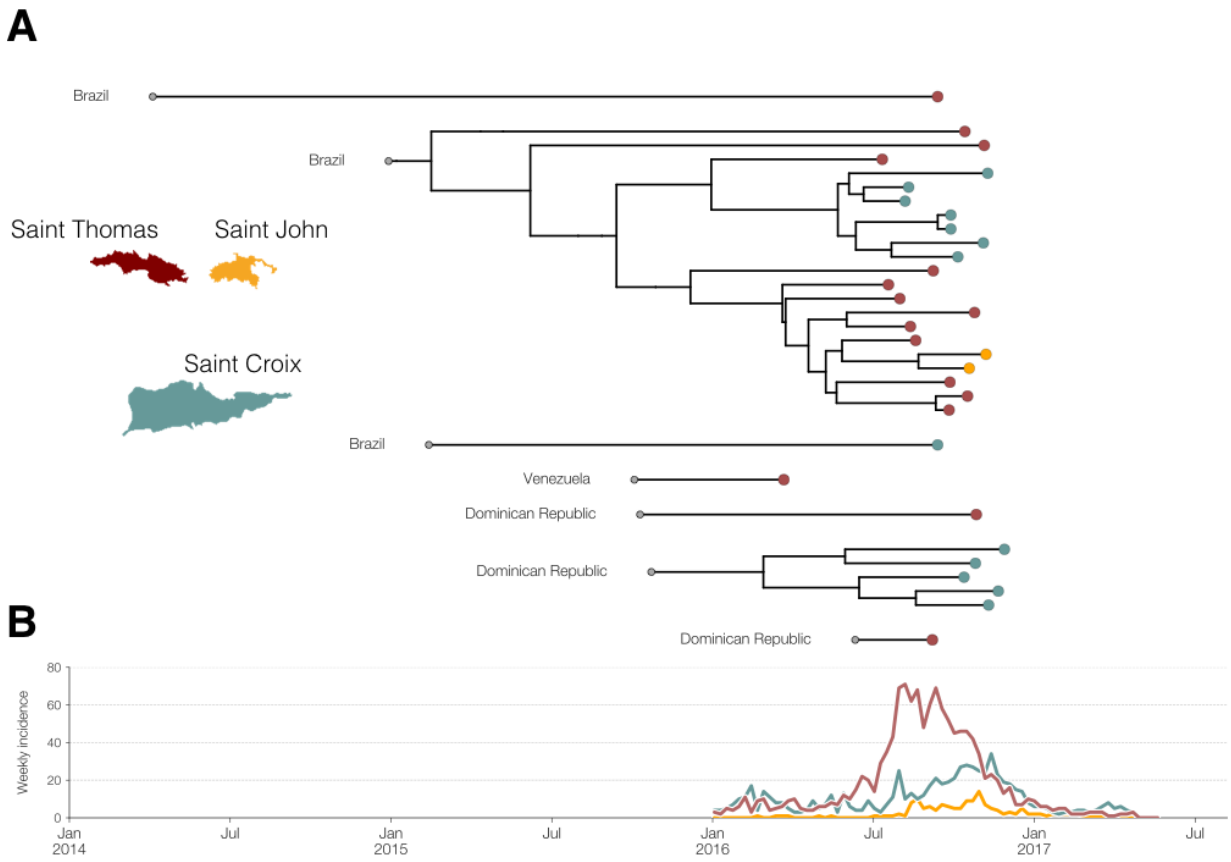

Figure 3.2. Naive phylogenetic reconstruction of ZIKV introductions to the USVI. (A) Distinct introductions of ZIKV to the USVI, where tips are colored to represent the island that they were sampled from (maroon: Saint Thomas, yellow: Saint John, teal: Saint Croix). The last inferred geographic source of the introduction is showing as a basal node for each introduction and is annotated with the most probable country of origin. The temporal scale of the tree matches that of the incidence curves, shown by island, in panel (B).

Subsampling genomic data such that demes are sampled in proportion to their true population size is one strategy for assessing the impact of sampling on the phylogeographic reconstruction and for inferring a more accurate reconstruction. However, given the minimal amount of genomic data available from Puerto Rico (6 genomes),

it was impossible to subsample the USVI genomic data sufficiently so as to have equitable sampling proportions between the two countries. Therefore, to assess the impact of DTA's assumptions on the phylogeographic reconstruction of ZIKV introductions to the USVI, we attempted to model phylogeographic patterns under two other phylogeographic models. These were the structured coalescent (as performed in [16] and a phylogeographic generalized linear model [42], which parameterizes the migration rate matrix between demes as the outcome variable of a regression model.

Both of these models failed to accurately model ZIKV movement from other countries in the Americas into the USVI. In the first case, the structured coalescent analysis consistently inferred that the root of the American ZIKV tree circulated in the USVI. This would indicate that the very first introduction of ZIKV to the Americas occurred into the USVI, and not Brazil, which stands in direct contrast to many other genomic epidemiological studies and case surveillance data. Thus, this particular pattern is almost certainly artifactual, and likely related to the structured coalescent demographic model, which assumes a constant population size through time. In reality, the American ZIKV epidemic grew exponentially, meaning that the viral population size at the root of the tree is initially small and then grows through time. We felt that this mismatch between the actual demographic process and the modeled demographic process likely led to the erroneous reconstruction. In this two deme model, which specifies circulation either within or outside of the USVI, the diversity of the non-USVI sequences indicates a larger population size. Thus, We hypothesized that the model placed the root of the tree in the USVI because the viral population size at the root is small, and the USVI deme with its more limited genetic diversity is inferred to be the smaller deme. Notably, the structured coalescent process can be modeled under other demographic models; the issue is simply related to implementation. Currently the only formal structured coalescent model implemented in BEAST assumes a constant population size.

In the second case, the phylogeographic generalized linear model (phyloGLM)

inferred near-zero migration rates in-to and out-of countries with minimal genomic sampling even when they experienced large epidemics as documented by case surveillance data. This surprised us; we had in fact selected this model in the hope of overcoming sparse genomic sampling. The phyloGLM parameterizes the migration rate matrix between demes as the outcome of a regression model that incorporates multiple predictive variables and their effect sizes. At the outset, we thought that parameterizing the migration rate matrix as the outcome of predictors such as ZIKV environmental suitability and air travel patterns between countries would provide a more accurate reconstruction of ZIKV movement, as it would borrow across data sources in the absence of deep genomic sampling. We thought that, in cases where countries had minimal genomic sampling, the other predictors of ZIKV transmission would give an accurate estimation of migration dynamics in-to and out-of a country. However, this was not how the phyloGLM model behaved. Rather than simply assuming that the lack of genomic data meant missing data, the phyloGLM inferred a zero probability of migration into any country that lacked sequence data, and the regression model was then fit to this outcome. As many ZIKV-affected countries had no genomic data at all, this issue yielded a migration rate matrix in which migration rates appeared to be governed mostly by genomic sampling intensity.

The failed reconstructions left us with two additional options to try. The first was to use a heuristic to overturn the reconstruction of deep internal nodes as circulating in the USVI. The second was to simply wait for data to be generated from Puerto Rico.

3.3.4 Inferred introductions to the USVI when timing is informed by surveillance

In the absence of additional data from other countries, we applied a heuristic; if an internal node was inferred to have circulated in the USVI prior to the first confirmed case of ZIKV in the USVI, we considered that reconstruction to be an artifact of DTA, and we dissolved that internal node. The resulting introductions are shown in

Figure 3.3, where the dashed line demarcates branch length from age of the dissolved internal node to January 2016, when ZIKV was confirmed to circulate in the USVI.

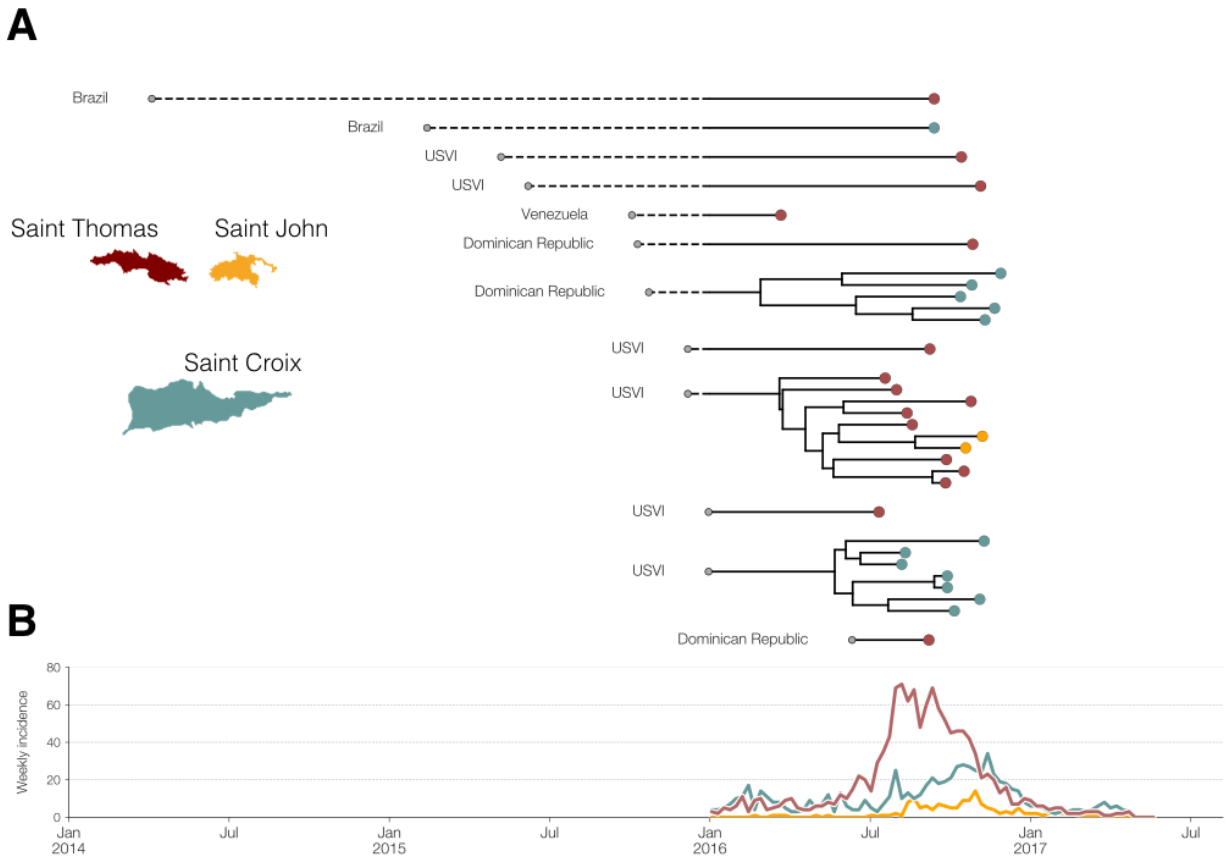

Figure 3.3. Phylogenetic reconstruction of ZIKV introductions to the USVI, accounting for bias in the reconstruction induced by sparse genomic sampling of other ZIKV-affected countries. (A) Distinct introductions of ZIKV to the USVI, but thresholded such that internal nodes inferred to be in the USVI prior to the first confirmed case in January 2016 are broken apart. Tips are colored to represent the island that they were sampled from (maroon: Saint Thomas, yellow: Saint John, teal: Saint Croix). Again, the incidence by island is shown in panel (B), and the temporal scale of both figures is aligned.

Two notable differences occurred when we instituted this threshold. Firstly, the number of discrete introductions of ZIKV to the USVI rose from 7 to 12. Secondly, the largest clade from Figure 3.2 was split such that long branches with tips sampled from Saint Thomas were considered separate introductions, as were the larger clades circulating in Saint Croix and Saint Thomas. This creates a different epidemiological understanding of ZIKV transmission within the USVI. Given the reconstruction in Figure 3.2, we might draw the conclusion that the most successful introduction into the USVI was introduced into Saint Thomas, after which it moved into Saint Croix, back into Saint Thomas, and from Saint Thomas to Saint John. In contrast, under the reconstruction in Figure 3.3, we would infer that ZIKV transmission within the USVI is characterized by introduction and circulation within one island, with minimal transmission between the different islands within the USVI.

3.3.5 Co-circulating transmission chains and variability in transmission success

One of the benefits of genomic epidemiological studies is that, by determining which infections are related to each other at high resolution, we can distinguish between multiple transmission chains that may co-circulate during an epidemic. In both analyses of introduction patterns to the USVI (Figure 3.2 and Figure 3.3), we see that not all introductions resulted in endemic transmission that we sampled. Rather, some introductions resulted in only a single sampled infection, while others led to endemic transmission and the accrual of genetic diversity. Most of these distinct transmission chains co-circulated during the period of peak incidence in the USVI (Figure 3.4), and therefore would likely not have been resolvable as separate introductions within case surveillance data.

Why does differentiating between these transmission chains matter? If all transmission chains contributed equally to the overall numbers of reported cases, or the overall genetic diversity observed within the USVI, then differentiating between transmission chains would likely not matter much. However, certain introduction events

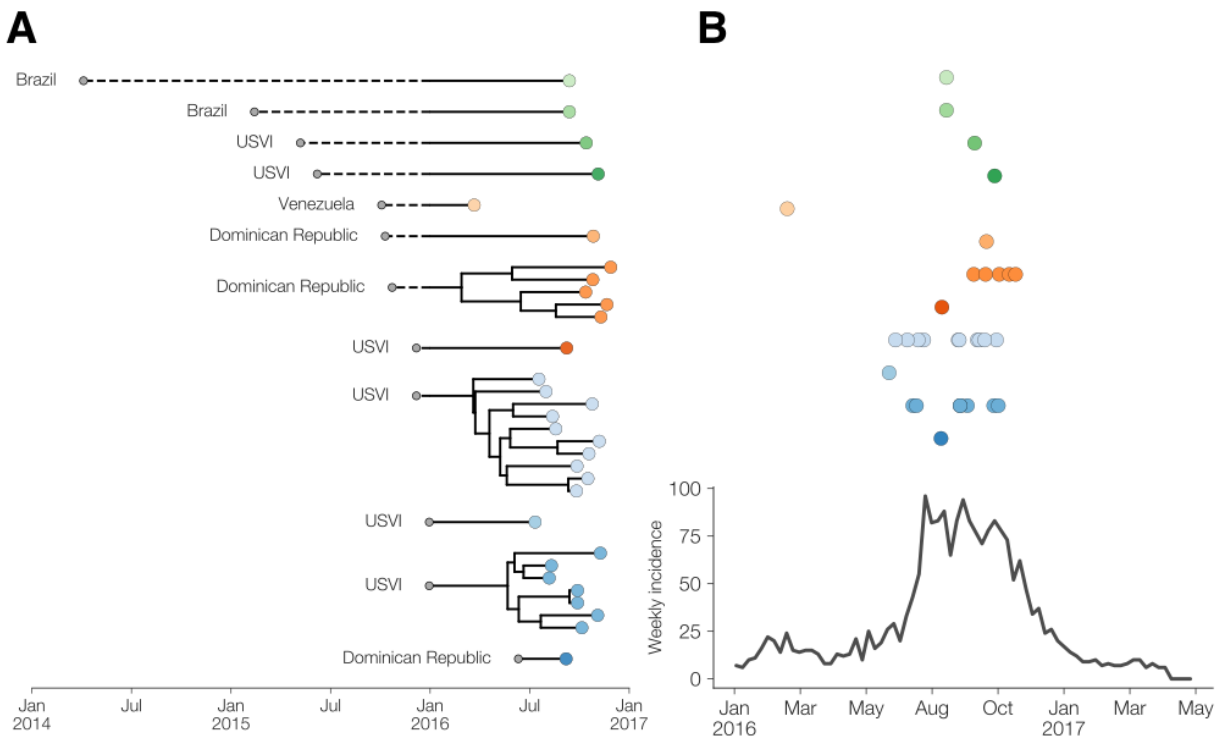

Figure 3.4. Co-circulation of multiple transmission chains resulting from separate introductions. (A) Distinct introductions of ZIKV to the USVI, thresholded as in Figure 3.3. Tips are colored to differentiate separate introductions, and thus discrete transmission chains. (B) Incidence of ZIKV across all three islands in the USVI. Circles above the incidence curve are positioned according to their sampling date, and are colored as in panel A. Y-axis positioning is meaningless and serves only to more easily distinguish the different transmission clusters.

yielded longer transmission chains than others. This raises the question of why some chains were more successful than others. If we could answer that, we might be better equipped to prepare for, or respond to, additional outbreaks of ZIKV.

It's possible that transmission chain success post-introduction is entirely stochas-

tic. However, I would hypothesize that given the seasonality of mosquito abundance in the USVI, the timing of the introduction may influence the probability of transmission post-introduction. If we had even higher geographic resolution we could perhaps also analyze where on island successful transmission chains circulated. Given that ZIKV is vector-borne, it is likely that socioeconomic status would affect transmission, with less access to resources being associated with greater vector density around the home. In this way, differentiation of co-circulating transmission chains could improve our ability to understand the factors facilitating transmission, which might also enable us to tailor control measures.

3.3.6 Inferred patterns of introduction with the addition of genomic data from Puerto Rico

In 2019, additional ZIKV genome sequences from Puerto Rico became available. These additional data provided a unique opportunity to return to the previous analysis and determine whether the addition of more sequence data would split apart clades circulating in the USVI. With the addition of the Puerto Rican data, and using DTA, we found that ZIKV was introduced to the USVI 8 separate times (Figure 3.5B). The inference of 8 introduction events is greater than the number inferred by DTA prior to the addition of Puerto Rican data, but less than the number of introductions inferred when assuming that introduction could not have occurred prior to detection by surveillance. This seems reasonable; there is significant evidence that ZIKV circulated cryptically [20, 30], and thus some introduction events likely did occur prior to the detection of the first confirmed case in the USVI.

The addition of the Puerto Rican genomes splits apart the largest clade that DTA inferred, which grouped genomes from Saint Croix and Saint Thomas together (Figure 3.2). As shown in Figure 3.6, there is evidence for three separate introduction events from Puerto Rico. One introduction event lead to ongoing transmission in Saint Croix, while there were two introduction events into Saint Thomas. One of

Figure 3.5. Phylogenetic reconstruction of ZIKV in the Americas and in the USVI with a larger dataset. (A) Temporally-resolved maximum likelihood reconstruction of 411 American ZIKV genomes and 18 Oceanian ZIKV genomes. (B) The same tree as above, but with all non-USVI branches and tips removed to clarify distinct introductions of ZIKV to the USVI. Vertical distances between clades have been modified for clarity.

these introduction events resulted in ZIKV circulation in Saint Thomas and subsequent movement to Saint John, which is only reachable by ferry from Saint Thomas.

The second introduction into Saint Thomas yielded no sampled transmission. This indicates that there was likely not significant ZIKV mixing between islands, and that most circulation on the individual islands resulted from introductions from outside of the USVI.

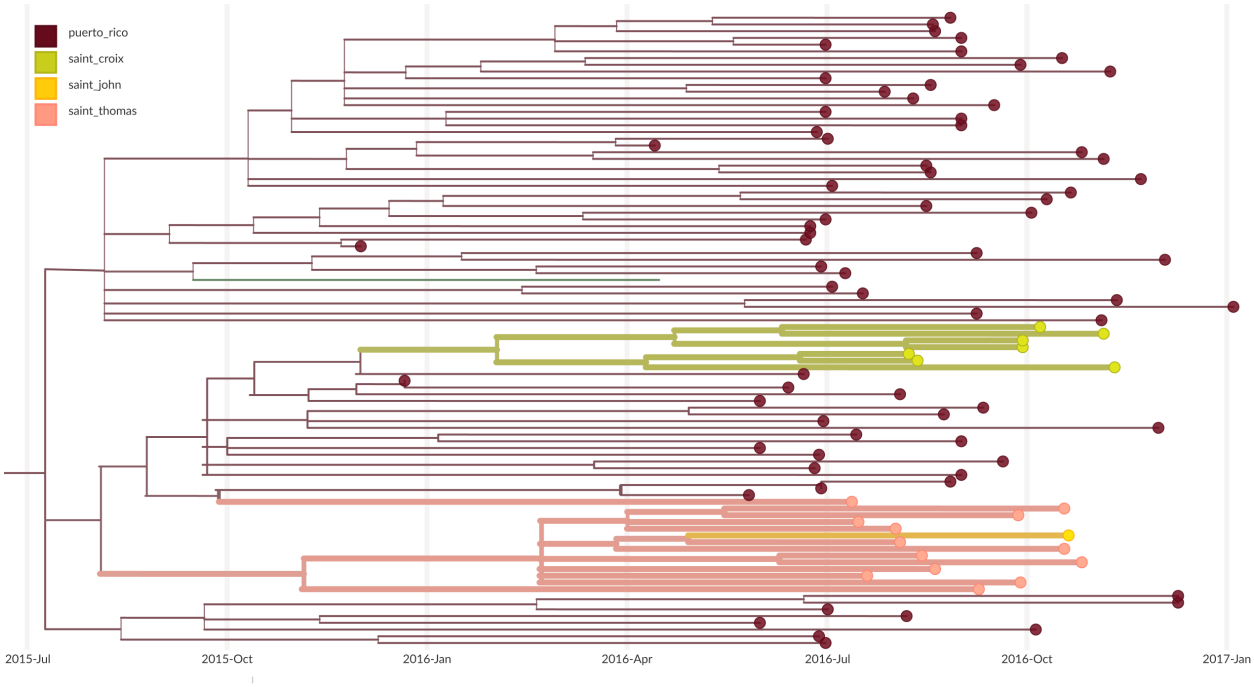

Figure 3.6. Phylogenetic reconstruction of introductions to the USVI from Puerto Rico. Temporally-resolved phylogenetic reconstruction of ZIKV introductions into the USVI, here, colored by the island from which a genome was sampled. We observe three introductions. Two introductions into Saint Thomas, and one introduction into Saint Croix.

The Dominican Republic was another highly sampled Caribbean country that appeared to be a source for ZIKV introductions to the USVI; one introduction into Saint Croix and one introduction to Saint Thomas nest within the genetic diversity of ZIKV in the Dominican Republic (Figure 3.7).

There are additional introductions to the USVI for which the transmission history is not well sampled. In these cases, branch lengths are long, and there may not be

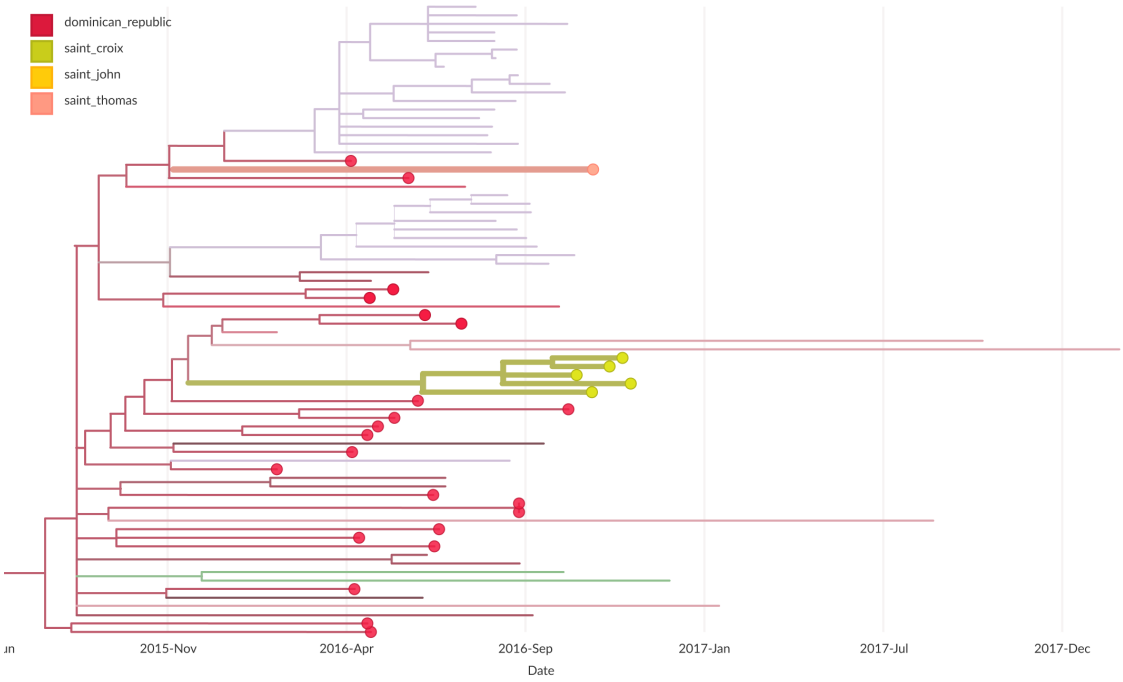

Figure 3.7. Phylogenetic reconstruction of introductions to the USVI from the Dominican Republic. Temporally-resolved phylogenetic reconstruction showing ZIKV introduction events into the USVI likely originating from the Dominican Republic. Here we see two introductions into the USVI: one introduction into Saint Thomas, and one introduction into Saint Croix.

a common ancestor between the USVI strain and another virus for as long as nine months. These long branches indicate remaining issues attributable to data paucity. Phylogeographic reconstructions will sometimes change with the addition of more data, not because they are inherently wrong, but because they can only reconstruct migration histories for which there are genomic data available. This property is why broad genomic datasets, sampled over time and across all affected geographies, are so critical to accurate genomic epidemiologic analysis.

Chapter 4

TEN RECOMMENDATIONS FOR SUPPORTING OPEN PATHOGEN GENOMIC ANALYSIS IN PUBLIC HEALTH SETTINGS

4.1 Introduction

Increasingly, public health officials are using pathogen genomic sequence data to support surveillance, outbreak response, pathogen detection, and diagnostics[1]. Sequencing cuts across traditional pathogen boundaries; for example, we can use it to distinguish cases of wild polio from vaccine-derived polio[6], or to predict the susceptibility of a tuberculosis infection to antibiotics[12], or to trace the source of a foodborne infection[1]. Because of its utility, public health agencies throughout the world are developing their capacity to perform genomic sequencing. However, with new data streams come new challenges; next generation sequencing (NGS) data must be transformed from its raw state to be interpretable, and many of the tools developed for sequence assembly and analysis are either expensive to license or require a high level of computational proficiency to use. This means that the capacity to perform genomic data assembly and analysis has not expanded as widely as the adoption of sequencing itself. In this chapter, I describe the current challenges that public health agencies face in supporting bioinformatics and genomic epidemiology, and provide recommendations for building a sustainable infrastructure that can be used across public health programs.

4.2 Methods

To investigate the current landscape of bioinformatics and genomic epidemiology in public health agencies, we conducted a series of long-form, semi-structured interviews with bioinformaticians, laboratory microbiologists performing sequencing, software engineers developing pipelines and workflow management software for public health, and epidemiologists acting upon inferences from genomic data. We aimed to get a broad perspective, interviewing individuals from different countries, working on a wide array of pathogens, and working in agencies with varied capacity for performing genomic analysis.

The interviews focused on the following topics: technical components of genomic analysis, considerations for genomic analysis specific to public health settings, and social issues surrounding genomic data. The interviews revealed various themes and consistent challenges related to supporting pathogen genomic analysis in public health agencies. Our recommendations seek to address those challenges, and describe strategies for building a sustainable, efficient, and effective bioinformatic infrastructure for the growing need in public health.

While we conducted interviews primarily with public health programs within the United States, our colleagues at the Africa Centres for Disease Control and Prevention (ACDC) led a concurrent effort to assess sequencing and bioinformatic capacity within African public health agencies. We reviewed each others' landscape analyses, finding many similar challenges within small public health institutions in the United States as exist in Africa. To ensure that our recommendations would be relevant across income settings, our colleagues at ACDC reviewed the recommendations outlined here for their appropriateness to public health settings in low- and middle-income countries.

4.3 Recommendations

4.3.1 Support data hygiene and interoperability via development and adoption of a consistent data model

The value of an isolate is dictated not just by its molecular characteristics. A sample needs context; who was the sample collected from, when was it collected, how was it collected? Without this information much of the value of the sample is lost, both from clinical reporting and data analysis standpoints. Despite the value of metadata, sequence records are frequently decoupled from the full constellation of epidemiologic data describing the sample. While there are various reasons why this occurs, breakdowns in data hygiene compromise the utility of the data and impact users' ability to interact with the data programmatically. This limitation poses a critical issue, as increasing amounts of data reduce public health officials' ability to manually interact with the data. Complete and structured data is fundamental to an informatic ecosystem that can work at scale.

This is a commonly recognized problem, and widely used genomic databases such as the Sequence Read Archive (SRA) have standards and formatting requirements for submissions that do improve data hygiene. However, there is still much room for improvement, especially because many sequences can map to a single set of epidemiologic data in a way that is complex and often hierarchical. As an example, imagine all the possible sequences that could be collected from a single individual infected with influenza. Patient demographic data, clinical data, and exposure information form an unchanging set of characteristics describing the individual at that point in time. However, that sample may yield many distinct pieces of genomic data. For instance, scientists can ascertain the consensus genome of the infecting strain by sequencing the clinical isolate. From that same dataset they may also have separate SNP calls describing within-host minority variants. Additionally, a lab may decide to culture the infecting strain and sequence the cultured isolates after different numbers

of passages. These scenarios yield sequences that are distinct from one another, and have different laboratory-associated data, but that share the same epidemiologic data. Ideally, the public health community would have a data structure and standardized vocabulary that accommodates and organizes these different data fields, linking them appropriately and annotating them consistently. We propose that the public health bioinformatics community, along with engagement from the major data repositories (e.g. NCBI, EBI, and DDBJ), develop common data models and adopt ontologies. To initiate this discussion, we describe an example of a data model below, and list ontologies that could be adopted more widely.

A data model for hierarchical genomic data.

As one example, the `https://nextstrain.org`[33], employs a custom database that canonicalizes pathogen genome data and associated metadata. Data are sourced from a variety of public databases (NCBI, GenBank, and ViPR) as well as from GitHub repositories if permitted by the owners. Because the data are pulled from various sources, different types of data are provided and data are often varied in format. Thus, the agglomerated dataset must be standardized according to a schema that can apply to all sequences. The Nextstrain data model (Figure 1) includes three major data fields: case, sample, and sequence. Each case record can have multiple samples, and each sample can have multiple sequences. Linkage between the fields is maintained by a case identifier and sample identifier that are logged as subfields. Within each field, subfields are tailored to record the most pertinent information to that field. For example, the case field contains information about host species, age, sex, symptoms, collection date, and geography. The sample field contains the case identifier, along with relevant information about the sample, such as collection date, collection medium (blood/urine/tissue), and culture information such as the cell line used and the number of passages that a sample underwent. This field also contains information about the lab that grew or tested the virus; this lab is often distinct from the clinic where the case presented. Finally, each sample may have multiple

sequences. The sequence field specifies the case identifier and the sample identifier, but again organizes information more pertinent to the sequences themselves, such as (1) what portion of the genome the sequence is from, (2) whether the sequence is a consensus sequence or a minor variant, (3) flags specifying whether the sequence is public or private, and (4) accession numbers if the sequences were pulled from a public database.

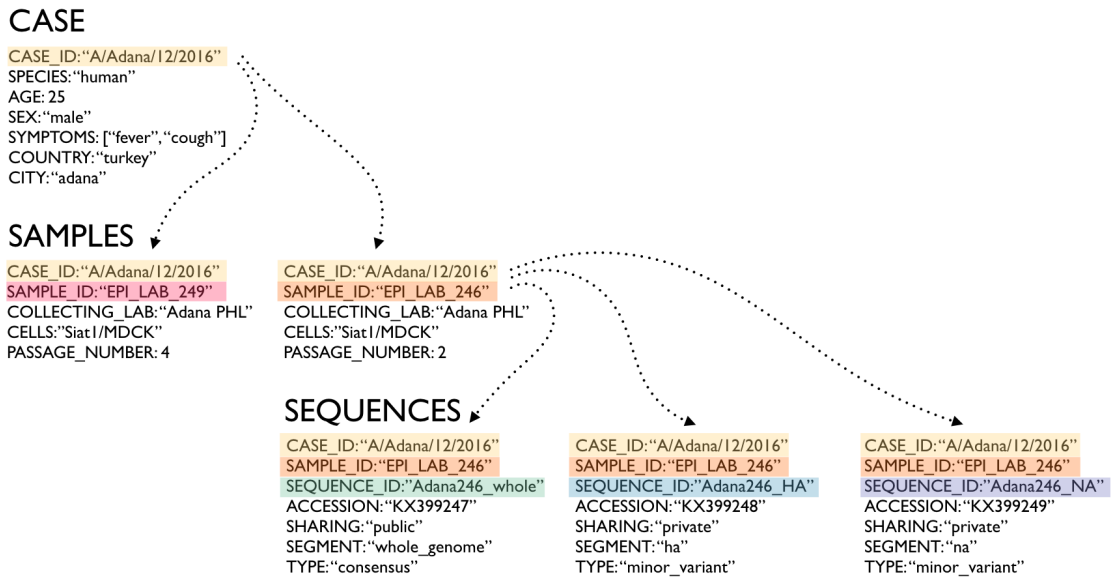

Figure 4.1. This schematic illustrates an example data model for hierarchical genomic data.

Ontologies for genomic epidemiology.

If widely adopted, the use of a common data model, or a set of common data models, will provide a unified framework for linking sequence data, clinical data, and epidemiologic information. However, to fully structure these data, public health programs will also need to adopt and/or develop ontologies that standardize free-form epidemiologic information about cases and their exposures. Two good examples of epidemiologic

ontologies are IRIDA's genomic epidemiology ontology, `https://genepio.org/`, and FoodON[11]. These ontologies create controlled, standardized vocabularies, which facilitate programmatic interaction with databases and enable users to automate quality control and analytic procedures.

4.3.2 Strengthen application programming interfaces (APIs)

Application programming interfaces, or APIs, are the mechanism by which users communicate with computers, code, and databases in an automated way. They are critically important for programmatic querying of databases, collation of disparate data sources, and communication between pieces of software within a greater ecosystem.

The relative paucity of consistent and well-documented APIs for software tools and databases affects public health bioinformatics in at least two ways. Firstly, the lack of APIs limits the scalability of bioinformatic analyses. Currently, querying genomic databases frequently requires human interaction via a web-based graphical user interface. However, with ever increasing amounts of data, the ability to manually explore, source, and distribute data will decline. Agencies will need to have tools for automated querying and communication, and the quality of APIs will directly affect the ease with which public health programs can perform these functions reproducibly and efficiently. Secondly, the lack of APIs leads to inefficient use of bioinformatician effort. When basic pipelines do not run automatically, or significant effort is required to link programs together into a pipeline, bioinformaticians spend large amounts of time writing interstitial code and managing file format conversions. Some interviewees noted that performing these tasks reduced their availability to do more sophisticated genomic analyses that might have greater public health utility.

The development and use of well-documented APIs will underlie the success of a software ecosystem within public health and cannot be an afterthought. Public health institutions should adopt API standards, and APIs should be developed in tandem with database or software development. For the many software programs and

databases that already exist, specific funding sources should be allocated to build or extend current APIs to function with the agreed-upon data models and adhere to adopted API standards.

Within the United States, openGSA, a division of the General Services Administration, develops and publishes APIs to provide open government data in machine-readable formats. Additionally, openGSA has developed standards that these government APIs must adhere to. These standards provide a concrete starting point in the development of APIs for genomic and epidemiologic databases, and are described in detail at `https://github.com/GSA/api-standards`. In brief, the standards provide the following guidelines:

(1) APIs should be RESTful, with clear, human-readable endpoints.

(2) They should return JSON objects for both API responses and error messages.

(3) APIs should be versioned, and they should be backward-compatible within a major version, though breaking changes can occur with major version changes.

(4) The API must have clear and readable documentation and allow users to report feedback or issues and ask questions.

(5) All APIs should use HTTPS.

4.3.3 Develop guidelines for management and stewardship of genomic data

The increasing abundance of longitudinally collected pathogen genomic sequence data is a valuable resource for public health. However, to fully realize the value of this data, programs will need to manage and care for the data in a unified manner. To this end, public health institutions should develop and adopt guidelines and standards for data

collection, annotation, archival, and reuse. These guidelines should be designed to ensure that data adhere to FAIR principles, such that archived data are **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable [73]. These principles and how they relate to genomic data are summarized below. Greater elaboration on data FAIRness is available in [73].

FINDABLE: Data observations need globally unique identifiers that are persistent. Data observations need to have rich descriptions of the data, including how the data were obtained. These observations would be specified by the data model discussed in Proposal 1. Data observations should be indexed.

ACCESSIBLE: Data should be retrievable using automated protocols that are open-access and universally implementable, and that provide authentication and authorization procedures. This principle is primarily met by the use of well-designed and documented APIs, as described in Proposal 2.

INTEROPERABLE: Data are represented with standardized vocabularies (ontologies) that also follow FAIR principles. This ensures that data are accessible and shareable across agencies.

REUSABLE: Data need to be richly described with relevant attributes about how they were generated and their provenance. Data released into open databases should have clear licenses describing how others can use the data.

Some of these principles are already being followed. The majority of agencies that we interviewed regularly submit raw sequencing reads to the SRA, which both archives the data for future use and publishes the data so that they can be found by other public health practitioners and scientists. Saving the raw data has the additional benefit of enabling users to completely re-run pipelines from start to finish, allowing comparison and beta testing of pipelines and analyses while protecting users from irreparable analytical mistakes[74].

While the consistent submission of raw reads to the SRA is an excellent first step, the community needs a formalized data stewardship framework that encompasses assembled genomic data as well. To promote findability and reusability, genomic assemblies should be annotated with information about how the assembly was made, such as what reference genome was used for mapping and what pipeline version was run. Submitted datasets should also be versioned. To ease the burden of archiving both raw reads and assemblies, submission mechanisms should be automated and fully integrated with bioinformatic assembly pipelines.

For further consideration of how to best preserve genomic data and curate high quality databases, please refer to Goodman et al (2014): Ten Simple Rules for the Care and Feeding of Scientific Data[28], and Hart et al (2016): Ten Simple Rules for Digital Data Storage[34].

4.3.4 Make bioinformatic pipelines fully open-source, broadly accessible, and transparent

Currently, one of the most frequently used software platforms for bioinformatic analysis in public health laboratories is BioNumerics (Applied Maths/bioMérieux). This commercial software package has played a central role in the development of PulseNet, a laboratory-based foodborne bacterial disease surveillance system that has been in use for more than 20 years. BioNumerics provides a bioinformatic analysis toolkit, but more importantly acts as a national database with specimen and process tracking and within-network data sharing. While many of the agencies that we interviewed appreciated the ease of use of the workflows and the integrated databasing capabilities of BioNumerics, multiple agencies also reported instances where the lack of modularity limited options for custom development and expansion, or where licensure costs impacted their ability to access the software. While BioNumerics provides developer resources, the field would benefit from moving towards a public health software ecosystem that supports the de-

velopment, adoption, maintenance, and hosting of a set of reference pipelines that are openly developed and completely accessible.

Ensuring that pipelines are accessible will require developers to consider pipeline users, who within public health may have limited to no formal training in bioinformatics. Currently, many public health agencies are undergoing a transition period in which they can generate large amounts of sequence data, but have more limited access to the technical expertise needed to assemble and analyze the data. We found that this need was generally harder to meet in smaller and lower-resource settings. Interviewees frequently mentioned that microbiologists or other laboratory technicians would volunteer their time to learn and transition to bioinformatic roles, but that the technical training necessary to build pipelines remained a significant hurdle. To help facilitate the standardized use of reference pipelines, public health agencies need a software ecosystem that readily provides access to easily deployed pipelines. To ensure that limited informatic training does not act as a barrier to use, reference pipelines should be containerized for easy deployment to a range of environments, wrapped to allow interaction via graphical user interfaces, and accessible via web-based entry. Additionally, pipelines should be completely open-source, make use of common, non-proprietary file formats, and be developed transparently in an environment that supports feedback and issue tracking. In our experience, many open-source projects and software are critically important, yet under-funded and developers over-taxed. We emphasize that this type of development effort will need to be appropriately incentivized, and will require large funding sources to support initial and sustained development and maintenance.

What might this deployment platform look like? We envisage a model where all reference pipelines and validation datasets are catalogued in a registry. Registry entries should provide information about the pipeline, such as what it does, what inputs it takes and outputs it provides, as well as information about where the pipeline is hosted and how to access it. These pipelines should be containerized so that individuals familiar

with command line interfaces can source and run a pipeline from a container hub (e.g. Docker, Singularity) using minimal shell scripts or pull and run commands. In addition, an open-source initiative could be funded to write and maintain graphical user interfaces for interacting with these same pipelines. Having graphical user interfaces will ensure that individuals with less familiarity with command line interfaces and programming can still access and use bioinformatic tools. Graphical user interfaces could either wrap the sourcing and running process, or wrap pipelines that are hosted directly on a specific public health server. These options are not mutually exclusive, and we imagine that both a broad registry and a more narrow shared computational service that hosts the most frequently used pipelines would co-exist. We emphasize that both registries and deployment platforms for shared computational services must host multiple generations of pipelines, clearly indicating which versions are vetted reference pipelines, which pipelines are beta tests of future generations, and which pipelines are deprecated. Having this versioning will allow pipeline improvement and development while maintaining access to the hardened reference pipelines.

The pipeline deployment platform should be developed as an open-source community project. Specific instances of the platform could be deployed on distinct compute infrastructures managed by various public health agencies or networks of agencies. The compute infrastructure could be cloud-based (see Proposal 7) or could use an in-house cluster. Creating and managing instances of the platform could be performed either by in-house computational personnel or could be performed as Software as a Service (SaaS), where SaaS is provided by a non-profit or a company that charges individual users for their compute. Notably, open-source platforms with SaaS options have been successful, as in the case of Arvados.

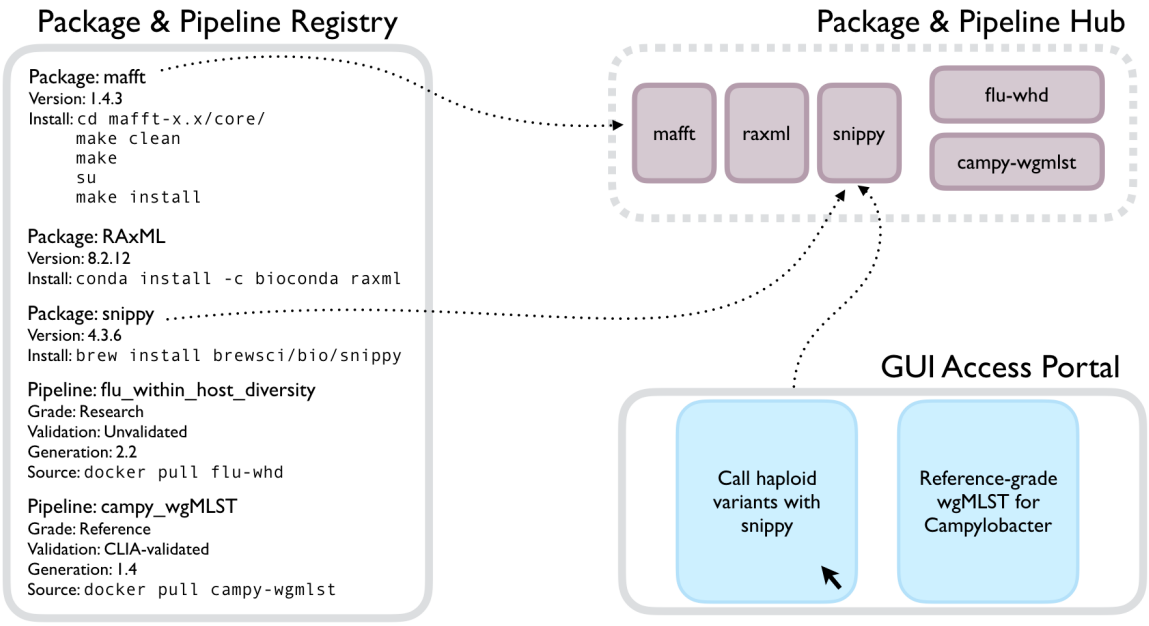

Figure 4.2. Three important components of this ecosystem: a package and pipeline registry, a hub where containerized packages and pipelines are hosted, and a graphical user interface access portal for running pipelines and packages.

The creation and maintenance of a bioinformatics infrastructure for the public health community will require ongoing social, political, and technical investment. In developing said platform, many different considerations must be tackled. These include, but are not limited to:

(1) developing and/or implementing validation criteria for bioinformatic analyses,

(2) developing new bioinformatics tools and pipelines,

(3) extending tools that already exist such that they work in the deployment platform environment,

(4) determining who will host pipelines and graphical interface portals to pipelines, and

(5) communicating the requirements of the community to developers and users. These requirements would include, for example, the data models, API standards, and pipeline documentation standards.

Governance of the deployment platform should focus on engaging with developers and users. This effort is necessary to ensure that development effort is sustainably supported, and to evaluate the effectiveness of the platform in providing open-access, low-barrier bioinformatics and genomic analyses.

4.3.5 Develop and support pipelines for data visualization, exploration, and automated analysis

Almost uniformly, interviewees wanted access to computational tools to visualize and explore genomic data. Their goals were frequently to understand their data better, to improve communication of genomic findings to surveillance epidemiologists, and to automate routine analyses and generation of reports. Currently, genomic visualizations such as phylogenetic trees, minimum spanning trees, SNP matrices etc., are often shared manually, by sending images over email. Surveillance epidemiologists may then manually add in epidemiologic data to look at how exposures or other relevant information correlate with the genomic data. This process is clunky and inefficient, and public health agencies would benefit from tools for automated data joining.

A variety of software for exploring and visualizing genomic analyses already exist. These include platforms such as Microreact, Nextstrain, and PHYLOViZ (discussed in more detail in the software section). Despite their value for understanding genomic data, these tools are variably implemented in public health settings. We found that the primary

impediments to deploying this software were insufficient technical knowledge to support an instance of the software, and the separation of epidemiologic and genomic data streams, which is done for security purposes. The lack of automated methods for rejoining genomic and epidemiologic data after bioinformatic processing limits the information that can be visualized on a phylogeny or other genomic data object, thereby reducing the utility of the visualizations. Development of data joining tools has been hampered at least in part due to the form of epidemiologic data, which may be variably complete and is often non-standardized, especially when collected during rapidly evolving outbreak investigations and response activities.

To ameliorate this situation, analytic and visualization pipelines ought to be separated from assembly pipelines. We emphasize this point because, while this division is relatively common in academic settings, the majority of bioinformatic pipelines we saw in public health programs ran from raw reads through to a genomic visualization in a monolithic series of computations. Taking a more functional, modular approach to bioinformatic software development will likely improve the flexibility, performance, scalability, and maintainability of public health bioinformatic applications and pipelines. Separating the assembly and analytic processes should also help overcome some of the issues of data classification and server security. Genomes can be assembled in the absence of epidemiologic data, and this can occur on lower security, scientific computing infrastructures. Then, subsets of epidemiologic data housed on secure servers can be joined with the assembled genomes, and the joined data objects can be run through visualization pipelines. Since the data should be sourced via API calls, different levels of security authorization can be required to source different components of the epidemiologic data with appropriate encryption and access controls. Analytical pipelines would export information for visualization in interactive browser-based portals. If high security is necessary, these could be served locally, and if not, they could be shared more widely over the web.

Notably, much of the infrastructure and code base for visualization pipelines has

already been developed; the challenge has been how to integrate and run these pipelines within public health agencies. To support this effort we need to: (1) standardize the structure of genomic data through the use of data models, (2) standardize epidemiologic data via adoption of ontologies, and (3) build the API infrastructure to source and join data streams while respecting security considerations. Again, we stress that the additional development effort needed to adapt current tools to work within a new software ecosystem, and to support them over time, will require sustained funding mechanisms.

4.3.6 Improve reproducibility of bioinformatic analyses

In Proposal 4 we described a deployment platform for public health bioinformatics. Here, we describe in greater detail how individual pipelines should be developed to ensure that they are highly reproducible. Consistent with common practice in academia, we found that many public health programs use similar, but distinct, pipelines for bioinformatic analysis; the wheel is frequently reinvented. While these pipelines all use a relatively narrow suite of the same open-source software programs, the lack of standardization across bioinformatic pipelines impacts the comparability of data and results across agencies. This lack of comparability was a major concern that was voiced frequently. The need for standardization and highly reproducible assays is particularly acute in public health. In contrast to academic settings, sequencing assays in public health need to be sufficiently robust and reproducible to meet government regulated standards.

In addition to stable reference pipelines, interviewees also expressed a need to use custom pipelines. Often this need was brought up by frontline health departments investigating questions of local public health importance. We imagine that there may in fact be a high degree of overlap in the types of questions an agency will investigate at the local level. Therefore, supporting wider access to non-reference pipelines may actually help to harmonize these analyses as well.

To have a high degree of reproducibility, agencies need to be able to use the same pipelines as each other, and the pipelines and their component software packages must be stably deployed. We discuss below possible strategies for supporting reproducibility. In particular, we consider how pipeline development can prioritize reproducibility, how the process of running pipelines can be made more reproducible via containerization and workflow management, and the need for rigorous auditing and validation to verify reproducibility.

Versioning.

When considering reproducibility, we must consider three aspects of pipelines: the data being assembled, individual programs being piped together, and the packaged pipeline as a whole. All reference datasets, component software programs, and whole pipelines should be version controlled. Versioning software components, in addition to full pipelines, allows all changes to be tracked and documented, and ensures that developers can roll-back undesired changes. Versioned pipelines can then also be cloned, allowing development of newer generations of the pipeline or offshoot pipelines without inhibiting access to the stable reference analysis. Facilitating the coexistence of reference and customized pipelines together is critical to widespread adoption of a single bioinformatic pipeline deployment platform.

Containerization.

Similar to versioning, developers should containerize pipelines and any software packages that are used outside of workflows (nested containers are not well supported at this time). We propose using containerized pipelines for the following reasons:

(1) Containerization increases the reproducibility of analyses because you can run the same pipeline in the exact same computing environment as someone else. This consistency in the compute environment limits issues where missing dependencies, or differences in versioning of dependencies, change the way a pipeline

runs.

(2) Containerized pipelines are shareable. Facilitating pipeline sharing enables agencies to run the exact same pipeline as each other, which is preferable to each agency attempting to build a similar pipeline from documentation.

(3) Containerization promotes software stability and reproducibility because a program can maintain replicate instances of a workflow. Being able to host old pipelines alongside new ones under development ensures that access to reference pipelines is maintained. Having replicate instances also allows developers to benchmark pipelines side-by-side. This ability to compare workflows systematically within the same environment is critical to ensuring that bioinformatic assays remain valid even as they are updated.

(4) Containerized pipelines can be run using automated workflow managers, a quality that improves reproducibility.

(5) Containerization reduces the burden of reproducibility; while someone must verify that they are using the correct version of a container, they do not then need to check the versioning of all component software and dependencies.

Containerizing software programs could be started fresh within public health, or could make use of other containerization projects, such as BioContainers, BioBoxes, or FlowCraft. Bioinformaticians within public health have also begun to containerize useful software, such as the library of docker builds maintained by the State Public Health Bioinformatics group (StaPH-B) (github.com/StaPH-B/docker-builds).

Containerized reference pipelines should be released as versioned generations. All generations of a pipeline should be concurrently hosted on the platform to ensure historical compatibility of bioinformatic analyses. New generations of pipelines should be thoroughly validated and released according to a stable release cycle. Additionally, beta versions should be released to the portal regularly to allow user testing and commenting.

Auditability.

Within a public health context, all transformations of data should be describable and recorded. Pipelines and workflows should create auditable reports that include the name and version of the program running, as well as which input parameters were used, especially if these vary from default. Pipeline runs should automatically store intermediate files in standardized formats. Having access to intermediate files is important as they can reveal the presence of discrepancies, and where they were introduced, within an analysis. Additionally, intermediate files help describe how data were transformed during the analysis.

Validation.

Despite previous work to develop validation datasets (described in [70]), many interviewees mentioned that they would benefit from further development of structured validation criteria for bioinformatic assembly pipelines. We suggest that agencies perform end-to-end proficiency testing of WGS protocols, including both the laboratory and bioinformatic portions of the assay. We imagine that agencies at higher levels of jurisdictional authority would be responsible for developing the validation metrics, given the need for these standards to be unified across all levels of public health. Finally, the bioinformatics deployment platform should clearly communicate which pipelines, at which generation, have been formally validated.

Workflow management.

One of best the strategies for writing reproducible and auditable pipelines is to design them as automated workflows. While pipelines can be written as single scripts, specifying pipelines in workflow languages creates self-documenting pipelines. To maximize portability, workflows could be written in Common Workflow Language (CWL), which would allow them to run on various deployment platforms such as Arvados, Terra(FireCloud), and eventually also on Galaxy. Specifying workflows in CWL also allows users to automatedly translate their workflows into other workflow languages. Alternatively, pipelines

could be written with other workflow systems, such as Snakemake or Nextflow. While potentially not as portable, these workflow systems have high uptake in biology, and may be more familiar to developers in public health.

4.3.7 Use cloud computing to improve the scalability and accessibility of bioinformatic analysis

To date, the adoption of cloud computing in public health has been hindered by issues with process, compliance, and acquisition of cloud services by governmental agencies at all levels of jurisdictional authority. However, as cloud services become increasingly feasible for government agencies to access, we expect the utility of these resources to increase. The ability to access cloud services will allow smaller jurisdictions, or those with limited infrastructure and resources, to support sophisticated bioinformatics capabilities without incurring significant capital or operational expenditures. This will be an important leveler for low and middle income countries, who can take advantage of a community-driven ecosystem of deployable and scalable bioinformatic tools and workflows.

Supporting accessibility.

Within the United States, adoption of BioNumerics gave frontline public health agencies greater autonomy to investigate diseases that were a priority at the local level. As we transition to an open-source software ecosystem, this autonomy needs to be maintained. Access to cloud-based bioinformatic analyses puts these tools at the frontline of public health, helping smaller public health agencies do more sophisticated analyses, and reducing response lags in outbreak scenarios. Using the cloud supports broad decentralized access to bioinformatics by ensuring that reference data, testing datasets, and pipelines are available from one place.

Using a cloud-based bioinformatic platform would also increase accessibility by de-

creasing economic burden on small or lower-resource labs. If using a cloud-based bioinformatic ecosystem, only one or a few high performance computing environments need to be managed. Thus not every institution needs to pay for server hardware or the highly remunerated workforce necessary to maintain a cluster, although they might have to pay for their usage of the cloud-based ecosystem unless centrally funded. Shifting to cloud computing converts capital expenses to operational expenses, which hopefully will make it easier to spend money on compute resources. While many agencies likely understand the traditional capital and operational expenditures associated with purchasing and maintaining servers, probably fewer currently have a good understanding of how cloud computing operational expenditures compound, and how to install necessary controls on them. Thus to ease expenditure concerns and smooth adoption, we propose that public health programs receive training on how cloud operational expenditures work, how to install controls, and how to train users who are purchasing resources.

Increasing capacity and efficiency.
Movement to a cloud-based analytic ecosystem allows dynamic scaling of compute resources, which allows the ecosystem to adapt to changing data storage needs, changing amounts of sequence analysis, and changes in the compute intensity of projects. While we absolutely face a growing need for compute power, that rise will have spikes and drops along the way. Cloud-based ecosystems can be designed to handle this variation elegantly, reducing the risk of compute underutilization or insufficient access to resources.

Cloud-based software ecosystems have been successfully used in academic settings to provide small labs with access to high performance computing and software, and developers could look at platforms such as the National Science Foundation supported XSEDE ecosystem as an example. That said, we recognize that there will be challenges to deploying cloud-based analyses for all possible public health scenarios. We imagine that the greatest challenges to overcome will be connectivity issues in lower resource settings and the need to update agency-specific data and patient privacy policies to permit cloud-

based computing. Overcoming these hurdles will require open communication between public health agencies and cloud computing providers, such that cloud services can be tailored to the needs and standards of public health agencies.

4.3.8 Support new infrastructure and software development demands with technical personnel

Adopting the proposals outlined above will require the support of a more technical, computationally-oriented workforce. In addition to bioinformaticians and genomic epidemiologists, supporting this infrastructure will require personnel with expertise in high-performance computing, cloud computing systems engineering, network/storage engineering, data science, and software development. Almost every program we interviewed mentioned that attracting and retaining this workforce was challenging for a number of reasons. Lower compensation, lack of access to newer technologies, and the frustrations of working within a government agency all affect workforce development.

What often does attract bioinformaticians and software developers to working in public health is the opportunity to have impact. The meaningfulness of work in public health and the ability to use one's expertise to tangibly improve people's lives are opportunities that private sector positions generally do not offer. Thus, in developing their technical workforce, public health agencies should highlight these factors when recruiting. Additionally, laboratory microbiologists are pivoting towards more bioinformatics-heavy roles, often by learning these new skills on their own. In addition to improving recruitment, public health agencies should support training programs that facilitate the transition from bench microbiology to bioinformatics. Due to their prior role, these individuals have an incredible wealth of knowledge about upstream sequencing process that can aid in troubleshooting and evaluating bioinformatic processes.

We found that successful recruitment of technical personnel into public health often occurred through connections with academic institutions. Generally, these relationships were forged when public health agencies reached out to access high-performance computing, or when they sought graduate-level students to work on informatic questions for thesis or practicum projects. Additionally, technical personnel can be recruited through fellowship programs. Successful examples include the CDC Public Health Informatics Fellowship Program and the APHL-CDC Bioinformatics Fellowship Program.

Recruitment is just one piece of workforce development, and interviewees mentioned that they also found it challenging to retain technical staff. Likely because bioinformatics and software engineering are new careers within public health, mechanisms for career advancement are not well developed at this time. We found that some agencies did not have formal job descriptions specific to computational disciplines, let alone competency and assessment criteria, or mechanisms to move into leadership roles. To sustain a computational workforce, public health agencies should create clear descriptions of the disciplines and job series for bioinformaticians, data scientists, and software engineers. We imagine that describing these positions and developing career trajectories will make public health a more attractive career option.

4.3.9 Improve the integration of genomic epidemiology with traditional epidemiology

From discussions with interviewees, we found that the degree of integration of genomic and traditional epidemiology varied highly across programs. Some divisions and institutions had open collaborations between bioinformaticians and epidemiologists, often with weekly meetings to discuss ongoing outbreaks or future surveillance directions. In other cases, we found that bioinformaticians had limited contact with epidemiologists; bioinformaticians would send out routine reports, but little information about how the genomic data were interpreted, or questions requiring follow-up, were communicated back. While

genomic and traditional epidemiology work synergistically, the training required to understand and analyze genomic data versus case data is distinct. Many individuals working in public health do not have both types of training. Thus, facilitating open communication and collaboration between bioinformaticians and epidemiologists is fundamental to supporting integrated surveillance systems.

We believe that integration of these domains can be improved by providing basic training in analyzing both types of data. Many of the bioinformaticians we spoke with said that they would appreciate having a better understanding of traditional epidemiologic methods. We interacted with fewer epidemiologists during our interviews, however those we spoke with commented that they found interpretation of genomic data challenging without specific training. Potential ways to provide this training include online and in-person courses. Interviewees mentioned that sustained courses, that met once a week over a longer period of time, were more helpful as this schedule allowed time to practice newly acquired skills and ask follow-up questions. In addition to providing further training, it might be helpful to have a team of genomic epidemiologists and bioinformaticians available to work with different agencies temporarily, providing training and support to onboard new techniques and systems for genomic epidemiologic analysis.

Integrating genomic and traditional epidemiology will require technical developments in addition to the social ones discussed above. From a technical standpoint, integration of genomic and epidemiologic data will require more sophisticated databasing approaches, included programmatic data sourcing and merging that respects security levels, use of ontologies to standardize data reporting formats for both surveillance data and genomic data, and machine-learning methods for data classification, tagging, and cleaning. The need to unify and harmonize systems to improve genomic epidemiology has the added benefit of developing a database infrastructure that could facilitate cross-agency data sharing. During the course of our interviews, interviewees frequently mentioned the need to see beyond human surveillance data in order to fully comprehend the source and

extent of an outbreak. Ideally, the use of standardized data models and ontologies, as well platforms accessible via cloud computing, would unify analytic platforms and standards in multiple agencies. While cross-agency integration would be politically challenging, it would exponentially improve public health programs' ability to perform surveillance within a One Health paradigm that recognizes the intersectionality of environmental, animal, and human health.

4.3.10 Develop best practices to support open data sharing

In an interconnected world where disease transmission occurs across borders, environments, and species, the best surveillance system would support data sharing across institutions and agencies, both within country and between countries. Every agency we interviewed understood the value of data sharing, and many had anecdotes where their inability to share or receive data hampered surveillance activities. Despite the recognized value of sharing data, putting it into practice is challenging. In part, this is due to the critical need to protect patient privacy. Public health programs rightfully must follow rules that govern how personally-identifiable information (PII) is shared. However, in practice these rules can make data sharing convoluted, since definitions of PII vary by disease incidence and geography, and different places have different laws and protections governing the use, storage, and transmission of PII. In order to develop a data sharing system that functions well for public health, we think that data sharing needs to:

(1) be easy to do, so that it is not a burden,

(2) occur along trusted channels, and

(3) be granular, so that access to different levels of data can be filtered based on security and legal constraints.

Large, diverse genomic datasets from many groups are greater than the sum of their

parts, and their utility has built momentum for greater data sharing within some sectors of public health. Perhaps the best example of this is PulseNet, a large, multi-agency network that supports within-network data sharing. In line with our proposals, PulseNet data sharing occurs along trusted channels, built on memorandums of understanding with each of the collaborating partners. These memorandums describe how data will be shared, with whom, and at what granularity, ensuring compliance with state and federal law. Importantly, the PulseNet sharing system is also relatively easy to use, and is integrated into BioNumerics. Detailed and complete data about a sample can be added to local BioNumerics databases, and subsets of that data can be shared with PulseNet via easy interaction with the BioNumerics graphical user interface.

While PulseNet's efforts have pushed data sharing forward immensely, we still find instances where released metadata are too coarse to provide meaningful genomic epidemiologic inference. To make improvements, we encourage a deeper conversation about identifiability, including what the risks are and what information we might lose by masking data. This discussion could be initiated by considering the following questions, adapted from Cologne et al[7]. What is the probability that someone could identify a case given the released metadata? What are the consequences of identification, should it occur? How might masking or omitting metadata affect their analytical utility? The answers to these questions will vary by pathogen, by disease incidence, by geography, and by host. As such, maintaining the integrity of data sharing will require frequent re-evaluation of the risk of identifiability, consequences of identification, and assessment of how analytical utility may be lost when masking data.

A secondary concern about data release is scooping, a process in which another group analyzes and/or publishes on data without permission of the data generators. This is a consistent concern with data sharing also found outside of public health within academia. While this concern is hard to allay, in our experience scooping is more rare that one imagines it will be. Additionally, as agencies develop their capacity to perform thorough

genomic epidemiologic analysis quickly, data analysis will occur on roughly the same time scale as data release, which should also reduce the risk of getting scooped.

We emphasize that the development of increased data openness in public health cannot be all or nothing; if it is, we will simply end up with a system where sharing is limited. Instead, we should identify consistent small steps that can be taken to improve the openness of data, with the hope that open data and integrated databases improve surveillance and outbreak response sufficiently to warrant their continued development and maintenance.

4.4 Current software platforms and programs

Whole genome sequencing (WGS) is now a routine component of molecular biology, and there are a wide variety of applications that transform, analyze, and visualize WGS data. However, many of these tools run from UNIX-based command line interfaces that require technical knowledge to use. Additionally, some bioinformatic processes may be computationally intensive, requiring access to high-performance computing and knowledge of how to use cluster or cloud-based server infrastructure. This creates a mismatch between the currently available infrastructure and workforce within public health and what would be ideal to support large pathogen genomics programs. As the frontline for outbreak response and surveillance, public health institutions have broad mandates for WGS, usually requiring large amounts of sequencing and analysis of many different pathogens. Despite this, many agencies currently have minimal access to expertise in high-performance computing and bioinformatics. This means that an ideal informatic ecosystem for public health would be sufficiently accessible and intuitive to use that individuals would need only minimal formal training to analyze and visualize genomic data.

Compared to the large numbers of individual bioinformatic applications, there are

fewer platforms that manage WGS data storage, host full pipelines for bioinformatic assembly, and provide visualization of the assemblies in a unified manner. We feel that the primary software hindrance to pathogen genomics in public health is not necessarily lack of access to bioinformatic tools, although this is certainly an additional problem in institutions where employees cannot access off-network computers. Rather, the greater need is to develop systems that perform sample management, automated storage and sharing, and host reproducible pipelines integrated with visualization and results sharing. By discussing what an ideal ecosystem might look like, we hope to improve the interoperability and usefulness of the platforms and software that already exist. We describe below some of the current platforms and tools that public health agencies regularly use.

4.4.1 Unified platforms for databasing and workflow management currently used in public health

BioNumerics.

Developed by Applied Maths (now bioMérieux) for standardized gel analysis in the era of pulsed field gel electrophoresis (PFGE), BioNumerics now also supports WGS analysis and is widely used in public health. BioNumerics provides a graphical user interface with access to multiple pipelines for genome assembly and analysis, including calling allele profiles and making minimum spanning trees. While interviewees appreciated the graphical user interface, they consistently mentioned that the greatest value of the software is the way it organizes and performs databasing, providing efficient sample management, archiving, and data sharing interfaces that are integrated with the bioinformatic pipelines. BioNumerics supports sourcing raw data from NCBI, pushing data to NCBI, and automated storage of various 'experiment types' (e.g. allelic profiles, SNP matrices) as these results are generated, in an easily searchable backend database. BioNumerics can run locally or make use of cluster and cloud-based servers. This means that users can access

high-performance computing or analyze data that can't be shared on their local system. Within the United States, this platform has served to create a distributed network where state and local-level agencies have relatively high autonomy, something that any new software ecosystem must also support.

While BioNumerics has helped to bridge the transition from PFGE to WGS, there are a variety of limitations to the software. Most critically, BioNumerics is not open-source, and licensure costs can be prohibitively expensive for small institutions and institutions in low and middle income countries. Additionally, pipelines in BioNumerics generally run straight through from raw data to a visualization endpoint. While this type of pipeline is easy to use, the lack of modularity limits the types of comparative genomic analyses that can be performed, and impacts the user's ability to flexibly change, and interact with, the genomic visualizations.

Integrated Rapid Infectious Disease Analysis (IRIDA).
Developed in Canada as a collaborative effort between the Public Health Agency of Canada, provincial public health agencies, and Canadian academic partners, IRIDA is a free, fully open-source software ecosystem for performing pathogen genomic analysis. The IRIDA platform performs data management and storage, and provides graphical user interfaces for bioinformatic pipelines executed by Galaxy. This provides users both the accessibility of an end-to-end workflow, and the flexibility to adapt or extend pipelines as desired. In addition to pipelines for bioinformatic assembly, IRIDA also supports integrated analytic pipelines that perform sequence typing, antimicrobial resistance prediction, and phylogenetic and phylogeographic reconstruction. Results and visualizations can be stored in IRIDA or transferred out to other applications and databases via a REST API. While IRIDA encourages and facilitates data sharing, independent instances of IRIDA can be run locally if data cannot be shared outside of an agency. Uniquely, IRIDA has a large emphasis on data standards, and will support the use of controlled vocabularies (ontologies) for genomic data and epidemiologic metadata, with the aim of

standardizing data reporting and facilitating inter-agency data harmonization.

IRIDA has many of the qualities that we would seek to have in a wider public health software ecosystem; indeed it was built specifically to fill this gap in Canada. However, one of the major challenges facing IRIDA, and one that other public health programs will have to consider as well, is the support model. Despite allowing decentralized genomic analysis, IRIDA is almost entirely centrally supported. Canada's reference laboratory, the National Microbiology Laboratory, provides the high-performance computing infrastructure for all provincial laboratories that use IRIDA, and is the only agency that can fully support software hosting and development in house. This centralized support model works to provide software and analytic infrastructure, but limits the degree of community support to continue growing the platform. This places a large amount of responsibility and ownership on a single group. Ideally, to transition this platform to a community-supported model, deployment of the ecosystem would be accompanied by targeted development of in-house capabilities at the regional laboratory level.

INNUENDO.

INNUENDO was developed by a large number of European partners with the goal of creating a unified platform for foodborne bacterial genomic assembly and analysis. The consortium brings together academic and governmental agencies from Finland, Estonia, Latvia, Portugal, Austria, and the Basque Autonomous Community in Spain. They also come from various sectors, including food safety, animal health, and human health. Similarly to IRIDA, INNUENDO is a fully open-source software platform that hosts pipelines for bioinformatic assembly and genomic analysis that are accessible via graphical user interfaces. INNUENDO also supports browser-based visualization, using a REST API for data transfer. INNUENDO uses pipelines built with flowcraft, an open-source application that allows easy, modular assembly of bioinformatic pipelines. Within this framework, each bioinformatic application called as part of the pipeline is containerized in Docker, ensuring greater reproducibility. These predefined workflows are also highly

auditable, creating run reports with versioning and command information. Similarly to IRIDA, INNUENDO has many of the qualities that we seek from a bioinformatic ecosystem for public health, and it should be looked at closely as a model.

PathogenWatch.

PathogenWatch is an easy-to-use platform for genomic surveillance, consisting of a datastore for genomic assemblies and metadata and a web client for processing and visualisation. Broadly focused on contextualising genomes within larger datasets and rapidly delivering results, PathogenWatch performs analytics on genomic assemblies, firstly identifying species, then performing species-specific analyses, such as multi-locus sequence typing (MLST), core genome MLST, the prediction of antimicrobial resistance profiles, and the inference of phylogenetic trees. PathogenWatch also hosts curated pathogen-specific genomic datasets that users can explore and use to contextualize their own data. PathogenWatch utilises a dockerised plug-in architecture allowing the inclusion of additional informatic pipelines created by the community. PathogenWatch provides an integrated database and visualization system.

4.4.2 Workflow platforms

EDGE.

Developed at the Los Alamos National Laboratory, EDGE is a bioinformatic workflow platform that provides users with a web-portal to access bioinformatic pipelines via a graphical user interface. This platform makes genomic data processing and analysis accessible to users without extensive bioinformatic experience.

Galaxy.

The Galaxy Project is an open-source, web-based bioinformatic workflow platform notable for its graphical user interface, which allows users to make reproducible bioinfor-

matic workflows even without knowledge of computer programming.

4.4.3 Application-specific platforms

IDseq.

IDseq is a centralized metagenomics platform that performs taxonomic identification of pathogens from uploaded FASTQ reads.

Mykrobe.

Mykrobe is an accessible, graphical user interface-based application that predicts the drug resistance profile of a pathogen given whole genome sequence information. Prediction is currently available for *Staphylococcus aureus* and *Mycobacterium tuberculosis*.

4.4.4 Visualization and data exploration software

Visualization platforms are already used to support public health surveillance. However, without computational expertise, it can be challenging to implement instances of the software. To make visualization software more accessible, these packages should be integrated into the greater bioinformatic ecosystem. For integration to work, we will need to have standardized assembly output formats and genomic data models that software applications can work with, as well as well-designed APIs to facilitate data transfer. Below we detail a subset of the genomic data visualization tools that are often used in public health.

Microreact.

Microreact is a flexible web application for linking and visualizing geographic, temporal, phylogenetic, network, and epidemiologic data. Microreact enables users to easily upload or link to files containing metadata, and/or tree and network files, to create a range

of visualizations, including trees decorated with metadata, maps, timelines, and tables. Visualizations in Microreact are easy to share via permanent web links, and can be linked to within publications. Furthermore, a comprehensive API enables the extension of Microreact to dynamic data. The application is used extensively by the European CDC, the US CDC, and Public Health England.

Nextstrain.

The Nextstrain software suite performs inference and visualization of maximum likelihood and time-resolved phylogenetic trees. It has the capacity to annotate tips of the tree given sample information, and also reconstruct states at internal nodes in the tree. Additionally, Nextstrain allows visualization of alignment and sequence characteristics, such as reconstructing nucleotide and amino acid mutations along each branch in the tree. The Nextflu subsidiary, which infers changes in antibody cross-reactivity across the tree, is used by the World Health Organization to inform vaccine selection, and by the CDC Influenza Division to support surveillance. Nextstrain is open-source, allowing independent groups to create custom Nextstrain builds for their own datasets.

PHYLOViZ.

Developed by the INNUENDO consortium, PHYLOViZ can infer allelic profiles and minimum spanning trees, and visualize associated metadata about pathogens on those trees. Additionally, the software will visualize distance matrices showing genetic distance between strains, and allows organization of that matrix by epidemiologic metadata. The software is accessible via a Java desktop application or browser-based application.

MicrobeTrace.

MicrobeTrace is a web application-based inference and visualization tool for drawing contact networks and transmission graphs given epidemiologic and genomic data. It allows inference and exploration of an underlying network structure, and can annotate nodes and edges with epidemiologic data. MicrobeTrace is used frequently by CDC to

support contact tracing efforts for HIV outbreaks. MicrobeTrace can also be used offline, which is an important quality for security considerations.

NCBI Pathogen Detection.

NCBI Pathogen Detection is a web-based application that facilitates automated genomic surveillance of 30 pathogens (29 bacterial and one yeast). This platform is heavily utilized for foodborne and health care acquired pathogen surveillance by US federal and state public health labs, along with international counterparts. The NCBI Pathogen Detection pipeline analyzes raw SRA data submitted to flagged surveillance projects, along with all assembled genomes submitted to the INSDC, totaling over a half million isolates at the time of this writing. The pipeline clusters isolates and infers phylogenetic trees that can be interactively explored in real-time to identify genetically linked isolates (e.g. from food, the environment, and human cases) or emerging antibiotic resistance.

4.5 Our vision of a potential software ecosystem

Given our proposals, and the software tools that currently exist, what do we envisage an open-source software ecosystem in public health might look like? We imagine that the system would benefit from being highly modular, with genomic assembly and data processing separated from genomic analysis and visualization processes. Splitting these processes will maintain efficiency while allowing flexibility, enabling many different analyses to be performed without having to rerun assembly pipelines. Importantly, separating the assembly and the analytic processes also ensures that output from the assembly pipelines is archived, an important extension to current archival practices that focus primarily on storing raw sequencing reads. The primary pieces of this ecosystem would be databases, APIs, pipelines, and scripts that move data around.

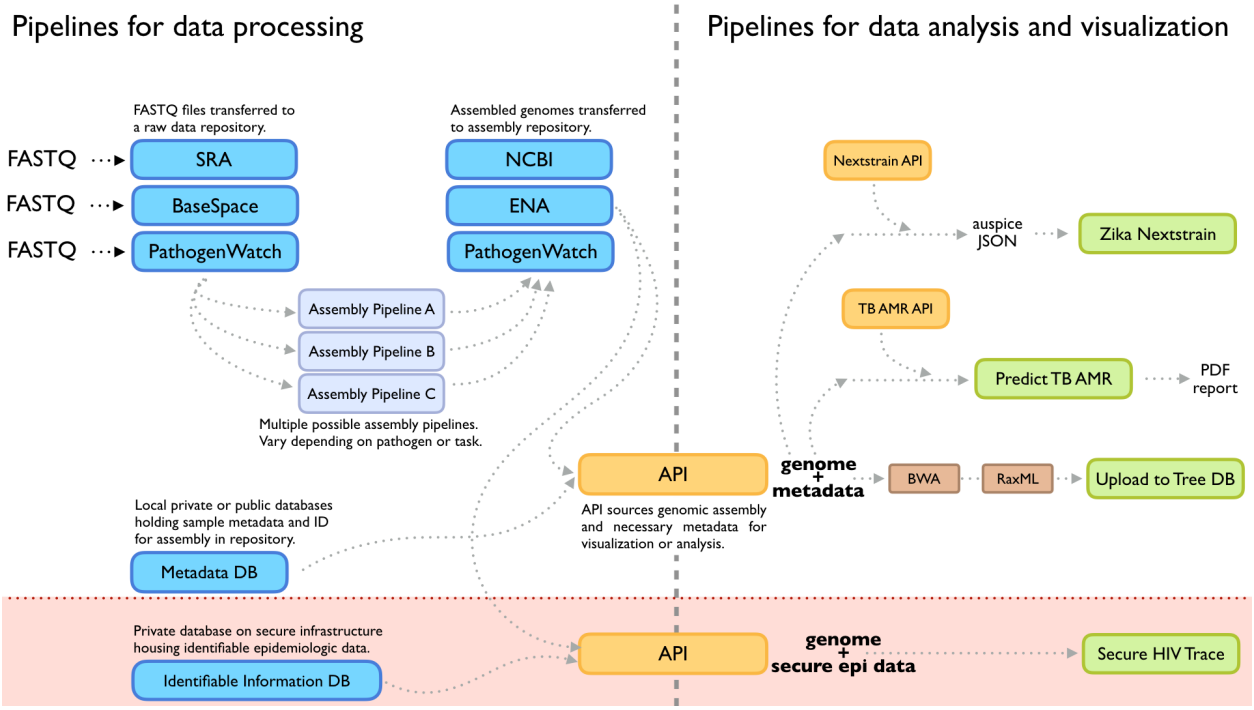

Figure 4.3. This schematic illustrates our vision of how an ecosystem in public health for bioinformatic assembly and genomic epidemiological analysis might look. We envisage a system where bioinformatic workflows are separated from genomic analysis and visualization workflows, with interaction and data sourcing mediated by APIs.

On the data assembly side, we consider three distinct types of databases: one for archiving or holding raw sequencing reads, one for archiving assembled data, and one for holding metadata about the samples. A variety of current databases could fill these positions, or the field could develop new databases if public health programs require additional utility. We imagine that the Sequence Read Archive would continue to serve as the primary raw reads database. But, if one has a metagenomic sample containing both pathogen and human reads, the reads could easily also be held in Illumina's BaseSpace platform instead. From here, raw reads could be assembled by one or more of the open-access pipelines; pipeline choice would be based around what type of assembly the

user needs. A final portion of the pipeline should be the automatic depositing of the genomic assembly into the relevant database for that assembly type. This database could be a relevant NCBI database (e.g. NCBI Nucleotide, NCBI Pathogen Detection), any database that is part of the International Nucleotide Sequence Database Collaboration (e.g. DDBJ, ENA), or a pathogen specific assembly database (e.g. GISAID, ViPR). The critical component of this databasing is ensuring that the accession identifier is deposited into a third database, the metadata database. Within our design, the metadata database would be an in-house relational database that facilitates sample tracking and houses all relevant clinical and laboratory data about the sample according to a well-defined schema that can also accommodate long form entries. Likely, the metadata database would be better to license than to build, however we do not have a specific databasing platform that we suggest at this time. Importantly, metadata databases could also be secured, and house relevant personally identifiable patient information collected during epidemiologic investigations. Having these data live separately from the genomic data ensures that PII can be kept private when necessary. Data linking would occur via API calls; calls to the metadata database would pull relevant sample information and the assembly accession number, allowing the assembly to be sourced from the genomic database. Various metadata+genomic data combinations could be sourced depending on what data fields are necessary for the desired analytic or visualization pipeline.

Once genomic assemblies and relevant metadata have been combined, they would be piped to various analytic workflows, such as predicting antimicrobial resistance, making specific data structures such as phylogenetic trees, or preparing datasets or data objects for serving to interactive data visualization platforms. We imagine that there will be a wide array of different visualization and analytic pipelines in use; good APIs and complete, standardized metadata are necessary to support that breadth. Some of these analytic pipelines may be completely containerized end-to-end workflows that produce visualizations or reports. Others could make data objects, such as phylogenies, and

submit these to a database for use in subsequent analyses, which could save compute time. Additionally, these pipelines could make API calls to external databases, such as antimicrobial resistance gene databases, thereby facilitating the integration of these new pipelines with tools and ecosystems that have already been built.

4.6 Conclusion

The shift toward extensive use of pathogen whole genome sequencing represents a turning point for public health agencies; agencies must pivot to accommodate a new data source that provides increased resolution for understanding disease dynamics, but that requires different tools and a changing workforce to support. While change poses challenges, we hope that these proposals provide direction that supports the public health community as it transitions.

Chapter 5

CONCLUSION: OPERATIONALIZING GENOMIC EPIDEMIOLOGY

In chapters 2 and 3 I discussed genomic epidemiological studies of ZIKV. These studies revealed differences in epidemiological dynamics between a large country affected early on in the epidemic, Colombia, and a small country where ZIKV arrived later during the epidemic, the USVI. While they help us understand ZIKV epidemiology, and they might help us to prepare for or understand subsequent outbreaks, they do not change the trajectory of the epidemic that occurred. In this way they are limited; they are not actionable in real-time.

To be useful in public health, we must conduct genomic surveillance and case-based surveillance on the same time scales, and and bring together applied epidemiological analysis of case data with comparative genomic analysis, such that analysis of both data streams can reinforce and evaluate each other. To move genomic epidemiology into this actionable realm, it may be useful to consider the genomic epidemiology pipeline (Figure 5.1), and ways in which we can move through it more efficiently.

Each of these pipeline components has been, and is being, iterated upon to improve the actionability of genomic epidemiological analysis. Ten years ago we would likely have brought samples back from the field to sequence in a lab at home. Now we have the ability to sequence samples in the field, using a molecular biology laboratory that one can pack in a suitcase [63]. We now routinely automate bioinformatic analyses through the use of workflow management software, and we can make data sourcing faster and more

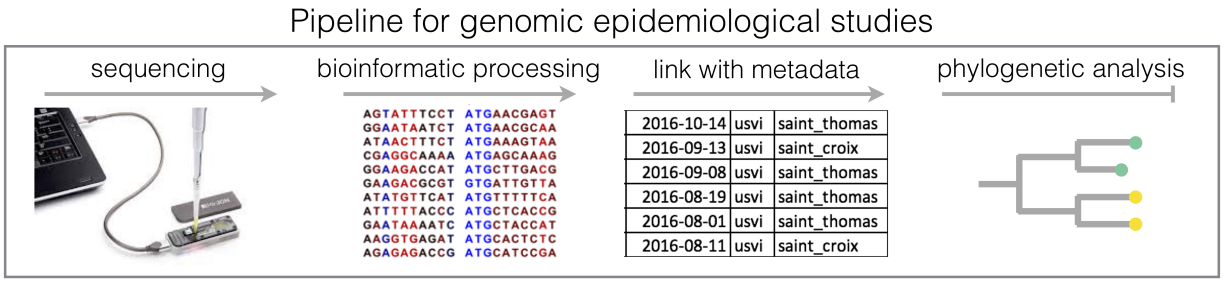

Figure 5.1. Broad steps in conducting a genomic epidemiological study. Conducting a genomic epidemiological study requires one to generate sequence data, assemble genomes from the raw sequencing data, collect and link additional epidemiologic data about the case to the sequence sampled from that case, and then finally jointly analyze the genomic data and metadata.

accurate with standardized data models and APIs. The field has developed, and continues to develop, new ways to make phylogenetic inference faster and more interpretable. And finally, as the utility of genomic data has become more clear, collaboration between traditional epidemiologists and genomic epidemiologists has become more routine, facilitating joint analysis of both data sources. These improvements are coming together to shift genomic epidemiology from an academic field to a routine part of applied public health.

One such example is the integration of genomic epidemiology into the outbreak response for Ebola virus disease in the Democratic Republic of the Congo (DRC). The DRC declared its tenth Ebola virus disease outbreak in July 2018, which has primarily circulated in the Nord Kivu province. In addition to standard epidemiologic surveillance and response efforts, the Institut National de Recherche Biomédicale (INRB) implemented a genomic surveillance system which includes viral whole genome sequencing from diagnostic specimens, bioinformatic analysis and data sharing through Nextstrain, and dissemination of genomic epidemiologic results to frontline public health workers. At the time of writing, the INRB had sequenced 675 Ebola virus genomes, representing an

unprecedented 20% of all laboratory-confirmed infections. For context, the 2013-2016 West Africa EVD epidemic was notable for sequencing 5% of reported EVD cases [15]. Most importantly, all of the sequencing, and most of the bioinformatic analysis, is being conducted within the DRC by Congolese scientists. The value of building capacity within-country is demonstrated not only by that achievement, but also by the sustainability of a system that can be shifted to other surveillance efforts as well. Indeed, using this same genomic surveillance system, INRB is now sequencing SARS-CoV-2 cases in the DRC.

The development of a genomic surveillance system within the DRC has enabled actionable information sharing between scientists and epidemiologists coordinating the day-to-day response on the time scales necessary to guide response efforts. This system has allowed us to monitor the genomic data for superspreading events, differentiate closely linked cases from potentially cryptic propagated transmission, infer regions that act as transmission sources or sinks, and detect EVD sexual transmission from survivors. Most importantly, via collaboration between the INRB and the DRC Ministry of Health, we have been able to communicate genomic findings to frontline epidemiologists, thereby directly informing outbreak response efforts.

The addition of genomic data to traditional epidemiologic data improves our ability to support contact tracing and evaluate interventions. Drawing inferences from multiple data sources can provide greater confidence in inferred epidemiologic dynamics, and also pinpoint weaknesses or erroneous findings in one data stream. It is my hope that we will increasingly integrate genomic and traditional surveillance, both in outbreak settings and in routine surveillance. I believe that unified databases linking epidemiologic information, laboratory information, and genomic data will support collaborative analysis between genomic and traditional epidemiologists. Their ability to work together, in real-time, will help further strengthen our public health surveillance systems.

BIBLIOGRAPHY

1. Gregory L Armstrong, Duncan R MacCannell, Jill Taylor, Heather A Carleton, Elizabeth B Neuhaus, Richard S Bradbury, James E Posey, and Marta Gwinn. Pathogen genomics in public health. *New England Journal of Medicine*, 381(26):2569–2580, 2019.

2. Maite Aubry, Anita Teissier, Michael Huart, Sébastien Merceron, Jessica Vanhomwegen, Claudine Roche, Anne-Laure Vial, Sylvianne Teururai, Sébastien Sicard, Sylvie Paulous, Philippe Desprès, Jean-Claude Manuguerra, Henri-Pierre Mallet, Didier Musso, Xavier Deparis, and Van-Mai Cao-Lormeau. Zika virus seroprevalence, French Polynesia, 2014–2015. *Emerging Infectious Diseases*, 23(4):669–672, 2017.

3. Trevor Bedford, Sarah Cobey, and Mercedes Pascual. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evolutionary Biology*, 11(1):220, 2011.

4. Trevor Bedford, Steven Riley, Ian G Barr, Shobha Broor, Mandeep Chadha, Nancy J Cox, Rodney S Daniels, C Palani Gunasekaran, Aeron C Hurt, Anne Kelso, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220, 2015.

5. Guilherme Calvet, Renato S Aguiar, Adriana S O Melo, Simone A Sampaio, Ivano de Filippis, Allison Fabri, Eliane S M Araujo, Patricia C de Sequeira, Marcos C L de Mendonça, Louisi de Oliveira, Diogo A Tschoeke, Carlos G Schrago, Fabiano L Thompson, Patricia Brasil, Flavia B dos Santos, Rita M R Nogueira, Amilcar Tanuri, and Ana M B de Filippis. Detection and sequencing of Zika virus from amniotic

fluid of fetuses with microcephaly in Brazil: a case study. *Lancet Infectious Diseases*, 16(6):653–660, 2016.

6. Centers for Disease Control and Prevention. Laboratory surveillance for wild and vaccine-derived polioviruses, January 2002– June 2003. *Morbidity and Mortality Weekly Report*, 52(38):913, 2003.

7. John Cologne, Eric J Grant, Eiji Nakashima, Yun Chen, Sachiyo Funamoto, and Hiroaki Katayama. Protecting privacy of shared epidemiologic data without compromising analysis potential. *Journal of Environmental and Public Health*, 2012, 2012.

8. Institute of Medicine Committee for the Study of the Future of Public Health, Division of Health Care Services. The Future of Public Health, 1988.

9. Esther Liliana Cuevas, Van T Tong, Nathaly Rozo, Diana Valencia, Oscar Pacheco, Suzanne M Gilboa, Marcela Mercado, Christina M Renquist, Maritza González, Elizabeth C Ailes, Carolina Duarte, Valerie Godoshian, Christina L Sancken, Angelica Maria Rico Turca, Dinorah L Calles, Martha Ayala, Paula Morgan, Erika Natalia Tolosa Perez, Hernan Quijada Bonilla, Ruben Caceres Gomez, Ana Carolina Estupiñan, Maria Luz Gunturiz, Dana Meaney-Delman, Denise J Jamieson, Margaret A Honein, and Martha Lucia Ospina Martínez. Preliminary report of microcephaly potentially associated with Zika virus infection during pregnancy - Colombia, January-November 2016. *Morbidity and Mortality Weekly Report*, 65(49):1409–1413, December 2016.

10. William E Diehl, Aaron E Lin, Nathan D Grubaugh, Luiz Max Carvalho, Kyusik Kim, Pyae Phyo Kyawe, Sean M McCauley, Elisa Donnard, Alper Kucukural, Patrick McDonel, et al. Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. *Cell*, 167(4):1088–1098, 2016.

11. Damion M Dooley, Emma J Griffiths, Gurinder S Gosal, Pier L Buttigieg, Robert

Hoehndorf, Matthew C Lange, Lynn M Schriml, Fiona SL Brinkman, and William WL Hsiao. FoodOn: A harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Science of Food*, 2(1):23, 2018.

12. Ronan M Doyle, Carrie Burgess, Rachel Williams, Rebecca Gorton, Helen Booth, James Brown, Josephine M Bryant, Jackie Chan, Dean Creer, Jolyon Holdstock, et al. Direct whole-genome sequencing of sputum accurately identifies drug-resistant Mycobacterium tuberculosis faster than MGIT culture sequencing. *Journal of Clinical Microbiology*, 56(8):e00666–18, 2018.

13. John W Drake. Rates of spontaneous mutation among RNA viruses. *Proceedings of the National Academy of Sciences*, 90(9):4171–4175, 1993.

14. Alexei J Drummond, Andrew Rambaut, Beth Shapiro, and Oliver G Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, 2005.

15. Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J Tatem, Guy Baele, Nuno R Faria, Daniel J Park, Jason T Ladner, Armando Arias, Danny Asogun, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, 544(7650):309–315, 2017.

16. Gytis Dudas, Luiz Max Carvalho, Andrew Rambaut, and Trevor Bedford. MERS-CoV spillover at the camel-human interface. *eLife*, 7:e31257, 2018.

17. Mark R Duffy, Tai-Ho Chen, W Thane Hancock, Ann M Powers, Jacob L Kool, Robert S Lanciotti, Moses Pretrick, Maria Marfel, Stacey Holzbauer, Christine Dubray, Laurent Guillaumot, Anne Griggs, Martin Bel, Amy J Lambert, Janeen Laven, Olga Kosoy, Amanda Panella, Brad J Biggerstaff, Marc Fischer, and Edward B Hayes. Zika virus outbreak on Yap Island, Federated States of Micronesia. *New England Journal of Medicine*, 360(24):2536–2543, 2009.

18. Ceiridwen J Edwards, Marc A Suchard, Philippe Lemey, John J Welch, Ian Barnes, Tara L Fulton, Ross Barnett, Tamsin C O'Connell, Peter Coxon, Nigel Monaghan, et al. Ancient hybridization and an Irish origin for the modern polar bear matriline. *Current Biology*, 21(15):1251–1258, 2011.

19. Antoine Enfissi, John Codrington, Jimmy Roosblad, Mirdad Kazanji, and Dominique Rousset. Zika virus genome from the Americas. *Lancet*, 387(10015):227–228, January 2016.

20. Nuno R Faria, Josh Quick, IM Claro, Julien Theze, Jacqueline G de Jesus, Marta Giovanetti, Moritz UG Kraemer, Sarah C Hill, Allison Black, Antonio C da Costa, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*, 546(7658):406–410, 2017.

21. Nuno Rodrigues Faria, Raimunda do Socorro da Silva Azevedo, Moritz U G Kraemer, Renato Souza, Mariana Sequetin Cunha, Sarah C Hill, Julien Thézé, Michael B Bonsall, Thomas A Bowden, Ilona Rissanen, Iray Maria Rocco, Juliana Silva Nogueira, Adriana Yurika Maeda, Fernanda Giseli da Silva Vasami, Fernando Luiz de Lima Macedo, Akemi Suzuki, Sueli Guerreiro Rodrigues, Ana Cecilia Ribeiro Cruz, Bruno Tardeli Nunes, Daniele Barbosa de Almeida Medeiros, Daniela Sueli Guerreiro Rodrigues, Alice Louize Nunes Queiroz, Eliana Vieira Pinto da Silva, Daniele Freitas Henriques, Elisabeth Salbe Travassos da Rosa, Consuelo Silva de Oliveira, Livia Caricio Martins, Helena Baldez Vasconcelos, Livia Medeiros Neves Casseb, Darlene de Brito Simith, Jane P Messina, Leandro Abade, José Lourenço, Luiz Carlos Junior Alcantara, Maricélia Maia de Lima, Marta Giovanetti, Simon I Hay, Rodrigo Santos de Oliveira, Poliana da Silva Lemos, Layanna Freitas de Oliveira, Clayton Pereira Silva de Lima, Sandro Patroca da Silva, Janaina Mota de Vasconcelos, Luciano Franco, Jedson Ferreira Cardoso, João Lídio da Silva Gonçalves Vianez-Júnior, Daiana Mir, Gonzalo Bello, Edson Delatorre, Kamran Khan, Marisa Creatore, Giovanini Evelim Coelho, Wanderson Kleber de Oliveira, Robert Tesh, Oliver G Pybus, Marcio R T

Nunes, and Pedro F C Vasconcelos. Zika virus in the Americas: Early epidemiological and genetic findings. *Science*, 352(6283):345–349, April 2016.

22. Oumar Faye, Ousmane Faye, Diawo Diallo, Mawlouth Diallo, Manfred Weidmann, and Amadou Alpha Sall. Quantitative real-time PCR detection of Zika virus and evaluation with field-caught mosquitoes. *Journal of Virology*, 10:311, October 2013.

23. Marco AR Ferreira and Marc A Suchard. Bayesian analysis of elapsed times in continuous-time Markov chains. *Canadian Journal of Statistics*, 36(3):355–368, 2008.

24. Paul EM Fine. The interval between successive cases of an infectious disease. *American Journal of Epidemiology*, 158(11):1039–1047, 2003.

25. Simon DW Frost and Erik M Volz. Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1548):1879–1890, 2010.

26. Mandev S Gill, Philippe Lemey, Nuno R Faria, Andrew Rambaut, Beth Shapiro, and Marc A Suchard. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724, 2013.

27. Marta Giovanetti, Teresa Milano, Luiz Carlos Alcantara, Laura Carcangiu, Eleonora Cella, Alessia Lai, Alessandra Lo Presti, Stefano Pascarella, Gianguglielmo Zehender, Silvia Angeletti, and Massimo Ciccozzi. Zika virus spreading in South America: Evolutionary analysis of emerging neutralizing resistant Phe279Ser strains. *Asian Pacific Journal of Tropical Medicine*, 9(5):445–452, May 2016.

28. Alyssa Goodman, Alberto Pepe, Alexander W Blocker, Christine L Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, et al. Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, 10(4), 2014.

29. Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332, 2004.

30. Nathan D Grubaugh, Jason T Ladner, Moritz U G Kraemer, Gytis Dudas, Amanda L Tan, Karthik Gangavarapu, Michael R Wiley, Stephen White, Julien Thézé, Diogo M Magnani, Karla Prieto, Daniel Reyes, Andrea M Bingham, Lauren M Paul, Refugio Robles-Sikisaka, Glenn Oliveira, Darryl Pronty, Carolyn M Barcellona, Hayden C Metsky, Mary Lynn Baniecki, Kayla G Barnes, Bridget Chak, Catherine A Freije, Adrianne Gladden-Young, Andreas Gnirke, Cynthia Luo, Bronwyn MacInnis, Christian B Matranga, Daniel J Park, James Qu, Stephen F Schaffner, Christopher Tomkins-Tinch, Kendra L West, Sarah M Winnicki, Shirlee Wohl, Nathan L Yozwiak, Joshua Quick, Joseph R Fauver, Kamran Khan, Shannon E Brent, Robert C Reiner, Jr, Paola N Lichtenberger, Michael J Ricciardi, Varian K Bailey, David I Watkins, Marshall R Cone, Edgar W Kopp, 4th, Kelly N Hogan, Andrew C Cannons, Reynald Jean, Andrew J Monaghan, Robert F Garry, Nicholas J Loman, Nuno R Faria, Mario C Porcelli, Chalmers Vasquez, Elyse R Nagle, Derek A T Cummings, Danielle Stanek, Andrew Rambaut, Mariano Sanchez-Lockhart, Pardis C Sabeti, Leah D Gillis, Scott F Michael, Trevor Bedford, Oliver G Pybus, Sharon Isern, Gustavo Palacios, and Kristian G Andersen. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature*, 546(7658):401–405, June 2017.

31. Nathan D Grubaugh, Mary E Petrone, and Edward C Holmes. We shouldn't worry when a virus mutates during disease outbreaks. *Nature Microbiology*, pages 1–2, 2020.

32. Mathilde Guerbois, Ildefonso Fernandez-Salas, Sasha R Azar, Rogelio Danis-Lozano, Celia M Alpuche-Aranda, Grace Leal, Iliana R Garcia-Malo, Esteban E Diaz-Gonzalez, Mauricio Casas-Martinez, Shannan L Rossi, Samanta L Del Río-Galván, Rosa M Sanchez-Casas, Christopher M Roundy, Thomas G Wood, Steven G Widen, Nikos Vasilakis, and Scott C Weaver. Outbreak of Zika virus infection, Chiapas state, Mex-

ico, 2015, and first confirmed transmission by Aedes aegypti mosquitoes in the Americas. *Journal of Infectious Diseases*, 214(9):1349–1356, November 2016.

33. James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

34. Edmund M Hart, Pauline Barmby, David LeBauer, François Michonneau, Sarah Mount, Patrick Mulrooney, Timothée Poisot, Kara H Woo, Naupaka B Zimmerman, and Jeffrey W Hollister. Ten simple rules for digital data storage. *PLoS Computational Biology*, 12(10), 2016.

35. Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

36. Edward C Holmes, Gytis Dudas, Andrew Rambaut, and Kristian G Andersen. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. *Nature*, 538(7624):193–200, 2016.

37. Instituto Nacional de Salud. Boletin epidemiologico semanal - semana epidemiologica 08 de 2016. `https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2016%20Bolet%C3%ADn%20epidemiol%C3%B3gico%20semana%208.pdf`. Accessed: 2016-3-11.

38. Kazutaka Katoh, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, Jul 2002.

39. John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

40. Robert S Lanciotti, Amy J Lambert, Mark Holodniy, Sonia Saavedra, and Leticia Del Carmen Castillo Signor. Phylogeny of Zika virus in the Western Hemisphere, 2015. *Emerging Infectious Diseases*, 22(5):933–935, May 2016.

41. John Lednicky, Valery Madsen Beau De Rochars, Maha El Badry, Julia Loeb, Taina Telisma, Sonese Chavannes, Gina Anilis, Eleonora Cella, Massimo Ciccozzi, Mohammed Rashid, Bernard Okech, Marco Salemi, and J Glenn Morris, Jr. Zika virus outbreak in Haiti in 2014: Molecular and clinical data. *PLoS Neglected Tropical Diseases.*, 10(4):e0004687, April 2016.

42. Philippe Lemey, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A Russell, Derek J Smith, Oliver G Pybus, Dirk Brockmann, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathogens*, 10(2), 2014.

43. Philippe Lemey, Andrew Rambaut, Alexei J Drummond, and Marc A Suchard. Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9):e1000520, 2009.

44. Philippe Lemey, Andrew Rambaut, John J Welch, and Marc A Suchard. Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27(8):1877–1885, 2010.

45. Justin Lessler, Lelia H Chaisson, Lauren M Kucirka, Qifang Bi, Kyra Grantz, Henrik Salje, Andrea C Carcelen, Cassandra T Ott, Jeanne S Sheffield, Neil M Ferguson, Derek A T Cummings, C Jessica E Metcalf, and Isabel Rodriguez-Barraquer. Assessing the global threat from Zika virus. *Science*, 353(6300):aaf8160, August 2016.

46. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

47. José M Malpica, Aurora Fraile, Ignacio Moreno, Clara I Obies, John W Drake, and Fernando García-Arenal. The rate and character of spontaneous mutation in an RNA virus. *Genetics*, 162(4):1505–1511, 2002.

48. Hayden C Metsky, Christian B Matranga, Shirlee Wohl, Stephen F Schaffner, Catherine A Freije, Sarah M Winnicki, Kendra West, James Qu, Mary Lynn Baniecki, Adrianne Gladden-Young, Aaron E Lin, Christopher H Tomkins-Tinch, Simon H Ye, Daniel J Park, Cynthia Y Luo, Kayla G Barnes, Rickey R Shah, Bridget Chak, Giselle Barbosa-Lima, Edson Delatorre, Yasmine R Vieira, Lauren M Paul, Amanda L Tan, Carolyn M Barcellona, Mario C Porcelli, Chalmers Vasquez, Andrew C Cannons, Marshall R Cone, Kelly N Hogan, Edgar W Kopp, Joshua J Anzinger, Kimberly F Garcia, Leda A Parham, Rosa M Gélvez Ramírez, Maria C Miranda Montoya, Diana P Rojas, Catherine M Brown, Scott Hennigan, Brandon Sabina, Sarah Scotland, Karthik Gangavarapu, Nathan D Grubaugh, Glenn Oliveira, Refugio Robles-Sikisaka, Andrew Rambaut, Lee Gehrke, Sandra Smole, M Elizabeth Halloran, Luis Villar, Salim Mattar, Ivette Lorenzana, Jose Cerbino-Neto, Clarissa Valim, Wim Degrave, Patricia T Bozza, Andreas Gnirke, Kristian G Andersen, Sharon Isern, Scott F Michael, Fernando A Bozza, Thiago M L Souza, Irene Bosch, Nathan L Yozwiak, Bronwyn L MacInnis, and Pardis C Sabeti. Zika virus evolution and spread in the Americas. *Nature*, 546(7658):411–415, June 2017.

49. Cynthia A Moore, J Erin Staples, William B Dobyns, André Pessoa, Camila V Ventura, Eduardo Borges da Fonseca, Erlane Marques Ribeiro, Liana O Ventura, Norberto Nogueira Neto, J Fernando Arena, and Sonja A Rasmussen. Characterizing the pattern of anomalies in Congenital Zika Syndrome for pediatric clinicians. *JAMA Pediatrics*, 171(3):288–295, March 2017.

50. Samia N Naccache, Julien Thézé, Silvia I Sardi, Sneha Somasekar, Alexander L Greninger, Antonio C Bandeira, Gubio S Campos, Laura B Tauro, Nuno R Faria,

Oliver G Pybus, and Charles Y Chiu. Distinct Zika virus lineage in Salvador, Bahia, Brazil. *Emerging Infectious Diseases*, 22(10):1788–1792, 2016.

51. Eduardo Martins Netto, Andres Moreira-Soto, Celia Pedroso, Christoph Höser, Sebastian Funk, Adam J Kucharski, Alexandra Rockstroh, Beate M Kümmerer, Gilmara Souza Sampaio, Estela Luz, Sara Nunes Vaz, Juarez Pereira Dias, Fernanda Anjos Bastos, Renata Cabral, Thomas Kistemann, Sebastian Ulbert, Xavier de Lamballerie, Thomas Jaenisch, Oliver J Brady, Christian Drosten, Manoel Sarno, Carlos Brites, and Jan Felix Drexler. High Zika virus seroprevalence in Salvador, northeastern Brazil, limits the potential for further outbreaks. *MBio*, 8(6), November 2017.

52. Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, January 2015.

53. Magnus Nordborg. Coalescent theory. *Handbook of Statistical Genetics*, 2:843–877, 2001.

54. Oscar Pacheco, Mauricio Beltrán, Christina A Nelson, Diana Valencia, Natalia Tolosa, Sherry L Farr, Ana V Padilla, Van T Tong, Esther L Cuevas, Andrés Espinosa-Bode, Lissethe Pardo, Angélica Rico, Jennita Reefhuis, Maritza González, Marcela Mercado, Pablo Chaparro, Mancel Martínez Duran, Carol Y Rao, María M Muñoz, Ann M Powers, Claudia Cuéllar, Rita Helfand, Claudia Huguett, Denise J Jamieson, Margaret A Honein, and Martha L Ospina Martínez. Zika Virus Disease in Colombia - preliminary report. *New England Journal of Medicine*, June 2016.

55. PAHO. Zika cumulative cases. `https://www.paho.org/hq/index.php?option=com_content&view=article&id=12390:zika-cumulative-cases&Itemid=42090&lang=en`. Accessed: 2018-9-10.

56. PAHO. Epidemiological alert: Zika virus infection. Technical report, PAHO, May 2015.

57. Beatriz Parra, Jairo Lizarazo, Jorge A Jiménez-Arango, Andrés F Zea-Vera, Guillermo González-Manrique, José Vargas, Jorge A Angarita, Gonzalo Zuñiga, Reydmar Lopez-Gonzalez, Cindy L Beltran, Karen H Rizcala, Maria T Morales, Oscar Pacheco, Martha L Ospina, Anupama Kumar, David R Cornblath, Laura S Muñoz, Lyda Osorio, Paula Barreras, and Carlos A Pardo. Guillain-Barré syndrome associated with Zika virus infection in Colombia. *New England Journal of Medicine*, 375(16):1513–1523, October 2016.

58. Rodrigo Pessôa, João Veras Patriota, Maria de Lourdes de Souza, Alvina Clara Felix, Nubia Mamede, and Sabri S Sanabani. Investigation into an outbreak of dengue-like illness in Pernambuco, Brazil, revealed a cocirculation of Zika, Chikungunya, and Dengue virus Type 1. *Medicine*, 95(12):e3201, March 2016.

59. John H-O Pettersson, John H O. Pettersson, Vegard Eldholm, Stephen J Seligman, Åke Lundkvist, Andrew K Falconar, Michael W Gaunt, Didier Musso, Antoine Nougairède, Remi Charrel, Ernest A Gould, and Xavier de Lamballerie. How did Zika virus emerge in the Pacific Islands and Latin America? *MBio*, 7(5), 2016.

60. Brett E Pickett, Eva L Sadat, Yun Zhang, Jyothi M Noronha, R Burke Squires, Victoria Hunt, Mengya Liu, Sanjeev Kumar, Sam Zaremba, Zhiping Gu, Liwei Zhou, Christopher N Larson, Jonathan Dietrich, Edward B Klem, and Richard H Scheuermann. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, 40(D1):D593–D598, 2011.

61. Miquel Porta. *A Dictionary of Epidemiology*. Oxford University Press, 2014.

62. Joshua Quick, Nathan D Grubaugh, Steven T Pullan, Ingra M Claro, Andrew D Smith, Karthik Gangavarapu, Glenn Oliveira, Refugio Robles-Sikisaka, Thomas F Rogers,

Nathan A Beutler, Dennis R Burton, Lia Laura Lewis-Ximenez, Jaqueline Goes de Jesus, Marta Giovanetti, Sarah C Hill, Allison Black, Trevor Bedford, Miles W Carroll, Marcio Nunes, Luiz Carlos Alcantara, Jr, Ester C Sabino, Sally A Baylis, Nuno R Faria, Matthew Loose, Jared T Simpson, Oliver G Pybus, Kristian G Andersen, and Nicholas J Loman. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols*, 12(6):1261–1276, June 2017.

63. Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, 2016.

64. Sonja A Rasmussen and Richard A Goodman. *The CDC Field Epidemiology Manual*. Oxford University Press, 2018.

65. Diana Patricia Rojas, Natalie E Dean, Yang Yang, Eben Kenah, Juliana Quintero, Simon Tomasi, Erika Lorena Ramirez, Yendi Kelly, Carolina Castro, Gabriel Carrasquilla, M Elizabeth Halloran, and Ira M Longini. The epidemiology and transmissibility of Zika virus in Girardot and San Andres island, Colombia, September 2015 to January 2016. *Eurosurveillance*, 21(28), July 2016.

66. Pavel Sagulenko, Vadim Puller, and Richard A Neher. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1):vex042, January 2018.

67. Gilberto A Santiago, Jesús Vázquez, Sean Courtney, Katia Y Matías, Lauren E Andersen, Candimar Colón, Angela E Butler, Rebecca Roulo, John Bowzard, Julie M Villanueva, and Jorge L Muñoz-Jordan. Performance of the Trioplex real-time RT-PCR assay for detection of Zika, dengue, and chikungunya viruses. *Nature Communications*, 9(1), 2018.

68. Marc A Suchard, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1), 06 2018. vey016.

69. Julien Thézé, Tony Li, Louis du Plessis, Jerome Bouquet, Moritz U G Kraemer, Sneha Somasekar, Guixia Yu, Mariateresa de Cesare, Angel Balmaseda, Guillermina Kuan, Eva Harris, Chieh-Hsi Wu, M Azim Ansari, Rory Bowden, Nuno R Faria, Shigeo Yagi, Sharon Messenger, Trevor Brooks, Mars Stone, Evan M Bloch, Michael Busch, José E Muñoz-Medina, Cesar R González-Bonilla, Steven Wolinsky, Susana López, Carlos F Arias, David Bonsall, Charles Y Chiu, and Oliver G Pybus. Genomic epidemiology reconstructs the introduction and spread of Zika virus in Central America and Mexico. *Cell Host and Microbe*, 23(6):855–864.e7, June 2018.

70. Ruth E Timme, Hugh Rand, Martin Shumway, Eija K Trees, Mustafa Simmons, Richa Agarwala, Steven Davis, Glenn E Tillman, Stephanie Defibaugh-Chavez, Heather A Carleton, et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ*, 5:e3893, 2017.

71. Richard A Urbanowicz, C Patrick McClure, Anavaj Sakuntabhai, Amadou A Sall, Gary Kobinger, Marcel A Müller, Edward C Holmes, Félix A Rey, Etienne Simon-Loriere, and Jonathan K Ball. Human adaptation of Ebola virus during the West African outbreak. *Cell*, 167(4):1079–1087, 2016.

72. Erik M Volz, Sergei L Kosakovsky Pond, Melissa J Ward, Andrew J Leigh Brown, and Simon DW Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430, 2009.

73. Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.

74. Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K Teal. Good enough practices in scientific computing. *PLoS Computational Biology*, 13(6), 2017.

75. Katherine S Xue, Louise H Moncla, Trevor Bedford, and Jesse D Bloom. Within-host evolution of human influenza virus. *Trends in Microbiology*, 26(9):781–793, 2018.

76. José Victor Zambrana, Fausto Bustos Carrillo, Raquel Burger-Calderon, Damaris Collado, Nery Sanchez, Sergio Ojeda, Jairo Carey Monterrey, Miguel Plazaola, Brenda Lopez, Sonia Arguello, Douglas Elizondo, William Aviles, Josefina Coloma, Guillermina Kuan, Angel Balmaseda, Aubree Gordon, and Eva Harris. Seroprevalence, risk factor, and spatial analyses of Zika virus infection after the 2016 epidemic in Managua, Nicaragua. *Proceedings of the National Academy of Sciences*, 115(37):9294–9299, September 2018.

77. Zhao Zhang, Libing Shen, and Xun Gu. Evolutionary dynamics of MERS-CoV: potential recombination, positive selection and transmission. *Scientific Reports*, 6:25049, 2016.

Appendix A

ADDITIONAL RESEARCH

Here I provide a brief description of additional research projects outside of my dissertation work that I have either co-led or had significant involvement in.

Understanding Mumps transmission in Washington state, 2016-2017

In 2016 and 2017, Washington state had a significant outbreak of Mumps virus. In collaboration with the Washington State Department of Health, Dr. Louise Moncla, a postdoctoral fellow in the Bedford Lab, and I co-led a project to sequence Mumps viral genomes and perform a genomic epidemiologic analysis of the outbreak. We are currently preparing a manuscript on which I am co-first author.

Using genomic epidemiology to support Ebola virus disease surveillance in the Democratic Republic of the Congo

During the Ebola virus disease outbreak occurring in Nord-Kivu province, Democratic Republic of the Congo, I travelled to Kinshasa to help train scientists at the Institut National de Recherche Biomédicale how to perform genomic epidemiological analysis of the Ebola virus genomes they had been sequencing. Since then, I have continued to support the effort remotely, developing situation reports describing genomic epidemiological findings that are written for, and distributed to, frontline public health workers. We are preparing a manuscript for publication that describes this genomic surveillance effort, on which I am a co-first author.

Investigating early dynamics of Zika in Brazil

As a part of a collaborative effort between scientists in Brazil, the UK, and the United States, I travelled to Brazil to sequence Zika virus genomes from diagnostic specimens collected in Northeastern Brazil, where Zika first circulated. Analysis of these data greatly refined our understanding of when Zika was introduced to Brazil, and how it spread throughout the country and into other parts of the Americas. This work led to a publication in Nature, on which I am a co-first author.

Phylogenetic analysis of an iatrogenic outbreak of HIV in Cambodia

In collaboration with researchers at the Institut Pasteur of Cambodia, I analyzed HIV genomes sequenced during an outbreak that occurred in Roka, Cambodia. Although the outbreak was primarily associated with needle-reuse by an unlicensed medical practitioner, the genomic analysis helped differentiate outbreak-associated cases from unrelated prevalent cases that were detected during enhanced surveillance in the area. I am a co-author on the paper, which was published in Clinical Infectious Diseases.