Using Intrahost Genetic Diversity to Understand RNA Virus Evolution and Transmission

by

Andrew L. Valesano

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Cellular and Molecular Biology)
2021

Doctoral Committee:

Associate Professor Adam Lauring, Chair
Professor Kathleen Collins
Associate Professor Andrew Tai
Associate Professor Christiane Wobus
Assistant Professor Robert Woods

Andrew Luke Valesano

avalesan@med.umich.edu

ORCID: 0000-0002-3649-8352

© Andrew Luke Valesano 2021

To M.

ACKNOWLEDGEMENTS

Adam Lauring, I could not have asked for a better experience in graduate school. Thank you for your advice, guidance, encouragement, and insight. You have been a tremendous example as a mentor and physician-scientist.

Will Fitzsimmons, M.S., thank you for your support and listening ear. Thank you for helping me get on my feet, get up, and make things happen.

Thank you, Kayla Peck, Danny Lyons, and David Jimenez-Vallejo for fun discussions of virus evolution. Thank you to JT McCrone for your example and helping me to grow my skills and confidence with computational methods. To Yuan Li, Emily Bendall, Sarah Arcos, Kalee Rumfelt, Derek Dimcheff, and all members of the Lauring Lab, old and new: thank you for your encouragement and uplifting spirits.

Thank you to Mike Famulare and Wes Wong for broadening my views of the interactions between virus evolution and epidemiology. Special thanks to Mami Taniuchi for introducing me to international infectious diseases research and many wonderful colleagues across the globe.

Thank you to Christiane Wobus and Bob Woods for sharing your expertise on viral pathogenesis and evolution. Kathy Collins and Andrew Tai, thank you for your examples and perspectives on developing as a physician-scientist.

I would not have made it to graduate school without the guidance of Aaron Best and Tahnee Prokopow. I am endlessly grateful for their encouragement and mentorship.

Thank you to everyone at the University of Michigan who has been so welcoming and made these past years a wonderful experience, especially the Medical Scientist Training Program, the Department of Microbiology and Immunology, and the Program in Cellular and Molecular Biology.

Most of all, thank you Mary Elizabeth for your unending love and support through this long program. Thank you for making every day a joy. Lastly, thank you, Madge, for everything.

This work was funded by NIH-T32-GM007863, the Bill and Melinda Gates Foundation, and grants to Adam Lauring from the NIH and the Burroughs Wellcome Fund.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

RNA viruses generate genetically diverse populations during acute infections within human hosts. Studying viral population dynamics in natural infections is important for understanding how new viral variants arise and spread. However, these dynamics are poorly understood in most viruses. In my thesis, I have comprehensively analyzed viral within-host evolution across three unique taxa.

In the first study, I defined the within-host diversity of influenza B viruses in a household cohort in southeastern Michigan. Using an experimentally validated next-generation sequencing approach, I found that influenza B viruses accumulate less genetic diversity compared to influenza A viruses. Similar to influenza A, I found that influenza B virus faces a stringent transmission bottleneck. These results suggest a complex relationship between viral mutation rates, intrahost diversity, and global rates of viral evolution.

In the second study, I investigated the early evolution of the live-attenuated oral polio vaccine (OPV) by sequencing samples from vaccine recipients and their close contacts in a field trial of polio vaccines. In contrast to endemic viruses, I found that OPV exhibits a significant amount of parallel evolution within primary vaccine recipients. I identified 19 sites under positive selection, most of which were previously thought to evolve neutrally. Between hosts, a tight transmission bottleneck limited the spread of adaptive mutations. These results demonstrate the distinct within-host dynamics of live-attenuated vaccines, highlight the role of transmission bottlenecks in constraining virus evolution, and offer valuable information for interpreting genetic surveillance data in the ongoing effort to eradicate polio.

In the third study, I defined the within-host variation of SARS-CoV-2 in hospitalized patients and infected healthcare workers during the first months of the pandemic. I sequenced known virus mixtures to show how viral load impacts the accuracy of variant identification. In contrast to several early reports, I found that intrahost diversity is low over the course of infections. I demonstrated that variants arise in parallel across individuals from separate transmission networks, which complicates the use of intrahost variants in transmission inference. These findings clarify conflicting results from previous work on SARS-CoV-2 variant generation and spread.

More broadly, these studies demonstrate the challenges of accurately detecting intrahost viral variants, illustrate how within-host studies can resolve ambiguities observed in global evolutionary dynamics, and provide important context for using intrahost variation in sequence-based transmission inference.

CHAPTER I

Introduction

Overview

The rapid evolution of RNA viruses has created many significant public health challenges. Influenza viruses regularly undergo antigenic change and escape from population immunity (Petrova and Russell, 2018); the diversification of human immunodeficiency virus, hepatitis C virus, and enteroviruses impedes the design of effective vaccines (Hedestam et al., 2008; Palmenberg et al., 2009); genetic reversion of the live polio vaccine thwarts global eradication (Kew and Pallansch, 2018); SARS-CoV-2 variants with greater transmissibility are actively emerging (Lauring and Hodcroft, 2021). Although the phenotypic consequences of virus evolution are often system-specific, the underlying evolutionary forces of mutation, selection, and drift are shared by all viruses. Understanding the fundamental pressures and constraints that RNA viruses face as they evolve within their natural hosts is critical for rational disease management, infection prevention, and public health measures.

It has long been recognized that RNA viruses exhibit high levels of genetic diversity as they replicate within infected hosts (Moya et al., 2004). Due to the lack of proofreading and repair activities of most RNA-dependent RNA polymerases, RNA viruses possess mutation rates that are orders of magnitude higher than other organisms. Rapid supply of *de novo* mutations during each round of replication results in the establishment of highly diverse viral populations. The earliest work in this field employed RNA bacteriophages to uncover fundamental principles of RNA virus mutation rates and population dynamics. More recently, "next-generation" sequencing (NGS) technologies have allowed systematic analysis of viral populations in human hosts (Lauring, 2020). Specifically, the high accuracy and throughput of Illumina sequencing has enabled the detection of rare within-host variants in primary human-derived samples. The bulk of this work has focused on human immunodeficiency virus (HIV) and influenza A virus (IAV), and there has also been substantial work on hepatitis C virus (HCV) and arboviruses such as dengue virus. This growing body of work has characterized the within-host evolution of multiple viruses and yielded important insights into viral evolutionary dynamics. However, there are important gaps that remain in understanding how viral intrahost diversity leads to evolutionary change on a global scale.

A key objective of studies on viral intrahost evolution is to understand the evolutionary pressures that act on viruses as they replicate in humans. Although mutations are the raw substrate for evolutionary change, the fate of any given mutation or genomic variant depends on the combined action of multiple evolutionary forces within infected hosts (Dolan et al., 2016). The opposing forces of natural selection and genetic drift are particularly important for understanding virus evolution in humans. Selection is a deterministic force that dominates when population sizes are large. In general, natural selection will increase the frequency of a variant that has a fitness benefit (positive selection) and decrease the frequency of a variant with a fitness defect (negative selection). The viral factors that determine fitness are multifactorial, but they often can be distilled to functions that enable better replication within a host and evasion of immunity, greater shedding from infected hosts, and greater transmission to new hosts. Mutation fitness effects are context dependent, such that a mutation may increase one aspect of viral fitness but decrease another. Opposite to selection is genetic drift, a stochastic force that causes random fluctuations in mutation frequency when population sizes are small. Decreases in population size, such as bottlenecks during transmission, can have drastic effects on the composition of viral populations and their subsequent evolution (McCrone and Lauring, 2018; Zwart and Elena, 2015). Detecting, quantifying, and comparing these forces across multiple relevant contexts is central to understanding how viruses evolve. However, it has been challenging to draw generalizable conclusions about virus evolution from many previous studies of intrahost diversity. Studies are often based on haphazardly collected specimens, such as convenience samples from a clinical laboratory. There can be systematic errors in identifying intrahost variants that obscure true biological diversity. These limitations hinder our ability to interpret what the results mean for understanding virus evolution and spread.

Herein, I will review the important factors in measuring virus intrahost populations by NGS. I will outline how these data can provide insight into virus evolution and transmission at the level of individual hosts. Lastly, I will introduce current challenges in genomic epidemiology and the opportunities that intrahost diversity offer for enhancing transmission inference.

Accurately quantifying virus intrahost diversity

Advent of high-throughput "next-generation" DNA sequencing (NGS) has provided the means to study viral intrahost diversity at whole genome scales. Most studies enrich viral genomes from clinical samples by RT-PCR, PCR, or hybridization-based methods. Enrichment of viral genomes allows the direct quantification of intrahost populations without intermediate passaging in cell culture, which can alter the composition of the population and introduce cell culture adaptations (McWhite et al., 2016). Strategies for RT-PCR amplification are specific to the genome structure. Segmented viruses like influenza virus are often amplified with primers targeted to the ends of genome segments. Non-segmented viruses are more difficult to amplify in a single amplicon; most studies use tiled amplicons in a sliding window across the viral genome. After enrichment, conventional NGS is then used to sequence the enriched DNA, map to virus reference genomes, and identify genetic variants that differ from the consensus sequence of the population. NGS reads are short, commonly 150-500 bases, and therefore are unable to comprehensively identify haplotype linkage. For RNA viruses, intrahost genetic variants are most often single nucleotide variants (iSNV), but intrahost insertions and deletions (indels) can also occur. Indel variants are more difficult to reliably detect and have been studied less than iSNV. Massively parallel sequencing methods can generate high read depth over a given loci, which allows for more precise estimates of a variant's frequency. Precision is a function of the read depth achieved. At a depth of 100x, precision cannot exceed 1% frequency by definition. A common practice is to only include variants that are supported by 10 or more reads. Therefore, read depth of 1000x is required to reliably detect variants present at 1% frequency. Greater read depth can identify rarer variants, provided that those variants were not lost during enrichment, library preparation, and sequencing.

Measuring intrahost diversity within clinical specimens faces several challenges due to multiple sources of error (Grubaugh et al., 2019a; McCrone and Lauring, 2016). These sources of error are not always accounted for in the published literature on viral intrahost diversity. Reverse transcription, PCR, and next-generation sequencing can introduce errors. Sequencing platforms differ in their base error rates; for example, nanopore-based sequencing has a higher base error rate compared to current sequencing-by-synthesis methods. Cross-contamination and mixed infections can also create the appearance of intrahost diversity (Cudini et al., 2019). Due to the errors introduced and propagated through enrichment, library preparation, and sequencing, there is often a predominance of errors at 0.5% frequency and below, regardless of read depth.

There are various genome amplification methods that can reduce error, but most are not compatible with whole genome sequencing from clinical specimens. Amplification with primers that have unique molecular identifiers can increase accuracy down to about 0.1%, but this is not viable for whole genomes (Jabara et al., 2011). New sophisticated methods of circular sequencing by rolling circle reverse transcription are highly accurate but require high RNA input that is not compatible with most clinical specimens (Acevedo et al., 2014). A simpler approach to reducing false positive variant calls is to perform technical replicates of a given specimen. True population variants should be identified in both technical replicates while spurious false positives are unlikely to randomly arise in all replicates. Previous experiments have shown that sequencing in duplicate can dramatically reduce false positives; however, technical replicates alone do not guarantee that false positives will not occur (Grubaugh et al., 2019a; McCrone and Lauring, 2016).

Specimen viral load has emerged as a key factor in the success of sample sequencing and the accuracy of variant calling. With low input genome copies, more effective PCR cycles occur which can propagate errors in the reverse transcription, resulting in more false positives. Low viral loads can also cause false negative errors for variants at low frequencies due to stochastic loss of variants during enrichment, library preparation, and sequencing. These errors are exacerbated in clinical specimens, which often harbor fewer viral genome copies compared to cell culture supernatants. With low input viral loads, a significant number of false positive variants can occur even in samples sequenced in duplicate.

After sequencing, there are many computational strategies for reducing error in variant identification. Multiple software packages for calling variants are available. There is no current consensus in the field on which variant caller is most accurate, and any given tool can have idiosyncratic error profiles. Studies often differ in the criteria used to identify within-host variants, which complicates direct comparisons of viral diversity across studies. Most approaches apply a common set of filter criteria, such as total and variant read depth, mapping quality, base quality (Phred) score, strand bias, and frequency. Some variant callers use more sophisticated statistical models. For example, deepSNV uses sequencing of a clonal plasmid control to model the local error rates across reads (Gerstung et al., 2012). LoFreq uses base quality scores and a Poisson-binomial model to test whether sequencing error alone could produce a given iSNV (Wilm et al., 2012).

Empirical benchmarking experiments are critical for establishing the accuracy of a given sequencing workflow. Sequencing of mock communities, where the true variant frequencies are known, has been widely used in bacterial 16S ribosomal RNA gene sequencing and metagenomic studies (Bowers et al., 2015). However, this approach has only recently been applied to virus intrahost sequencing. A few previous studies have benchmarked the sensitivity and specificity of different variant callers with experimental data, including influenza A virus, West Nile virus, and Zika virus (Grubaugh et al., 2019a; McCrone and Lauring, 2016). Experimental validation of variant identification by sequencing mock populations that closely replicate the diversity and viral load of corresponding clinical specimens is now the field standard. However, there is no "one size fits all" for the overall sequencing and analysis approach. Specific viral RNA enrichment strategies, variant callers, and iSNV filter criteria may necessarily differ based on the viral system and the analysis goals.

In my thesis, I have grounded each analysis of viral intrahost diversity in empirically validated methods for detecting intrahost variants. In Chapter II, I relied on previous studies in the lab on influenza A virus to inform variant calling for influenza B virus. In Chapters III and IV, I performed additional experiments to benchmark and validate the accuracy of our variant calling approaches for poliovirus and SARS-CoV-2. These efforts were crucial for robust biological interpretation of the results.

Evolutionary inference from within-host data

Accurate characterization of viral populations is merely the starting point for understanding their evolutionary dynamics in humans. A major challenge of these studies is making appropriate inferences about evolutionary forces on viruses and making reasonable conclusions about the underlying biological significance. Common mistakes in this area are haphazardly collecting samples with limited clinical metadata, improperly controlling for sequence errors, and making outsized conclusions about identifying exceedingly rare variants.

Obtaining samples from a well-defined set of patient samples with relevant metadata is crucial for the downstream analysis and biological inference. To observe changes in variant frequency over time or shared variation between transmission networks, it is necessary to obtain longitudinal samples from a large cohort of individuals. Clinical trials, observational studies, and field studies of virus infection and epidemiology are rarely designed with analysis of intrahost variation in mind. Therefore, access to samples is usually post-hoc, which fundamentally limits the analysis. However, there are notable exceptions in the literature. Household cohorts, which have traditionally been used to measure vaccine effectiveness, have been recently used to study viral transmission dynamics and estimate transmission bottlenecks (McCrone et al., 2018). Infection challenge studies, while not necessarily representative of natural infections, have allowed frequent sampling from the same individuals (Leonard et al., 2016). A recent placebo-controlled clinical trial of influenza vaccines allowed researchers to isolate the effects of vaccination on IAV intrahost diversity (Debbink et al., 2017). Although these studies are difficult to conduct, they are the most powerful for capturing various forms of natural selection and genetic drift.

Even with access to a rich sample set, it can be difficult to detect positive and negative selection within hosts. There are many approaches to detecting natural selection in sequence data, but few are suited to the within-host level. The frequency distribution of mutations usually holds indirect evidence of purifying selection. In nearly all studies of intrahost diversity, there are more variants at lower frequency levels due to negative selection, and fewer mutations survive to reach intermediate frequencies. However, this “snapshot” view offers little for biologically useful insights.

A common method for identifying selection is calculating the dN/dS ratio, which relates the number of nonsynonymous mutations to the number of synonymous mutations, corrected for the respective number of sites (Kryazhimskiy and Plotkin, 2008). While synonymous mutations can have significant fitness effects, non-synonymous mutations are often deleterious and removed by negative selection. Across a gene or at a given codon, a dN/dS ratio of > 1 suggests positive selection, while dN/dS ratio of < 1 suggests negative selection. However, dN/dS and related tests are not well suited for intrahost datasets. This test cannot assess mutations in noncoding regions, which can be sites of significant selection in viruses that depend on RNA secondary structures or other elements for critical functions. On the short timescales of single infections, dN/dS ratios can be artificially elevated because negative selection has not had sufficient time to act on deleterious mutations. Conversely, dN/dS ratios may not be sensitive enough to detect recent positive selection at a locus or sites under weakly positive selection.

A complementary strategy for detecting positive selection is to look for a preponderance of mutations in a given genome region or at a specific locus (Gutierrez et al., 2019). The general idea is that if a mutation occurs more times than expected by chance alone, then that suggests the mutation is favored by natural selection. This phenomenon is often referred to as convergent evolution or parallel evolution. Strictly speaking, those terms have distinct but similar definitions. Parallel mutation in consensus-level phylogenetic analyses may suggest sites where there is selection for the same phenotype, when different evolutionary lineages "converge" on the same genetic solution. This approach has been used to identify positive selection at the consensus level in influenza viruses, vaccine-derived polioviruses, and has been central to current investigations of SARS-CoV-2 variants of concern (VOC) (Escalera-Zamudio et al., 2020; Hodcroft et al., 2021; Stern et al., 2017). Detecting the same intrahost variant in multiple infections is suggestive of positive selection. Unlike dN/dS, this approach does not have standard statistical framework. The simplest method is to detect intrahost mutations that have known fitness effects, established in previous studies or experimental work. Similarly, enrichment of mutations in genome regions with known functions can suggest positive selection. For example, enrichment of non-synonymous mutations in antigenic regions of the influenza hemagglutinin protein might suggest selection for better viral entry or escape from antibody neutralization. There could also be enrichment at the level of a gene that has multiple sites experiencing mild positive selection. A more direct assessment of positive selection might include identifying mutations that occur independently across a cohort of individuals not related by transmission. This is powerful evidence supporting positive selection of a mutation, but it can be difficult to use this method for quantifying the degree of fitness advantage that a mutation confers. Lastly, appearance of an intrahost variant multiple times in samples that temporally precede an increase in the consensus frequency of that variant may suggest the action of positive selection. However, it is often difficult to separate this from the influence of genetic drift at the within-host or population level.

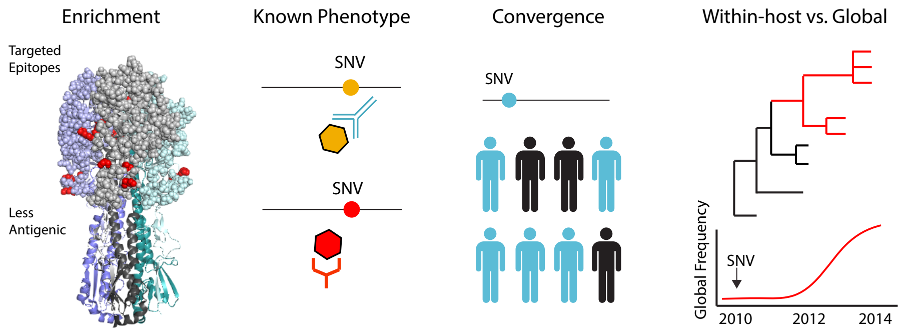


Figure 1.1. Strategies for detecting positive selection with within-host sequencing data. Figure from Lauring 2020, Annual Reviews in Microbiology. Within-host data may reveal enrichment of mutations on key epitopes of viral proteins, such as influenza hemagglutinin shown here. Some mutations may have known phenotypes from laboratory experiments, such as changes in antigenicity or receptor binding. Convergence is the appearance of the same mutation independently across individuals. Comparing the dynamics of mutations from the within-host to the global scale can suggest evidence of positive selection.

A common conundrum when looking for enriched or parallel mutations is distinguishing what is occurring by chance and what is influenced by natural selection. How many times a mutation could occur by chance in a given intrahost dataset is highly dependent on the context of the study and the viral system. Recent work on influenza A in immunocompromised individuals and avian influenza viruses have used permutation tests to assess whether shared mutations in hemagglutinin occurred more often than by chance alone (Moncla et al., 2020; Xue et al., 2017). However, these statistical tests are not without their limits and may not be generalizable to all study designs.

Another challenge in interpreting intrahost datasets is evaluating the impact of genetic drift, which is commonly underappreciated relative to natural selection. Drift can manifest in two important ways in individual infections: stochastic fluctuations in variant frequency during replication within a host or spread to different anatomical compartments, and reduction in population size during transmission.

Within hosts, genetic drift influences variant frequency simply due to random sampling (Dolan et al., 2016). Therefore, a rise in frequency of a specific variant is not necessarily evidence of natural selection. This is an important force to account for in inferences of selection. Genetic drift is amplified when population sizes are small. This is usually much smaller than the number of individuals in a population, or for viruses, the number of physical or infectious particles within an individual host. Genetic drift is thought to play a major role in within-host evolution for many viruses (Lequime et al., 2016; McCrone et al., 2018; Nelson et al., 2006). However, the effective population size of RNA viruses in acute infections is poorly understood. This remains an important area for future innovation.

Transmission bottlenecks can have profound impacts on the rates of virus evolution and the contribution of variants generated within hosts to global evolutionary dynamics (Xue et al., 2018). Bottlenecks decrease the genetic diversity, and thus the effective population size, of a population. This decrease in the effective population size increases the influence of stochastic effects on variant frequency. This can have different effects on viral fitness, depending on the context. Transmission bottlenecks may cause stochastic loss of beneficial mutations, while fixing neutral or even deleterious ones (McCrone and Lauring, 2018). Bottlenecks may also benefit viral populations by purging defective viral genomes or other deleterious elements. These bottlenecks may undo the momentary gains of intrahost mutations that were generated during a given infection by limiting their spread to new hosts. Consequently, if mutations do not transmit to new hosts, they will not have the opportunity to increase and fix at a global population level. While the transmission bottleneck is primarily a force that acts between two individual hosts, it can have large impacts on the global rate of viral adaptation by decreasing the likelihood that a mutation arising within a host will spread throughout a population.

The size of the transmission bottleneck for a given virus is an important question. There may not be a single value that describes the bottleneck for all transmission events or all epidemiological contexts, but it is important to understand what happens most often during viral transmission. If a bottleneck is consistently tight, or stringent, this will greatly increase the role of genetic drift at the individual-to-individual level. This will decrease the efficiency of selection in the population at large. However, if a bottleneck is consistently wide and allows transmission of high levels of genetic diversity to a new host, this will permit natural selection to exercise a stronger role. The transmission bottleneck has been measured for very few viruses in humans. HIV and HCV are thought to experience tight bottlenecks during transmission (Kariuki et al., 2017; Wang et al., 2010). The first studies on influenza A virus had conflicting evidence on the bottleneck size, but the most recent and robust studies have demonstrated a narrow bottleneck (McCrone et al., 2018; Poon et al., 2016; Xue and Bloom, 2019a). A similar picture has emerged for SARS-CoV-2, although there are conflicting reports (Lythgoe et al., 2021; Popa et al., 2020). These conflicting estimates highlight the challenges of intrahost variation at all levels: study design, accuracy of variant identification, and interpretation of shared variants. These questions have been at the center of my thesis, relevant to each chapter and viral system we studied.

In Chapters II and III, I used rigorous statistical methods and longitudinal sampling to identify sites under positive selection. I used household-based studies of viral transmission to estimate the transmission bottleneck size for influenza B virus and the oral polio vaccine. I used these estimates to demonstrate the influence of transmission bottlenecks on the spread of positively selected mutations.

Applications to genomic epidemiology

The most common goal of studying viral intrahost variation has been to gain insight into the general evolutionary dynamics of viruses. Recently, however, there has been increased attention on potential applications in genomic epidemiology and sequence-based transmission inference (Villabona-Arenas et al., 2020). Consensus-level genomic epidemiology has been widely used in the past several years for tracking the movements of viral lineages through human populations (Grubaugh et al., 2019b). Due to the intensive resources and analysis required for genome sequencing, this is usually a retrospective exercise. However, outbreaks of Ebola virus in West Africa and Zika virus in South and Central America were the first instances of real-time sequencing and epidemiologic inference at a large scale (Di Paola et al., 2020; Grubaugh et al., 2017). These tools have broad applications in public health and infection prevention. Genome sequencing can help disentangle transmission clusters in high-density settings, such as in hospitals between patients and healthcare workers (Meredith et al., 2020). They can also monitor the frequency of genomic variants associated with important phenotypic changes, like antigenicity or transmissibility. These efforts can help identify risk factors for transmission, define the local epidemiology of circulating viruses, and enhance contact tracing.

The high mutation rate of RNA viruses and the resolution of genomic sequencing allows for fine-scale analysis of transmission. Based on sequence similarity and phylogenetic relatedness, sequencing can easily rule out transmission linkage of individuals if the genetic distance between two genomes is high. Conversely, identical or near-identical genome sequences plus linkage by traditional contact tracing is powerful evidence of transmission linkage. However, there is a limit to the resolution achieved by genomic epidemiology approaches that are now standard (Villabona-Arenas et al., 2020). A cluster of cases may have identical consensus genomes, making it difficult to determine the exact order of the chain of transmission. This problem is especially difficult for pathogens with lower mutation rates, such as DNA viruses and bacteria (Martin et al., 2018). One potential solution that has been explored largely for bacterial genomics is comparing patterns of intrahost variation (Worby et al., 2014). The concept in its most basic form is that if intrahost genetic variation is shared between transmission pairs but not other individuals, then that might suggest transmission linkage between those individuals. The potential utility of shared intrahost variants for this purpose has been modeled in generalized statistical frameworks and explored in several bacterial pathogens, such as tuberculosis (Maio et al., 2018; Worby et al., 2017). This application of intrahost diversity has received even more interest throughout the current SARS-CoV-2 pandemic, in which the pathogen of interest exhibits extremely low consensus diversity.

Little is known about the validity and utility of viral intrahost diversity in genomic epidemiology. Its validity depends on a complex integration of the topics discussed in this Introduction. In order to make high-quality transmission inferences based on shared intrahost variation, there are important criteria that must be met. The quality of the epidemiologic metadata must be high, such that there is a known set of possible transmission pairs in the cohort. The sequencing and variant identification must be highly accurate; otherwise, spurious and systematic errors could create the appearance of shared variation when there is none. The idiosyncrasies of within-host evolution are also important. There must be a sufficient amount of genetic diversity that is generated during an acute infection and passed to new hosts through a sufficiently large transmission bottleneck (Worby et al., 2017). Lastly, the base rate of parallel mutation in the cohort must be sufficiently small. If there is a large amount of parallel mutation and a tight transmission bottleneck, then it could be more likely that pairs who are not transmission linked share variants compared to actual transmission pairs. Particularly for a pathogen with emerging variants like SARS-CoV-2, these dynamics could differ by time and evolutionary lineage. In sum, the applications of intrahost diversity for genomic epidemiology is not amenable to a "plug-and-chug" approach. The success of this endeavor depends entirely on a fundamental understanding of how viruses evolve within hosts and the technologies used to measure this evolution.

In Chapter IV, I investigated SARS-CoV-2 intrahost diversity and its consequences for sequence-based transmission inference. I found that in the context of a tight transmission bottleneck, the low intrahost diversity and relatively high frequency of parallel mutation in SARS-CoV-2 intrahost populations creates difficult complications for its application in genomic epidemiology.

CHAPTER II

Influenza B Viruses Exhibit Lower Within-Host Diversity Than Influenza A Viruses in Human Hosts

Note: This chapter is a modified version of the published article:

Valesano AL, Fitzsimmons WJ, McCrone JT, Petrie JG, Monto AS, Martin ET, Lauring AS. 2020. "Influenza B viruses exhibit lower within-host diversity than influenza A viruses in human hosts." J Virol 94:e01710-19.

The American Society of Microbiology grants all authors full rights to reuse their articles in dissertations.

Introduction

Influenza viruses rapidly mutate and evolve through selection, genetic drift, and reassortment (Moya et al., 2004). At a global scale, influenza A virus (IAV) and influenza B virus (IBV) evolve under strong positive selection driven by pressure for escape from pre-existing population immunity (Nelson and Holmes, 2007; Rambaut et al., 2008). Selection of new antigenic variants contributes to reduced effectiveness of seasonal influenza vaccines, necessitating annual updates of vaccine strains (Yamayoshi and Kawaoka, 2019). IAV and IBV both undergo seasonal antigenic drift and share a similar genomic architecture, but their ecology and evolution differ in important ways (Petrova and Russell, 2018). While IBV accounts for roughly one-third of influenza's burden of morbidity and mortality (Paul Glezen et al., 2013; Thompson et al., 2003), it circulates only in humans and seals and is considered to be a lower pandemic risk than influenza A (IAV) due to its limited animal reservoirs. Like IAV, there are co-circulating, antigenically distinct lineages of IBV that are included in the quadrivalent influenza vaccine. Two lineages of IBV diverged in the 1980s, B/Victoria/2/87-like and B/Yamagata/16/88-like, here referred to as B/Victoria and B/Yamagata, respectively (Rota et al., 1990).

IBV evolves more slowly than IAV on a global scale and has a lower rate of antigenic drift, but the reasons for this are poorly understood (Petrova and Russell, 2018; Yamashita et al., 1988). Similar evolutionary forces are involved in the antigenic evolution of both IAV and IBV, generally characterized by non-synonymous substitutions at antigenic sites in the surface hemagglutinin (HA) protein (Chen and Holmes, 2008; Shen et al., 2009) and reassortment within and between lineages (Dudas et al., 2015; Langat et al., 2017; Vijaykrishna et al., 2015). The IBV polymerase has a lower mutation rate relative to IAV (Nobusawa and Sato, 2006). However, it is unclear whether the slower global evolution of IBV is driven by its lower mutation rate or other differences in selection at the global scale.

All new seasonal influenza variants are ultimately derived from *de novo* mutations within individual hosts (Xue et al., 2018). Therefore, understanding how new variants arise within individuals and transmit between them is essential to defining how novel viruses spread in host populations. For example, if the relative mutation rate is a major factor underlying the global evolutionary differences across IAV and IBV, we might also expect to see differences in their within-host dynamics. We and others have used next-generation sequencing to investigate the within- and between-host evolutionary dynamics of IAV in humans (Debbink et al., 2017; Dinis et al., 2016; Leonard et al., 2016; McCrone et al., 2018; Xue and Bloom, 2019a; Xue et al., 2018). We have found that there is little accumulation of intrahost variants during acute infections of immunocompetent individuals (Leonard et al., 2016; McCrone et al., 2018), and we have not found evidence of changes in intrahost diversity by vaccination status or other proxies for immunological history (Debbink et al., 2017; Han et al., 2018; McCrone et al., 2018). The IAV transmission bottleneck is stringent (McCrone et al., 2018), which generally means that few variants that arise within hosts are able to transmit. Together, these studies suggest that positive selection of novel variants is an inefficient process in IAV-infected hosts, contrasting with its patterns of significant positive selection at the global level (Xue and Bloom, 2019b). Despite the importance of intrahost processes to influenza virus evolution, these dynamics have not been systematically investigated in IBV.

Here we use next-generation sequencing to define the within-host diversity of IBV populations from individuals enrolled in the Household Influenza Vaccine Evaluation (HIVE) study, a community-based household cohort initiated in 2010. We apply a previously-validated sequencing approach and bioinformatic pipeline (Debbink et al., 2017; McCrone and Lauring, 2016; McCrone et al., 2018) to identify intrahost single-nucleotide variants (iSNV) arising during infection with B/Victoria and B/Yamagata viruses. We find that IBV has significantly lower intrahost diversity than IAV, consistent with its lower mutation rate and slower rate of evolution. We analyze shared iSNV across 15 genetically validated household transmission pairs and find that, like IAV, IBV is also subject to a tight genetic bottleneck at transmission. These data provide the first systematic evaluation of the genetic architecture of IBV populations during natural human infection and provide insights into the comparative epidemiology and evolution of influenza viruses.

Methods

Description of the HIVE cohort

The HIVE study is a prospective, community-based household cohort in Southeastern Michigan based at the University of Michigan School of Public Health (Monto et al., 2014, 2019; Ohmit et al., 2013, 2016; Petrie et al., 2013, 2017). The cohort was initiated in 2010, with enrollment of households with children occurring on an annual basis and an active surveillance period lasting from October through May. In 2014, active surveillance was expanded to take place year-round. Participating adults provided informed consent for themselves and their children, and children ages 7-17 provided oral assent. Individuals in each household were followed prospectively for acute respiratory illness, defined as two or more of the following: cough, fever or feverishness, nasal congestion, chills, headache, body aches, or sore throat. Study participants meeting the criteria for acute respiratory illness attended a study research clinic at the University of Michigan School of Public Health where a combined throat and nasal swab, or a nasal swab only for children less than three years old, was collected by the study team. Beginning in the 2014-2015 season, study participants with acute respiratory illnesses took an additional nasal swab at home at the time of illness onset, collected either by themselves or by a parent. The study was approved by the Institutional Review Board of the University of Michigan Medical School.

Viral detection, lineage typing, and viral load quantification

We processed upper respiratory specimens (combined nasal and throat swab or nasal swab) for confirmation of influenza virus infection by reverse transcription polymerase chain reaction (RT-PCR). We extracted viral RNA with either QIAamp Viral RNA Mini Kits (Qiagen) or PureLink Pro 96 Viral RNA/DNA Purification kits (Invitrogen) and tested samples using the SuperScript III Platinum One-Step Quantitative RT-PCR System with ROX (Invitrogen) and primers and probes for universal detection of influenza A and B (CDC protocol, 28 April 2009). Specimens positive for influenza virus were tested using subtype/lineage primer and probe sets, which are designed to detect influenza A (H3N2), A (H1N1)pdm09, B (Yamagata), and B (Victoria). An RNAseP primer/probe set was run for each specimen to confirm specimen quality and successful RNA extraction.

We quantified the viral load in each sample by RT-qPCR using primers specific for the open reading frame of segment 8 (NS1/NEP): forward primer 5'-TCCTCAACTCACTCTTCGAGCG-3', reverse primer 5'-CGGTGCTCTTGACCAAATTGG-3', and probe 5'-(FAM)-CCAATTCGAGCAGCTGAAACTGCGGTG-(BHQ1)-3'. Each reaction contained 5.4 μL of nuclease-free water, 0.5 μL of each primer at 50 μM, 0.1 μL of ROX dye, 0.5 μL SuperScript III RT/Platinum Taq enzyme mix, 0.5 μL of 10 μM probe, 12.5 μL of 2x PCR buffer master mix, and 5 μL of extracted viral RNA. To relate genome copy number to Ct value, we used a standard curve based on serial dilutions of a plasmid control, run in duplicate on the same plate.

Amplification, library preparation, and sequencing

We amplified viral cDNA from all eight genomic segments using the SuperScript III One-Step RT-PCR Platinum Taq HiFi Kit (Invitrogen). Each reaction contained 5 μL of extracted viral RNA, 12.5 μL of 2x PCR buffer, 2 μL of primer cocktail, 0.5 μL of enzyme mix, 5 μL of nuclease-free water. The primer cocktail was a mixture of B-PBs-UniF, B-PBs-UniR, B-PA-UniF, B-PA-UniR, B-HANA-UniF, B-HANA-UniR, B-NP-UniF, B-NP-UniR, B-M-Uni3F, B-Mg-Uni3F, B-M-Uni3R, B-NS-Uni3F, and B-NS-Uni3R (sequences and proportions are listed in ref. (Zhou et al., 2014)). The thermocycler protocol was: 45 ˚C for 60 min, 55 ˚C for 30 min, 94 ˚C for 2 min, then 5 cycles of 94 ˚C for 20 s, 40 ˚C for 30 s, 68 ˚C for 3 min 30 s, then 40 cycles of 94 ˚C for 20 s, 58 ˚C for 30 s, 68 ˚C for 3 min 30 s, and a final extension of 68 ˚C for 10 min. We confirmed IBV genome amplification by gel electrophoresis. We sheared amplified cDNA (100-500 ng) on a Covaris ultrasonicator with the following settings: time 80 sec, duty cycle 10%, intensity 4, cycles per burst 200. We prepared sequencing libraries with NEBNext Ultra DNA Library Prep kits (NEB) and sequenced them on an Illumina NextSeq with 2x150 paired end reads (mid-output run, v2 chemistry). To increase the specificity of variant identification, samples with a viral load between 10^3 and 10^5 genome copies/μL of transport media were amplified and sequenced in duplicate. Samples amplified from B/Victoria and B/Yamagata plasmid clones were included on each sequencing run to account for sequencing errors. The plasmids used in the control reactions were generated by segment-specific RT-PCR from clinical samples of B/Victoria and B/Yamagata strains from the 2012-2013 season followed by gel extraction and TOPO-TA cloning (Invitrogen). The sequence of each plasmid was determined by Sanger sequencing. We generated the plasmid control amplicons included on each Illumina sequencing run using the same multiplex amplification protocol, but with cloned plasmid DNA as the template.

Identification of iSNV

Intrahost single-nucleotide variants (iSNV) were identified using a previously-described analytic pipeline (McCrone and Lauring, 2016). We identified iSNV in samples that had an average genome coverage greater than 1000x and a viral load greater than 10^3 genome copies per microliter of transport media in the original sample. Sequencing adapters were removed with cutadapt (Martin, 2011) and reads were aligned to the sequences derived from the B/Victoria and B/Yamagata plasmid controls with Bowtie2 (Langmead and Salzberg, 2012). Duplicate reads were marked and removed with Picard and samtools (Li et al., 2009). Putative variants were identified with the R package deepSNV using data from the clonal plasmid controls of each sequencing run (Gerstung et al., 2012). Minority iSNV (<50% frequency) were identified using the following empirically-derived criteria: deepSNV p-value <0.01, average mapping quality >30, average Phred score >35, and average read position in the middle 50% (positions 37 and 113 for 150 base pair reads). For samples processed in duplicate, we used only variants that were present in both replicates; the frequency of the variant in the replicate with greater coverage at that site was used. Lastly, variants with frequency <2%, which have a higher false positive rate from RT-PCR and/or sequencing errors, were not included in downstream analyses.

In our previous work on IAV, we found that there were multiple sites with mutations that were essentially fixed (>0.95) relative to the plasmid control and in which the base in the plasmid control was therefore identified as a minority variant in the sample (McCrone et al., 2018). At these sites, deepSNV is unable to estimate the base-specific error rate and cannot distinguish true minority iSNV; however, we found that we could accurately identify minority variants at these sites at a frequency of 2% or above (McCrone et al., 2018). This frequency threshold was incorporated into the pipeline for iSNV identification at these sites. Therefore, we identify intrahost variants with frequencies between 2-98%. Minority iSNV are the subset of these variants with a frequency between 2-50% relative to the sample consensus, which we use as a metric of within-host diversity. For each minority iSNV, we identify the majority iSNV present at a frequency of 50-98%. Any sites that were monomorphic after applying quality filters were assigned a frequency of 100%. Nucleotide diversity (π) was calculated using identified iSNV in each sample using the formula described in Zhao and Illingoworth (Zhao and Illingworth, 2019).

Data and code availability

Raw sequence data, with human content filtered out, are available at the NCBI Sequence Read Archive under BioProject accession number PRJNA561158. Code for the variant identification pipeline is available at http://github.com/lauringlab/variant_pipeline. Analysis code is available at http://github.com/lauringlab/Host_level_IBV_evolution.

Results

We used high depth-of-coverage sequencing to define the intrahost genetic diversity in IBV-positive samples collected from individuals in the HIVE, a prospective, household cohort in southeastern Michigan that follows 200-350 households annually (Table 2.1). This cohort provides an opportunity to investigate natural infections and transmission events in a community context. Individuals that meet symptom-based criteria for an upper respiratory illness during the surveillance period undergo collection of nasal and throat swabs for molecular detection of respiratory viruses by RT-PCR. Starting in 2014-2015, individuals also provided a sample collected at home prior to subsequent collection of a second specimen at the on-site clinic.

Table 2.1. Influenza B viruses over seven seasons in a household cohort.

	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
Households	328	213	321	232	340	227	208
Participants	1441	943	1426	1049	1431	996	890
Vaccinated n(%)[a]	934 (65)	554 (59)	942 (66)	722 (69)	992 (69)	681 (68)	611 (69)
IBV Positive Individuals[b]	45	7	49	4	44	11	30
B/Yamagata	1	3	38	4	34	5	26
B/Victoria	37	0	10	0	10	6	4
IBV Positive Households[c]							
Two Individuals	10	2	5	0	11	2	4
Three Individuals	0	1	1	0	1	0	2
High Quality NGS Data[d]	13	2	20	1	32	11	20

[a] Self-reported or confirmed receipt of vaccine prior to the specified season.

[b] RT-PCR confirmed infection.

[c] Households in which two individuals were positive within 7 days of each other. In cases of trios, the putative chains could have no pair with onset >7 days apart.

[d] Samples with $>10^3$ genome copies per µl of transport medium, adequate amplification of all 8 genomic segments, and average sequencing coverage $>10^3$ per nucleotide.

Over seven seasons (2010-2011 through 2016-2017) and 8176 person-seasons of observation, we identified 111 individuals infected with B/Yamagata and 67 infected with B/Victoria (Table 2.1). Several households had clusters of infections of two or three IBV-positive individuals within 7 days of each other, suggestive of within-household transmission. Because variant identification is sensitive to input viral titer (McCrone and Lauring, 2016), we first measured viral loads of all available IBV-positive samples by RT-qPCR (Figure 2.1A). Any samples with a viral load below 10^3 copies/μL were not submitted for sequencing. For samples with a viral load in the range of 10^3-10^5 copies/μL, we performed two independent RT-PCR reactions and sequenced replicate libraries on separate sequencing runs. We sequenced samples with viral loads above 10^5 copies/μL of transport media in a single replicate. From the available IBV-positive samples, we were able to obtain sequence data on 106 samples from 91 individuals, consisting of 35 individuals infected with B/Victoria and 56 infected with B/Yamagata (Table 2.1).

We identified intrahost single nucleotide variants (iSNV) using our previously validated bioinformatic pipeline. As in our previous work, we report iSNV at frequencies of 2% or above, for which we have well-defined sensitivity and specificity (McCrone et al., 2018). We consider sites with >98% frequency to be essentially fixed, setting the frequency at those sites to 100% (see Materials and Methods). We achieved a mean coverage of 10,000x per sample across most genome segments, with generally lower coverage on segments encoding NP and NS (Figure 2.1B). We restricted our analysis of iSNV to samples with an average genome coverage of greater than 1000x, which includes 99 of the original 106 sequenced samples.

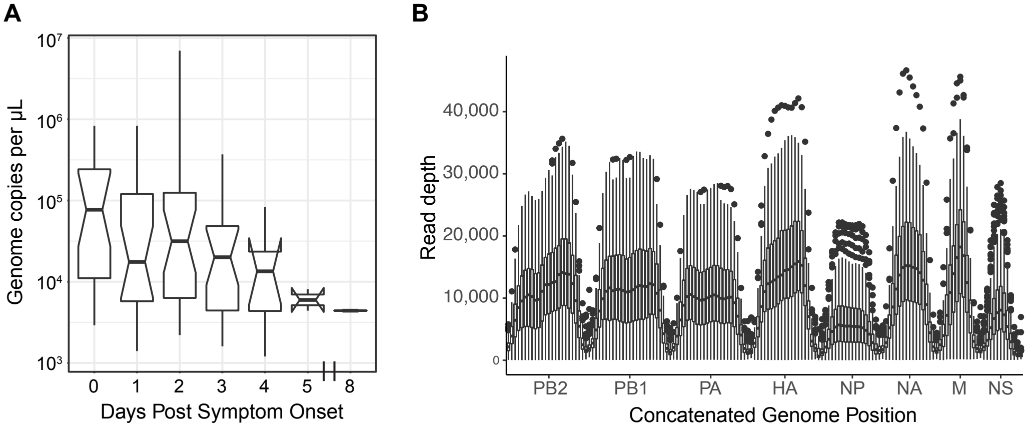


Figure 2.1. Viral load and sequencing coverage. (A) Boxplot of viral load (genome copies per microliter of swab transport media, y-axis) by day of sampling relative to symptom onset (x-axis). The boxes display median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range. (B) Sequencing coverage is plotted with read depth on the y-axis and location within a concatenated influenza B virus genome on the x-axis. The mean coverage for each sample was calculated over a sliding window of size 200 and a step size of 100. The data are displayed for all samples at each window as a boxplot, showing the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range; all values outside this range are shown as individual points.

Within-host genetic diversity of IBV in natural infections

All samples exhibited low genetic diversity. The vast majority had no iSNV above the 2% cutoff. Of the 99 samples with high-quality NGS data, 70 had no minority iSNV, 17 had one iSNV, 7 had two iSNV, and 3 samples had 3 iSNV (median 0, IQR 0-1; Table 2.2). Two outliers had a large number of iSNV, with 8 and 20 iSNV. These two samples came from the same individual, with one collected at home and the second at the study clinic two days later. Most of the iSNV in these two samples were present at similar frequencies, 3-5% in the home sample and 17-23% in the clinic sample (Table 2.3), both of which were sequenced in duplicate on separate Illumina runs. The high number of mutations present at similar frequencies is suggestive of a mixed infection with distinct haplotypes or strains as opposed to *de novo* mutations arising on a single genetic background. The iSNV in the home-collected sample are all found in the subsequent clinic-collected sample, each with a similar change in frequency across the two samples. This further supports the conclusion that these mutations are on the same genome in a mixed infection with two distinct strains.

Table 2.2. Identified iSNV, excluding samples from one putative mixed infection.

				Viral			Amino		
Enrollee	Specimen	Season	Lineage	Load[a]	Gene	Nucleotide[b]	Acid[c]	Frequency	Vaccinated[d]
50207	MH15919	16/17	B/Victoria	3.50E+03	M	A650G	N207S	0.024	IIV4
331001	MH2671	12/13	B/Victoria	3.20E+05	NA	C276T	L73F	0.086	LAIV3
331001	MH2671	12/13	B/Victoria	3.20E+05	PA	A2047G	K671R	0.474	LAIV3
330171	MH3227	12/13	B/Victoria	7.00E+06	NA	A385C	N109T	0.162	IIV3
330171	MH3227	12/13	B/Victoria	7.00E+06	PA	C1982T	A649A	0.461	IIV3
301587	M53957	10/11	B/Victoria	3.30E+04	HA	G1603A	G522R	0.081	No
301587	M53957	10/11	B/Victoria	3.30E+04	NA	A863C	T268T	0.038	No
301202	M54308	10/11	B/Victoria	4.40E+04	PA	C1037T	N334N	0.195	No
50003	MH10403	14/15	B/Victoria	8.20E+04	NS	A103C	T18T	0.063	No
50004	MH10404	14/15	B/Victoria	8.20E+04	NP	A577G	N171S	0.057	No
50004	MH10404	14/15	B/Victoria	8.20E+04	NA	A1457G	L466L	0.223	No
50004	MH10404	14/15	B/Victoria	8.20E+04	PA	G1617A	V528M	0.497	No
50424	HS1876	14/15	B/Victoria	1.60E+03	NP	G1191A	D376N	0.034	IIV4
50051	HS1909	14/15	B/Victoria	1.90E+03	M	G709A	E227K	0.343	Yes, Unk
50004	HS1788	14/15	B/Victoria	8.30E+05	NP	A577G	N171S	0.054	No
50004	HS1788	14/15	B/Victoria	8.30E+05	PA	G1617A	V528M	0.389	No
50004	HS1788	14/15	B/Victoria	8.30E+05	NA	A1457G	L466L	0.045	No
50312	HS2019	15/16	B/Victoria	2.00E+05	NP	G987A	V308I	0.420	No
50312	HS2019	15/16	B/Victoria	2.00E+05	PA	G1346A	E437E	0.467	No
51123	HS2680	15/16	B/Victoria	3.50E+05	NP	G1511A	R482R	0.344	No
320779	MH0776	11/12	B/Yamagata	3.40E+05	NP	A735G	S223S	0.023	IIV3
320779	MH0776	11/12	B/Yamagata	3.40E+05	PB2	G661A	R211R	0.222	IIV3
51092	MH10076	14/15	B/Yamagata	1.20E+04	PB1	A223G	I66V	0.116	IIV4
50650	MH16167	16/17	B/Yamagata	5.20E+04	PA	T2019C	L662L	0.373	IIV4
50650	MH16167	16/17	B/Yamagata	5.20E+04	PB1	C345T	A106A	0.159	IIV4
331060	MH3065	12/13	B/Yamagata	3.70E+05	PA	A1912G	K626R	0.051	LAIV3
331397	MH4247	12/13	B/Yamagata	2.40E+04	PB2	A676G	R216R	0.370	IIV3
330459	MH4289	12/13	B/Yamagata	2.10E+05	HA	G1102A	A355T	0.024	IIV3
330460	MH4364	12/13	B/Yamagata	2.10E+05	PB2	G520A	V164V	0.032	IIV3
50006	MH16139	16/17	B/Yamagata	1.20E+05	HA	T728C	F230S	0.148	No
331471	MH2216	12/13	B/Yamagata	8.60E+04	PB2	G1936A	Q636Q	0.024	No
331470	MH2246	12/13	B/Yamagata	1.20E+04	PA	G1535A	A500A	0.029	No
331470	MH2246	12/13	B/Yamagata	1.20E+04	PB2	A2253G	K742R	0.124	No
331470	MH2246	12/13	B/Yamagata	1.20E+04	PB2	C769T	H247H	0.023	No
331364	MH4166	12/13	B/Yamagata	2.80E+04	HA	C746T	T236I	0.037	No
331364	MH4166	12/13	B/Yamagata	2.80E+04	PA	G1298A	L421L	0.093	No
UM41536	MH6592	13/14	B/Yamagata	2.00E+04	PB1	G1893A	R622R	0.022	No
51093	HS1747	14/15	B/Yamagata	3.90E+04	PA	G1433A	L466L	0.087	IIV4
50419	HS3214	16/17	B/Yamagata	5.40E+05	PB1	C345T	A106A	0.046	IIV4
51121	HS3258	16/17	B/Yamagata	1.10E+04	PB1	A2079G	E684E	0.022	No

[a] Viral load measured by RT-qPCR, expressed in genome copies per microliter of transport medium.

[b] Consensus nucleotide followed by position on reference genome and variant nucleotide.

[c] Consensus amino acid followed by codon position on reference genome and variant amino acid.

[d] Self-reported or confirmed receipt of vaccine prior to the specified season. IIV4, quadrivalent inactivated; LAIV3, trivalent live attenuated; IIV3, trivalent inactivated; Unk, vaccine product unknown.

Table 2.3. Identified iSNV in one vaccinated individual[a] with a putative mixed infection.

Specimen[a]	Viral Load[b]	Gene	Nucleotide[c]	Amino Acid[d]	Frequency
HS1875	2.90E+04	HA	G1061A	R341K	0.029
HS1875	2.90E+04	NP	G1666A	G534D	0.027
HS1875	2.90E+04	NA	G1210A	G384D	0.054
HS1875	2.90E+04	NA	A798G	S247G	0.052
HS1875	2.90E+04	NS	G1004A	V100V	0.045
HS1875	2.90E+04	PB2	G817A	V263V	0.036
HS1875	2.90E+04	PB2	A1231C	I401I	0.032
HS1875	2.90E+04	PB2	A793G	E255E	0.035
MH10536	3.80E+04	HA	G1061A	R341K	0.236
MH10536	3.80E+04	HA	C366T	C109C	0.213
MH10536	3.80E+04	M	A114G	L28L	0.189
MH10536	3.80E+04	NP	G1666A	G534D	0.193
MH10536	3.80E+04	NP	G1257A	R398R	0.166
MH10536	3.80E+04	NA	G1210A	G384D	0.239
MH10536	3.80E+04	NA	C1286T	Y409Y	0.218
MH10536	3.80E+04	NA	C1319T	C420C	0.217
MH10536	3.80E+04	NA	G1148A	R363R	0.225
MH10536	3.80E+04	NA	A798G	S247G	0.233
MH10536	3.80E+04	NA	T816C	F253L	0.06
MH10536	3.80E+04	NS	G1004A	V100V	0.185
MH10536	3.80E+04	NS	G596A	V183I	0.190
MH10536	3.80E+04	NS	T469A	V140V	0.198
MH10536	3.80E+04	NS	T66C	M6T	0.173
MH10536	3.80E+04	PA	G1279A	S415N	0.188
MH10536	3.80E+04	PB1	T1932A	S635S	0.214
MH10536	3.80E+04	PB2	G817A	V263V	0.217
MH10536	3.80E+04	PB2	A1231C	I401I	0.218
MH10536	3.80E+04	PB2	A793G	E255E	0.221

[a] Enrollee number 50425. HS Indicates home specimen and MH indicates clinic specimen, both from same individual
[b] Viral load measured by RT-qPCR, expressed in genome copies per microliter of transport medium.
[c] Consensus nucleotide followed by position on reference genome and variant nucleotide.
[d] Consensus amino acid followed by codon position on reference genome and variant amino acid.

We examined how within-host diversity changes by day of sampling during IBV infections, as the virus population rapidly expands and contracts. As we have previously shown that specimen viral load can affect the sensitivity and specificity of variant identification (McCrone and Lauring, 2016), we sought to control for this variable in our analysis. Although viral load generally decreased with time after symptom onset (Figure 2.1A), we found that within-host diversity as measured by number of identified minority iSNV did not vary with viral load (Figure 2.2A; p = 0.2996, adjusted r-squared = 0.0009) or with day of infection (Figure 2.2B; p = 0.62, ANOVA). The frequencies of the identified iSNV were consistent across replicate libraries from the same samples, indicating that our measurements of iSNV frequency are precise (Figure 2.2C).

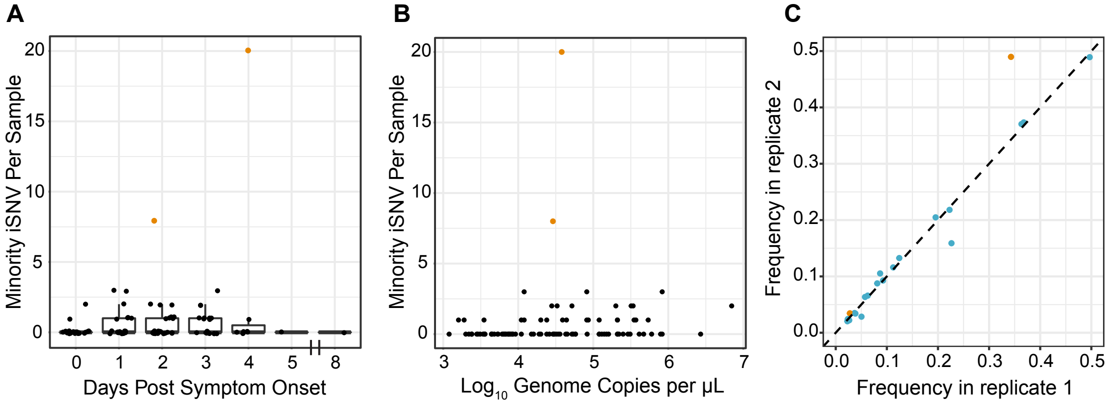


Figure 2.2. Intrahost minority SNV by day post-symptom onset and viral load. (A) Number of minority iSNV per sample is plotted on the y-axis by day post symptom onset on the x-axis. Data are displayed as boxplots representing the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ±1.5 times the interquartile range. The raw data points are shown in black overlaid on top of the boxplots; points from mixed infection samples are shown in orange. (B) Scatterplot relating the number of minority iSNV per sample on the y-axis to the $\log_{10}$ of viral load, in genome copies per microliter, on the x-axis. Data points from the mixed infection are shown in orange. (C) Frequency of minority iSNV in samples sequenced in duplicate. Orange dots represent variants identified in samples with viral load of 10^3 – 10^4 genome copies per microliter and blue dots represent variants in samples with viral load of 10^4 – 10^5 genome copies per microliter.

We detected minority iSNV across all eight genome segments (Figure 2.3). The ratio of nonsynonymous to synonymous iSNV was 0.74, which given the excess of nonsynonymous sites across the genome, suggests significant purifying selection. There was only one minority iSNV present in more than one individual; we identified a variant encoding a synonymous mutation in PB1 in two individuals from separate households infected with B/Yamagata in the 2016-2017 season. We did not identify any nonsynonymous minority iSNV in the known antigenic sites of IBV hemagglutinin, which suggests that positive selective pressure for variants that escape antibody-mediated immunity is not particularly strong within hosts. We found that there is no difference in the distribution of the number of iSNV per sample between vaccinated and non-vaccinated individuals (Figure 2.4A). During the first few seasons of the study, some individuals received trivalent vaccines, which contain only one of the two IBV lineages. We therefore repeated this analysis, excluding 3 individuals for whom we had no information about specific vaccine product and re-classifying 6 individuals who received trivalent vaccines and were infected with a lineage not included in that season's trivalent formulation as "unvaccinated." We again found no difference in the number of iSNV between groups (MWU test, p = 0.9103). Together, these data indicate that vaccine-induced immunity is not a major diversifying force for IBV within hosts in our study population. This is consistent with our previous work on IAV in the HIVE as well as a randomized-controlled trial of vaccine efficacy (FLU-VACS), both of which showed no difference in intrahost diversity based on same-season vaccination status (Debbink et al., 2017; McCrone et al., 2018). Intrahost diversity was similar between B/Victoria and B/Yamagata virus populations (Figure 2.4B), consistent with our previous comparison of subtype A/H3N2 and A/H1N1 viruses (McCrone et al., 2018).

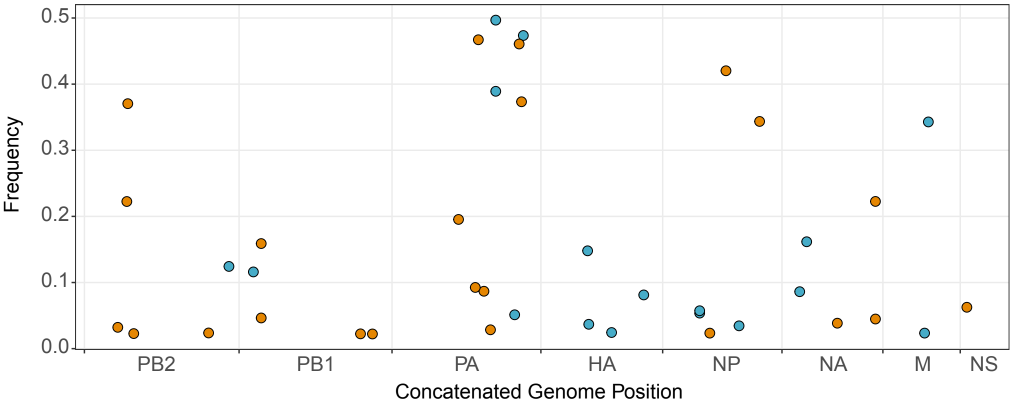


Figure 2.3. Intrahost SNV frequency by genome position and mutation type. All minority (<50%) iSNV from 99 samples are displayed with their frequency on the y-axis and their position within a concatenated influenza B virus genome on the x-axis. Synonymous mutations are shown in orange and nonsynonymous mutations in blue.

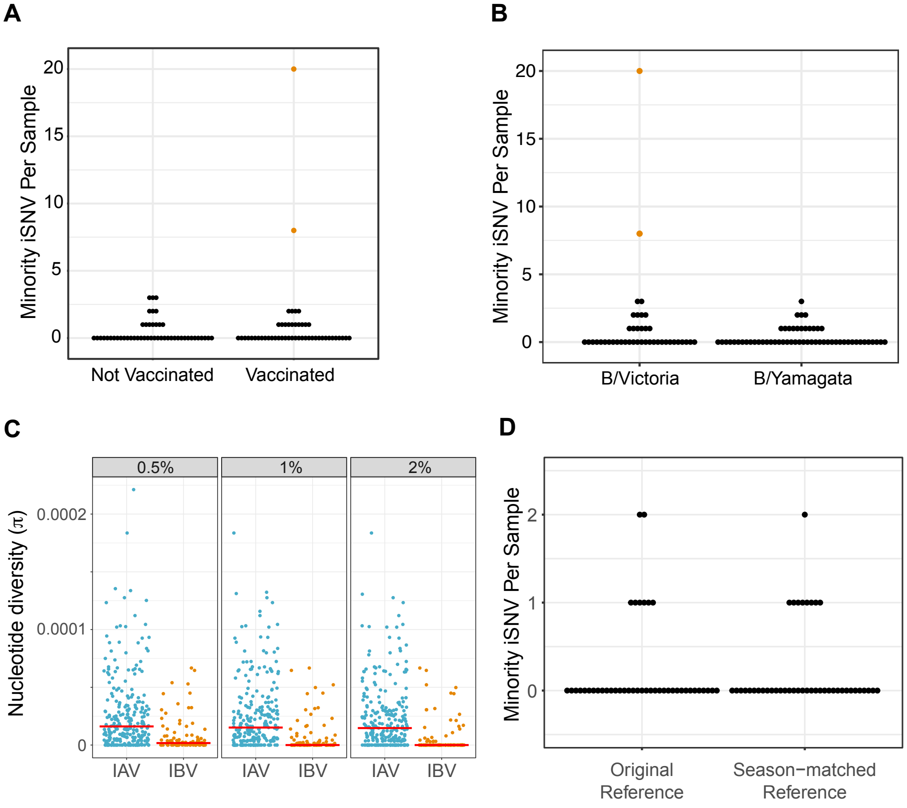


Figure 2.4. Intrahost SNV by vaccination status and IBV lineage. (A) Numbers of minority iSNV per sample across all 99 samples are shown (y-axis) by current-season vaccination status of the host (x-axis). Samples from the mixed infection are shown in orange. (B) Numbers of minority iSNV per sample are shown (y-axis) by IBV lineage (x-axis). Samples from the mixed infection are shown in orange. (C) Pairwise nucleotide diversity (π, y-axis) by influenza virus type (x-axis), stratified by iSNV frequency cut-off (top). Medians are shown as red lines. Data for influenza A virus are from 243 samples described in McCrone et al. 2018. Data on influenza B virus are from 97 high-quality samples in the present study. Samples from mixed infections in both studies are excluded. (D) Numbers of minority iSNV in 43 of the 99 high-quality samples (y-axis), consisting of B/Yamagata from the 2014/2015 season, B/Victoria from the 2015/2016 season, and B/Yamagata from the 2016/2017 season based on alignments to the original references from the 2012/2013 season vs. season-matched reference genomes (x-axis).

We compared the within-host genetic diversity of IBV to our previously published data on IAV from the HIVE cohort (McCrone et al., 2018). Here, IBV exhibits lower within-host diversity compared to IAV (Table 2.4; $p < 0.001$, MWU test). IBV samples had a lower median number of minority iSNV (median 0, IQR 0-1) compared to IAV (median 2, IQR 1-3); 71% of IBV samples contained no minority iSNV compared to 20% of IAV samples. The difference in within-host diversity was robust to relaxation of the iSNV frequency cutoff to 1% and 0.5%. IBV also exhibited a lower within-host diversity as measured by nucleotide diversity (π), which takes both the number and the frequency of iSNV into account (Figure 2.4C and Table 2.4). To ensure our results were not an artifact of overly stringent quality thresholds, we also identified minority iSNV with less conservative read mapping quality (MapQ) and base quality (Phred) scores. We identified the same set of minority iSNV with a MapQ cut-off of 20 as with the original cutoff of 30. Similarly, reduction of the Phred base-quality cutoff to >25 in addition to a MapQ score cutoff of >20 resulted in only 20 more minority iSNV, eight of which were found in the individual with a mixed infection. The other additional 12 minority iSNV were dispersed across specimens and did not significantly change the overall distribution of within-host diversity. We also examined whether our results were biased by use of a single B/Yamagata and B/Victoria reference for alignment and variant calling, which were both drawn from the 2012-2013 season (see Materials and Methods). We realigned sequence data from 43 of the original 99 samples to season-specific reference genomes isolated in southeastern Michigan. We found that the overall alignment rate for any given specimen was similar between the original reference and the new season-matched reference. Variant identification based on the new references and the original quality thresholds resulted in the same distribution of within-host diversity, although the identity of some iSNV was different (Figure 2.4D).

Table 2.4. Within-host diversity of IAV versus IBV.

	Minority iSNV Richness			Nucleotide Diversity (π)		
Frequency Cutoff	IAV[a]	IBV[a]	p-value (MWU test)	IAV[a]	IBV[a]	p-value (MWU test)
2%	1 (1 - 3, 0.23)	0 (0 - 1, 0.73)	< 0.001	1.47e-05 (3.56e-06 - 3.28e-05, 0.23)	0 (0 - 3.07e-06, 0.73)	< 0.001
1%	2 (1 - 3, 0.20)	0 (0 - 1, 0.56)	< 0.001	1.52e-05 (3.77e-06 - 3.37e-05, 0.20)	0 (0 - 4.16e-06, 0.56)	< 0.001
0.5%	2 (1 - 4, 0.16)	1 (0 - 2, 0.32)	< 0.001	1.61e-05 (3.96e-06 - 3.46e-05, 0.16)	1.73e-06 (0 - 6.4e-06, 0.32)	< 0.001

[a] Median (IQR, proportion at zero)

Together, these results indicate that our measurements of within-host diversity are robust to several technical aspects of variant identification, which are unlikely to account for the lower observed diversity of IBV. Because these data are from the same cohort and were generated using the same sequencing approach and analytic pipeline as our previous IAV datasets, the observed differences likely reflect true biological differences between IAV and IBV.

Identification of household transmission pairs

We compared viral diversity across samples from individuals in the same household to investigate the genetic bottleneck that influenza B viruses experience during natural transmission. Over the seven influenza seasons, thirty-nine households in the HIVE cohort had two or more individuals positive for the same IBV lineage within a 7-day interval (Table 2.1). This epidemiologic linkage is suggestive of transmission events but does not rule out co-incident community acquired infection (McCrone et al., 2018). We identified 16 putative transmission pairs for which we sequenced at least one sample from each individual. In one of these pairs, the putative recipient was the individual with a mixed infection. The donor did not have evidence of a mixed infection based on number of iSNV, which would imply that the recipient may have been infected twice or that the second virus was lost from the donor by the time of sampling. This pair was excluded from the between-host analysis, leaving 15 putative transmission pairs for which we have high-quality sequencing data on both donor and recipient influenza populations.

We used our sequencing data to determine which of these epidemiologically linked household pairs were actual IBV transmission pairs. We generated maximum likelihood phylogenetic trees for samples from the two IBV lineages using the concatenated coding consensus sequences.

Phylogenetic analysis provided genetic evidence that the 15 epidemiologically-linked pairs were indeed true transmission pairs, as epidemiologically-linked pairs were found nearest each other in each tree (Figure 2.5A and 2.5B; vertical bars with household ID). We also validated these transmission pairs by analyzing the genetic distance across viral populations. True transmission pairs should have genetically similar populations exhibiting low genetic distance, while individuals with coincident community acquisition are more likely to have populations with a higher genetic distance. We compared the genetic distance between epidemiologically-linked household pairs and random community pairs from the same season and infected with the same IBV lineage, using L1-norm as measurement of genetic distance (Figure 2.5C). The distribution of random community pairs functions as a null model of genetic distances among locally circulating strains. All of the 15 putative transmission pairs fell on the tail of this distribution, below the 5th percentile of the community pair L1-norm distribution, indicating that they are true transmission pairs (Figure 2.5C). While the L1-norm is a function of both the consensus sequence and the iSNV, this signal was predominantly driven by consensus differences, as reflected in the phylogenetic analysis.

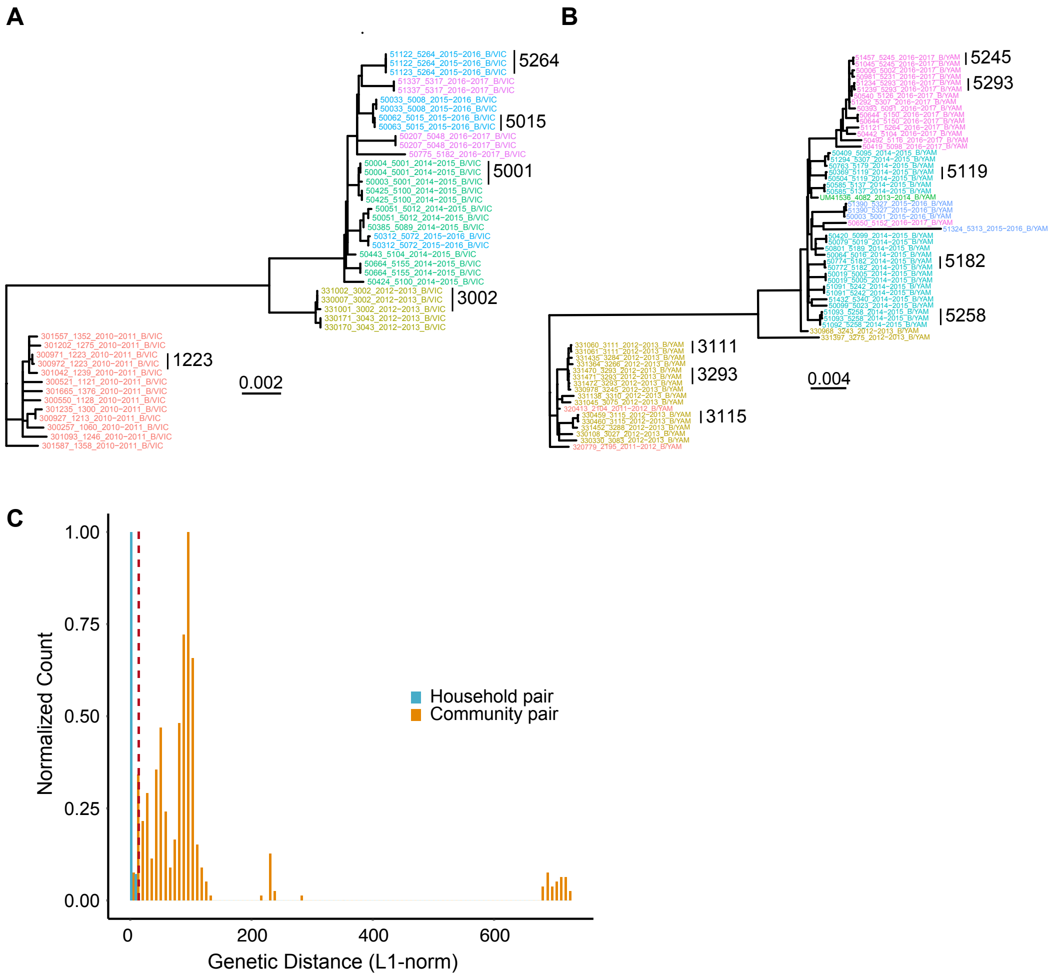


Figure 2.5. Identification of household transmission pairs. Maximum likelihood phylogenetic tree of all B/Victoria (A) and B/Yamagata (B) samples from this study. Concatenated consensus coding sequences were aligned with MUSCLE and phylogenetic trees constructed with RAxML. Tip labels are denoted as enrollee ID, household ID, season, and lineage, separated by underscores; tip labels are color-coded by season. (C) Histogram of genetic distance, as measured by L1-norm, between household pairs and random community pairs from the same season and lineage. The bar heights for each group are normalized to the maximum for each group for comparison. Community pairs are shown in orange and household pairs shown in blue. The dotted red line indicates the 5th percentile of the community pair distribution.

Comparison of viral diversity across transmission pairs

Transmission bottlenecks restrict the genetic diversity that is passed between hosts. With a loose transmission bottleneck, many unique genomes will be passed from donor to recipient. Because this will allow two variants at a given site to be transmitted, sites that are polymorphic in the donor are more likely to be polymorphic in the recipient. However, in the case of a tight or stringent bottleneck, sites that are polymorphic in the donor will likely be either fixed or absent in the recipient. We have previously demonstrated that influenza A experiences a tight transmission bottleneck of 1-2 unique genomes (McCrone et al., 2018). Across our 15 IBV transmission pairs, we found no sites that were polymorphic in the donor and recipient (Figure 2.6). Intrahost SNV present in the donor were either fixed (100%) or absent (0%) in the recipient. These data suggest a stringent transmission bottleneck for influenza B, similar to that of influenza A. As there were fewer samples, transmission pairs, and iSNV in our IBV dataset, we were unable to obtain a robust and precise estimate of bottleneck size.

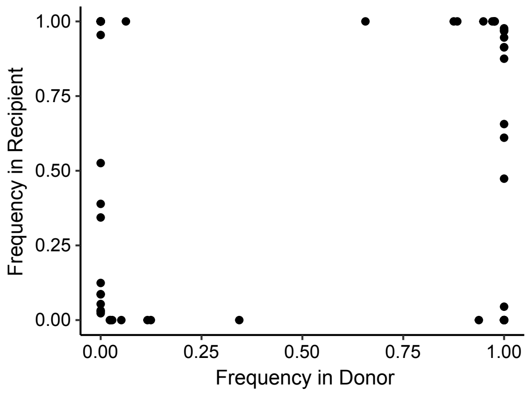


Figure 2.6. Shared diversity across household transmission pairs with influenza B virus. Intrahost SNV for 15 validated transmission pairs using samples closest to the time of transmission (inferred based on day of symptom onset). Each iSNV is plotted as a point with its frequency in the recipient (y-axis) versus its frequency in the donor (x-axis).

Discussion

Here we define the within-host genetic diversity of IBV in natural infections by sequencing 106 samples collected over 8176 person-seasons of observation in a household cohort. Because the HIVE study prospectively identifies individuals with acute respiratory illness regardless of severity, these samples capture IBV dynamics in a natural setting, reflective of infections occurring in the community. We show that within-host diversity of IBV is remarkably low, with most samples displaying no intrahost variants above our level of detection. We also find that IBV experiences a tight transmission bottleneck, limiting the diversity that is passed between hosts. IBV exhibits significantly lower within-host diversity compared to IAV. These findings reflect the slower relative evolutionary rate of IBV compared to IAV.

Our findings are largely consistent with what has been observed in IAV infections in humans (Debbink et al., 2017; Dinis et al., 2016; Leonard et al., 2016; McCrone et al., 2018). We found that only a minority of samples contain iSNV, the majority of which encode synonymous changes, consistent with a predominance of purifying selection within hosts. If immune-driven selective pressures were sufficiently strong to drive positive selection of antigenic variants at the individual level, we would expect to see enrichment of variants in antigenic regions. However, variants were no more common in the antigenic proteins, hemagglutinin and neuraminidase, and we found no intrahost variants in known antigenic regions of hemagglutinin. We also found that the extent of within-host diversity did not vary with current-season vaccination status, further suggesting that immune selection is not particularly strong within hosts (Debbink et al., 2017; Han et al., 2018; McCrone et al., 2018). Our data suggest that selective sweeps occur infrequently at the individual level, with selection only evident over a broader scale of time and space (McCrone et al., 2018; Nelson et al., 2006). We recognize, however, that it is possible for individual level selective pressure to vary in magnitude by age, locale, influenza infection history, or immune status (Lee et al., 2019).

We do find that there are important differences in the within-host evolution of IAV and IBV. IBV displays significantly lower within-host diversity compared to IAV. Since measurements of within-host diversity can vary based on host population, sequencing approach, and variant calling algorithm (Grubaugh et al., 2019a), a strength of our study is that our comparison is based on samples from the same cohort with the same sequencing approach and analytic pipeline. In both of our studies, we have sequenced swab samples directly without prior culture, accounted for the confounding effect of viral load, and used a standardized, empirically-validated analytic pipeline for variant identification (McCrone and Lauring, 2016). The only difference in methodology between these two studies is the multiplex amplification primers. In both viruses, these primers target the highly conserved ends of influenza virus genome segments, making it unlikely that this factor would drastically alter the amplification efficiency of within-host variants. Our analytic pipeline includes rigorous quality criteria to reduce false positives that can be introduced by amplification and Illumina sequencing. Importantly, these empirical quality criteria did not mask diversity actually present in these samples, strengthening the conclusion that IBV exhibits lower within-host diversity compared to IAV.

The most likely biological explanation for IBV's lower within-host diversity is its *de novo* mutation rate, which is thought to be at least two-fold lower than that of IAV (Nobusawa and Sato, 2006). Viral mutation rates are critical to the diversification of rapidly evolving viruses within hosts. Under a neutral model, the number and frequency of minority variants is dependent on the mutation rate and demographics of the population (Xue et al., 2018). In such a model, the expected number of variants is highly sensitive to variation in the mutation rate across the range commonly estimated in RNA viruses. In light of our results, a more thorough comparison of mutation rates across influenza viruses is needed.

Another possible factor underlying IBV's reduced diversity is the mutational robustness of the IBV genome relative to IAV. If IBV were less robust to mutation, stronger negative selection on multiple genes in IBV could result in more limited within-host diversity, perhaps located to certain regions of the genome. However, we found that the distributions of iSNV across IAV and IBV genomes are relatively similar. Furthermore, we have previously shown that the distribution of mutational fitness effects in influenza A/WSN/33/H1N1 matches that of other RNA and ssDNA viruses (Visher et al., 2016). Given that viruses across families with vastly different genomic architecture have similar mutational robustness, this is unlikely to account for the differences in within-host diversity between IAV and IBV.

We find that IBV experiences a stringent genetic bottleneck between hosts. A stringent transmission bottleneck places a constraint on the rate of adaptation of viral populations within and between individual hosts. Population bottlenecks reduce the effective population size, which increases random genetic drift and decreases the efficiency of selection (McCrone and Lauring, 2018). This results in a reduced ability of selection to fix beneficial mutations and to remove deleterious ones, which can decrease population fitness. However, there are potential evolutionary advantages to stringent bottlenecks, including removal of defective interfering particles (Vignuzzi and López, 2019; Zwart and Elena, 2015). While we were not able to estimate the size of the transmission bottleneck as precisely as IAV, it is likely that the bottleneck size is comparable across the two viruses given the similarities in their transmission routes and ecology in the human population. Data from many more transmission pairs will be necessary for a more robust estimate.

Together, our results are consistent with the slower rate of global evolution observed in IBV lineages compared with both seasonal A/H1N1 and A/H3N2 (Bedford et al., 2015; Chen and Holmes, 2008; Langat et al., 2017; Vijaykrishna et al., 2015). We suggest that a lower intrinsic mutation rate leads to reduced within-host diversity. With a comparably tight bottleneck, fewer *de novo* variants will rise to a level where they can be transmitted and spread through host populations. Combined with a lower incidence of IBV versus IAV, this would result in fewer variants that eventually spread and influence global dynamics. However, further investigation in larger populations will be required to evaluate the within-host dynamics of both types of seasonal influenza viruses and how they contribute to larger-scale evolutionary patterns.

Acknowledgements

We acknowledge the individuals enrolled in the HIVE study for their participation. We thank Maria Virgilio for technical assistance. This work was supported by NIH R01 AI 118886, NIH R21 AI 141832 and a Burroughs Wellcome Fund Investigator in the Pathogenesis of Infectious Diseases award to ASL. The HIVE cohort was supported by CDC U01 IP 001034 to ASM and ETM, and CDC U01 IP000474 and NIH R01 097150 to ASM. ALV was supported by NIH T32 GM 007863.

CHAPTER III

The Early Evolution of Oral Poliovirus Vaccine is Shaped by Strong Positive Selection and Tight Transmission Bottlenecks

Note: This chapter is a modified version of the published article:

Valesano AL, Taniuchi M, Fitzsimmons WJ, Islam MO, Ahmed T, Zaman K, Haque R, Wong W, Famulare M, Lauring AS. 2021. "The early evolution of oral poliovirus vaccine is shaped by strong positive selection and tight transmission bottlenecks." Cell Host & Microbe 29(1):32-43.

Elsevier grants all authors full rights to reuse their articles in dissertations.

Introduction

Genetic reversion and the associated loss of attenuation in oral poliovirus vaccine (OPV) strains are major barriers to achieving global poliovirus eradication (Kew et al., 2005). In areas of low vaccine coverage, OPV can evolve into circulating vaccine-derived polioviruses (cVDPV) that cause cases of poliomyelitis that are indistinguishable from those caused by wild polioviruses (WPV) (Jenkins et al., 2010; Kew et al., 2002; Pons-Salort et al., 2016). Of the three OPV serotypes, the Sabin type 2 is responsible for most cVDPV outbreaks (Burns et al., 2014; Kew and Pallansch, 2018). Following the eradication of wild type 2 polioviruses, the Global Polio Eradication Initiative switched routine immunization schedules from trivalent OPV (tOPV, Sabin types 1, 2, and 3) to bivalent OPV (Sabin types 1 and 3) to reduce the risk of future cVDPV2 outbreaks. However, monovalent type 2 OPV (mOPV2) is still used to combat cVDPV2 outbreaks. While the global replacement of tOPV with bOPV has reduced the presence of OPV2 in surveillance samples (Blake et al., 2018), cVDPV2 outbreaks remain a major problem. This is partly due to the continued reliance on monovalent Sabin type 2 OPV (mOPV2) to control new cVDPV2 outbreaks. Nearly half of the cVDPV2 outbreaks observed after the withdrawal of tOPV resulted from a previous mOPV2 intervention response (Macklin et al., 2020).

The emergence of cVDPV is a recurrent evolutionary process that exhibits a high degree of parallel, or convergent, evolution. Most data on cVDPV come from poliovirus isolates in cases of acute flaccid paralysis or environmental surveillance (Burns et al., 2013; Famulare et al., 2016; Shaw et al., 2018; Stern et al., 2017). A recent study of type 2 cVDPV sequences from multiple outbreaks in five countries identified a limited number of sites under positive selection across independent lineages (Stern et al., 2017). Three mutations – A481G, U2909C (VP1-I143T), and U398C – were inferred to be under the strongest selection pressure and precede subsequent substitutions. The A481G and U398C mutations are located in the 5' noncoding region and are functionally important to RNA structures in the internal ribosome entry site (IRES). All three mutations are known molecular determinants of attenuation, occur within the first two months after vaccination, and are associated with increased virulence in animal models (Famulare et al., 2016; Macadam et al., 1991, 1993; Muzychenko et al., 1991; Ren et al., 1991; Stern et al., 2017). For these reasons, they are referred to as "gatekeeper" mutations that initiate the process of attenuation loss. Although phylogenetic studies have provided important information on the evolutionary trajectories of cVDPV, they are limited in their ability to resolve the exact timing of gatekeeper mutations and may lack power to detect natural selection due to sampling bias (Geoghegan and Holmes, 2018). Isolates of cVDPV have undergone months or years of evolution prior to isolation and lack a definitive link to the time of vaccine administration, further limiting our understanding of the early evolution of OPV in humans.

Investigating the evolutionary dynamics within individual hosts can complement phylogenetic studies of virus evolution (Lauring, 2020). Complex evolutionary processes that take place within the span of a single infection cannot be resolved by standard consensus sequencing. Individual mutations, most often single nucleotide variants, arise within infected hosts and change in frequency according to the forces of natural selection and genetic drift. Studying how viruses evolve at this scale can uncover genomic sites under selective pressure, clarify the relative roles of selection and drift in viral evolution, and can inform sequence-based diagnostic and surveillance tools (Dolan et al., 2016; Holubar et al., 2019).

Various approaches have been used to study the molecular epidemiology of poliovirus and to monitor OPV stocks for reversion (Neverov and Chumakov, 2010; Sarcey et al., 2017), but few have been purposed for measuring viral diversity within naturally infected hosts. Routine surveillance for VDPV involves sequencing only the region encoding the capsid protein VP1 (Kilpatrick et al., 2011). Many approaches for whole genome sequencing rely on amplification of viral isolates in cell culture, which may not accurately preserve the diversity present in the original specimen (Montmayeur et al., 2017). Other high-throughput sequencing approaches that aim to measure within-host diversity have targeted only a specific portion of the poliovirus genome (Sahoo et al., 2017). While some have sequenced viral genomes or specific genomic regions from asymptomatic vaccine recipients (Boot et al., 2007; Dedepsidis et al., 2006; Sanden et al., 2009; Stern et al., 2017), we lack a comprehensive characterization of the early evolutionary dynamics of OPV within vaccinated individuals and during transmission to their close contacts.

Here we use whole genome, deep sequencing of stool samples from a clinical trial of OPV to elucidate the early evolution of polioviruses within and between human hosts. We developed an approach for sequencing OPV genomes directly from primary stool samples and validated its accuracy for identification of intrahost single nucleotide variants (iSNV). We applied this approach to samples from a recent trial that investigated the effect of tOPV cessation on the transmission of type 2 OPV (Taniuchi et al., 2017). The trial included a defined point of introduction of monovalent type 2 OPV (mOPV2) and weekly longitudinal sampling of vaccine recipients and their household contacts; it therefore represents an opportunity to investigate the early evolutionary dynamics of OPV2 in a community setting. We identify several mutations under strong positive selection, most of which are located in the capsid proteins and the 5' noncoding region. By comparing viral diversity across household transmission pairs, we find that mOPV2 experiences a narrow transmission bottleneck which may limit the spread of mutations that are strongly selected within hosts. These results connect the within-host selection of mutations with the dynamics of viral transmission and enhance our understanding of cVDPV evolution.

Methods

Clinical trial information and ethics

The clinical trial, including all aspects of sample collection and viral load measurements, is described in full in a prior publication (Taniuchi et al., 2017). The study was done according to the guidelines of the Declaration of Helsinki. The protocol for the clinical trial was approved by the Research Review Committee (RRC) and Ethical Review Committee (ERC) of the icddr,b and the Institutional Review Board of the University of Virginia. It is registered at ClinicalTrials.gov, number NCT02477046.

Sample collection and viral load quantification

Stool sample collection, nucleic acid extraction, and viral load measurements were performed and described in a prior publication (Taniuchi et al., 2017). Briefly, stool samples were collected, placed at 4˚C, and delivered to the icddr,b laboratory in Matlab within 6 hours. Samples were then aliquoted and stored at -80˚C until shipment on dry ice to the icddr,b laboratories in Dhaka. Total nucleic acid (TNA) from approximately 200 grams of stool was extracted with the QIAamp Fast DNA Stool mini kit and OPV was detected and quantified by RT-qPCR with serotype specific primers. TNA samples were shipped on dry ice to the University of Virginia and stored at -80 ˚C until processing for sequencing.

Primer design

We designed primers to amplify all three serotypes of OPV in overlapping amplicons covering the poliovirus genome. We used PrimerDesign-M (Yoon and Leitner, 2015) to determine sites of conservation across the three poliovirus types and identify potential primer sequences, allowing for ambiguous bases. We selected primers such that the four segments overlapped by at least 500 bp and manually curated each primer to have similar melting temperatures. We empirically tested various primer candidates for amplification on type 1 and type 2 poliovirus RNA templates. The primers used for genome amplification are listed in Table 3.2.

Amplification and sequencing

We amplified viral cDNA in four amplicons using a two-step RT-PCR protocol. We performed reverse transcription using the SuperScript III First-Strand Synthesis System (ThermoFisher). Each reaction contained 1.13 μL of 50 ng/μL random hexamers, 0.37 μL oligo-dT, 1.5 μL 10 mM dNTP mix, 12 μL of template total nucleic acid from stool, 3 μL of 10x RT Buffer, 6 μL 25 mM MgCl2, 3 μL 0.1M DTT, 1.5 μL RNase Out, and 1.5 μL of SuperScript III RT enzyme. The mixture of template, primer, and dNTPs was heated at 75˚C for 15 minutes to denature RNA secondary structure and placed directly on ice for > 1 minute. The enzyme and buffer mixture were then added on ice. The thermocycler protocol was: 25˚C for 10 min, 50˚C for 50 min, 85˚C for 5 min, and hold at 4˚C. The four overlapping segments were amplified by PCR with the primers listed in Table S1. The PCR reactions were as follows: 10 µL 5x HF Buffer, 1 µL 10 mM dNTP mix, 0.25 µL 100 µM forward primer, 0.25 µL 100 µM reverse primer, 33 µL nuclease free water, 0.5 µL of Phusion DNA Polymerase (NEB), and 5 µL of template cDNA. The thermocycler protocol was: 98˚C for 30 sec, 40 cycles of 98˚C for 10 sec, 59.5˚C for 30 sec, 72˚C for 2 min, then 72˚C for 2 min for final extension, and hold at 4˚C. The four segments for each sample were pooled in equal volumes (18 µL of each segment for a total pooled volume of 72 µL). Pooled amplicons were purified with Agencourt AMPure XP magnetic beads, using 1.8X volume of beads (129.6 µL of beads for 72 µL pooled PCR product). The sample was eluted into 40 µL of nuclease-free water. The purified PCR products were quantitated by Quant-iT PicoGreen dsDNA High Sensitivity Assay. A limited number of PCR products were spot-checked by gel electrophoresis. A plasmid control was prepared by applying the PCR protocol to a template of OPV2 in a plasmid. The sequence of the plasmid was determined by Sanger sequencing and was identical to the OPV2 GenBank reference (AY184220.1). One plasmid control was included in each pooled library, beginning at the library preparation stage, to account for sequencing errors and batch effects. Samples between 9 x 10^5 copies/gram and 4.5 x 10^7 copies/gram of OPV2 were amplified and sequenced in duplicate to improve the specificity of within-host variant identification. Libraries were prepared for Illumina sequencing with the Nextera DNA Flex Library Preparation kit according to the manufacturer's instructions, using Nextera DNA CD Indexes (96 samples). Eight pooled libraries were prepared in total and sequenced on an Illumina MiSeq (2x250 reads, v2 chemistry).

Benchmarking of variant identification

To determine the sensitivity and specificity of variant identification, we sequenced mock populations of a mixture of two viruses using the protocol described above. The viruses used were wild-type Mahoney type 1 poliovirus and the type 1 OPV strain, which differ by 66 mutations within the amplified regions. The consensus sequences of each viral stock were confirmed by Sanger sequencing. Viral RNA was extracted from each stock with the QIAamp Viral RNA Mini Kit (Qiagen) and viral RNA were mixed in equal concentrations at 0%, 1%, 2%, 5%, 10%, and 100% WT in OPV1. Virus mixtures were diluted to genome copy concentrations of 4.5 x 10^4 copies/μL, 9 x 10^3 copies/μL, 9 x 10^2 copies/μL, and 9 x 10^1 copies/μL (copies/gram is related to copies/μL by a factor of 10^3). To simulate the complex mixture of nucleic acid present in our samples, we performed the dilutions of viral populations in total nucleic acid extracted from stool from deidentified human donors (a gift of Pat Schloss, University of Michigan). The mixtures were then amplified by the protocol described in the section above (*Amplification and sequencing*). A plasmid control was generated from the OPV1 plasmid clone in the same way as described in the section above (*Amplification and sequencing*). The pooled library was generated using the Nextera DNA Flex Library Preparation Kit and sequenced on an Illumina MiSeq (2x250 reads, v2 chemistry), including the OPV1 plasmid control to account for batch effects and errors. To more carefully estimate the sensitivity of variant identification at various coverage levels, mapped reads were randomly down sampled to approximately 1000x, 500x, and 200x coverage evenly across the genome. Then within-host variants were identified using an analytic pipeline previously used for influenza viruses (McCrone and Lauring, 2016; McCrone et al., 2018), which depends on a clonal plasmid control to account for batch effects and local errors in Illumina sequencing.

Processing sequence data

Sequencing adapters were removed with cutadapt (Martin, 2011) and reads were aligned to all three OPV reference genomes (AY184220.1, AY184221.1, V01150.1) using bowtie2 (Langmead and Salzberg, 2012) with the –very-sensitive option. Duplicate reads were removed with Picard and samtools (Li et al., 2009). Consensus bases were identified at sites with 10x coverage or greater. For samples sequenced in duplicate, the replicate with the higher coverage at a given site was used to assign the consensus base. Each biological sample was assigned to a group based on depth and evenness of coverage across the OPV2 reference. Mean coverage was calculated across non-overlapping 50 bp bins across the genome region amplified by the four amplicon segments. Samples were considered variant-quality if they had an average coverage greater than 200x in every bin. Samples that had coverage of 10x or greater at every site were considered consensus-quality. We aligned the consensus sequences with the OPV2 reference with the MUSCLE algorithm (Edgar, 2004). For the dN/dS analysis, we used the PAML software version 4.8 (Yang, 2007). For gene-wise dN/dS analysis, we calculated a single value for omega across each gene with codeml using model M0. To identify sites with evidence of positive selection, we compared the likelihood of models M2 vs M1 with a chi-squared test and identified sites with omega greater than 1 with the Bayes empirical Bayes method.

Identification of within-host variants

We identified within-host variants in any 50 bp window with greater than 200x mean coverage, even if fewer than four segments were successfully amplified and sequenced. Within-host variants on the OPV2 genome were identified with the R package deepSNV (Gerstung et al., 2012), using the OPV2 plasmid control to account for sequencing errors and strand bias. Minor iSNV (< 50% frequency) in the cohort samples were filtered using the following criteria: deepSNV p-value < 0.01, average mapping quality > 20, average Phred score > 35, and average read position in the middle 75% of the read (positions 31 and 219 for 250 bp pair reads). For samples sequenced in duplicate, we only used variants identified in both samples; we assigned frequency using the sample that had higher coverage at the site. We only identified iSNV present at a frequency of > 5%. Sites that were monomorphic after applying these filter criteria were assigned a frequency of 100%. The analytic pipeline was used to determine the position of each base in the coding sequence of the viral polyprotein and assign it as synonymous or non-synonymous relative to the sample consensus.

To obtain haplotype information specifically for VP1-143, codon frequency was identified by finding all reads that spanned the codon, filtering by MapQ > 20 and Phred score of each base > 20, counting the number of reads within each codon, and dividing by the number of reads that passed the quality filters. Samples with quality read depth less than 150 were excluded.

Permutation test for parallel mutations

We quantified the probability that mutations would arise in parallel across a given number of individuals by implementing a permutation test. We first assumed that all sites were equally likely to mutate. We found the number of mutations that occurred in 83 individuals with variant-quality samples relative to the OPV2 reference above a frequency of 5%. We used this distribution to draw sites randomly across the genome, accounting for the length of the region amplified in our assay and excluding primer binding sites. We then found the number of sites shared by a given number of individuals. We ran this permutation 1000 times and calculated the p-value as the number of permutations with a number of shared sites equal to or greater than the observed data for a given group (e.g. mutations shared by two individuals, etc.). We simulated constraint on the mutability of genomic sites by restricting the fraction of sites available to mutate. We chose a fraction available of

60% to reflect the known distribution of fitness effects in poliovirus based on experimental data (Acevedo et al., 2014).

Estimation of the transmission bottleneck

Models for estimating the transmission bottleneck were implemented as described in our prior work on influenza virus (McCrone et al., 2018). In the presence-absence model, we assessed whether donor iSNV are found in the recipient. We assumed perfect detection of transmitted iSNV and that the probability of transmission of donor iSNV are determined by the measured frequency at the time of sampling. We modeled the probability of transmission as a binomial sampling process, depending on the donor iSNV frequency and the bottleneck size (N_b). We used maximum likelihood optimization to estimate the bottleneck size distribution, assuming bottlenecks across pairs follow a zero-truncated Poisson distribution. In the beta-binomial model, we relaxed the assumption of perfect detection of iSNV in the recipient by accounting for false-negative variant calls and stochastic loss below our detection threshold. We use our benchmarking data to supply the sensitivity of variant identification by frequency and titer, rounding down to the nearest titer threshold (e.g. 4.5 x 10^4 copies/μL, 9 x 10^3 copies/μL, etc.). We assume sensitivity in each range is the same as that of the titer threshold.

Data and code availability

The raw sequence reads for the Matlab samples and the benchmarking experiment are available on the NCBI Sequence Read Archive in BioProject PRJNA637613. Reads aligning to the human genome were filtered out by the SRA. The code for the primary analysis of within-host variants is publicly available at https://github.com/lauringlab/variant_pipeline. The rest of the code for data analysis and generation of the figures was written in R version 3.5.0 and python2.7 and is publicly available on GitHub at https://github.com/lauringlab/Poliovirus_Intrahost.

Quantification and statistical analysis

We performed various statistical tests on the data, all of which are described in the Results and Method Details. Unless otherwise noted, statistical tests were performed in R version 3.5.0. We used a multiple linear model to measure the effects of time since vaccination and viral load on specimen iSNV richness shown in Figure 3.2B (n = 101 specimens). We used a linear regression model to quantify the precision of iSNV frequency measurements across 11 variant-quality specimens sequenced in duplicate (Figure 3.7A). We applied a beta regression model (R package "betareg") with a logit link function to the mutation frequency data shown in Figure 3.3A and Figure 3.9B for three mutations. For samples with a frequency of 0 or 1, we adjusted their frequency by 10^{-7} in order to apply the beta regression model. For the dN/dS analyses, we used all complete consensus genomes from mOPV2 vaccine recipients (n = 157 genomes). For dN/dS calculations, we used PAML version 4.8 (Yang, 2007). For codon-specific dN/dS analyses, we used the Bayes empirical Bayes method (Yang et al., 2005). The permutation test for parallel mutations and the transmission bottleneck analyses are described in the Method Details.

Results

We used high depth of coverage sequencing on the Illumina platform to define the within-host diversity of mOPV2 in samples from vaccinated individuals and their household contacts (Taniuchi et al., 2017). These samples were collected as part of a cluster-randomized trial of OPV in the rural Matlab region, where the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) has conducted demographic and public health research since the 1960s (Alam et al., 2017). The trial assessed the impact of tOPV withdrawal on OPV community transmission by randomizing 67 villages to three different vaccination schedules: tOPV, bOPV followed by one dose of IPV, and bOPV followed by two doses of IPV. The trial then implemented a coordinated mOPV2 vaccination campaign over the course of one week, targeting 40% of children under 5 years of age. Shedding of OPV types 1-3 from 800 individuals across the three arms was monitored by quantitative RT-PCR of weekly stool samples. Transmission was measured by monitoring stool samples in household contacts of mOPV2 recipients. We selected 497 specimens from the vaccination campaign period for genome amplification and sequencing, prioritizing those with a stool viral load > 10^6 copies/gram.

Sample sequencing and assessment of genome coverage

We sequenced 416 samples from 219 mOPV2 recipients and 81 samples from 52 household contacts (Figure 3.1A). We amplified poliovirus genomes as overlapping RT-PCR amplicons using degenerate primers that recognize all three OPV serotypes (Table 3.2). We performed separate RT-PCR reactions for each segment and pooled them prior to library preparation (see Materials and Methods). Given that low viral titer influences the accuracy of within-host variant identification (McCrone and Lauring, 2016), we amplified and sequenced samples with an OPV2 viral load between 9 x 10^5 copies/gram and 4.5 x 10^7 copies/gram in duplicate. These cutoffs are based on the distribution of viral loads across the cohort and our empirically defined viral load cut-offs for influenza virus (McCrone et al., 2018). We amplified and sequenced several samples below 9 x 10^5 copies/gram that were collected from household contacts.

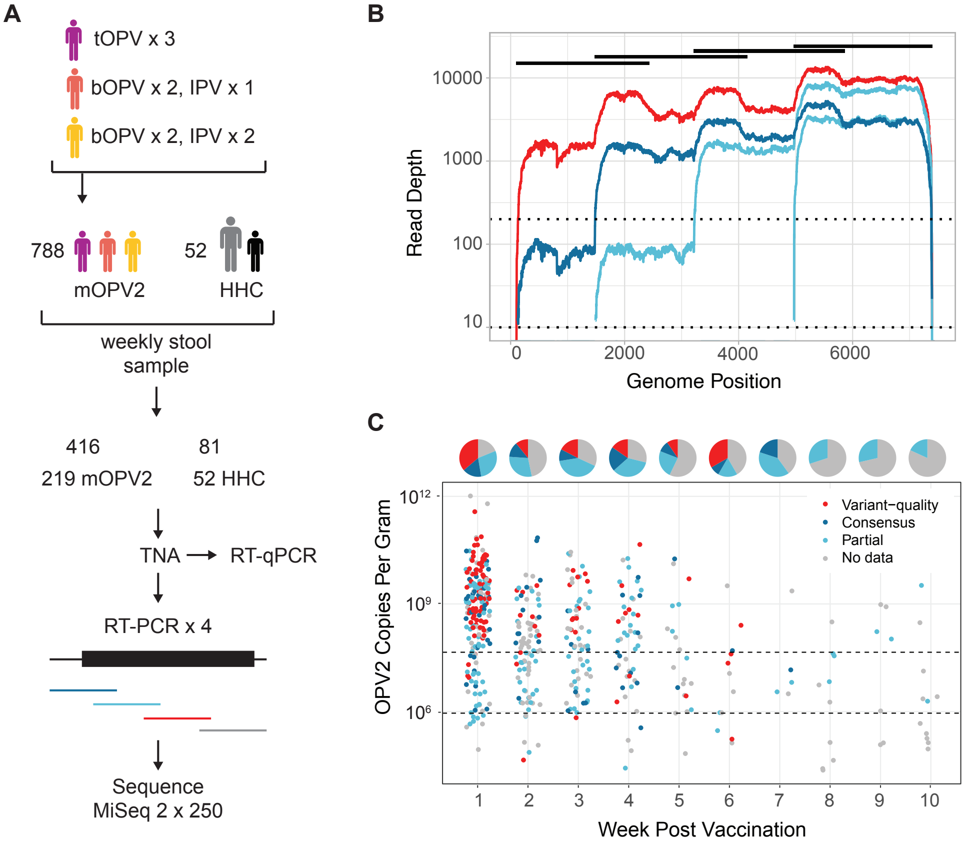


Figure 3.1. Overview of study and sequence data. (A) Schematic of study design and sample processing. The clinical trial had three arms with lead-up vaccination as indicated. tOPV = trivalent OPV, bOPV = bivalent OPV, IPV = inactivated polio vaccine. All individuals (n=788) then received mOPV2. Stool samples were collected weekly from mOPV2 recipients and household contacts (HHC). Only 52 household contacts had detectable shedding of OPV2. Total nucleic acid (TNA) was extracted from stools. Poliovirus genomes were amplified from each sample as four overlapping RT-PCR amplicons. For each sample, these amplicons were pooled and prepared for sequencing. (B) Line graph of sequencing coverage of four selected samples in three coverage groups. Log10 of coverage depth on the y-axis and genome position on the x-axis. One variant-quality sample shown in red, one consensus-quality sample shown in dark blue, and two partial-genome samples shown in light blue. Amplicons are shown as black bars (top). Dotted lines show cutoffs at 200x and 10x used for defining coverage groups. (C) Coverage groups of samples sequenced in this study. Each sample is shown as a point with OPV2 copies per gram of stool on the y-axis and week post-vaccination on the x-axis. Pie charts above each week indicate proportion of samples with variant-quality data (red), consensus quality data (dark blue), partial genome sequence data (light blue), no data (grey). Region in between the dotted lines shows the samples that were sequenced in duplicate.

Depth of coverage across the OPV2 genome for a given sample was uneven (Figure 3.1B). The 3' end generally exhibited coverage at least an order of magnitude greater than the 5' end, which contains the highly structured IRES (Lévêque and Semler, 2015). Three hundred twenty-seven samples had greater than 10x coverage of at least one of the four amplicons (Figure 3.1C, light blue points), and 179 samples had greater than 10x coverage across the whole genome (Figure 3.1C, dark blue points). We identified 111 samples with > 200x coverage (Figure 3.1C, red points), 81 samples with > 500x coverage, and 48 samples with > 1000x coverage across the genome, based on averages across a 50-bp sliding window. The majority of samples that yielded at least partial OPV2 genome coverage were collected in the first two months following vaccination (Figure 3.1C), which is consistent with the known shedding duration of Sabin type 2 (Famulare et al., 2018). Most individuals were represented by only one sample, although a subset of individuals had multiple longitudinal samples with at least partial genome data (Figure 3.7).

Empiric evaluation of variant calling criteria

We benchmarked the accuracy of our variant calling criteria for iSNV identification by sequencing defined mixtures of WPV1 (Mahoney strain) and OPV1 in stool-derived total nucleic acid (TNA) with viral genome concentrations ranging from 9 x 10^4 copies/gram to 4.5 x 10^7 copies/gram of stool. These concentrations were tailored to match those of the sequenced clinical samples. We then calculated the sensitivity and specificity of iSNV identification at various thresholds of input concentration, sequencing coverage, and iSNV frequency (Table 3.3). At a coverage depth of 200x, we reliably identified the expected single nucleotide variants at 5% frequency with 95% sensitivity at all genome copy inputs. However, at this coverage level, sensitivity was weaker for low frequency variants. We found that the number of false positives was low when the viral load was greater than 4.5 x 10^7 genome copies/gram. Specificity declined at viral loads below this cutoff, with a false positive rate of ~1% at all coverage levels. While some of these false positives can be filtered with various criteria, such as base and mapping quality, performing technical replicates proved to be the most effective approach for removing false positive variants. When considering only variants that were identified in both sequencing replicates, specificity dramatically improved even at low viral loads. Overall, these data validate our approach for poliovirus sequencing and demonstrate high sensitivity and specificity for variants above 5% frequency in the majority of sequenced samples. Therefore, we identified iSNV above a frequency threshold of > 5% in 111 samples that had > 200x coverage by sliding window across the genome, denoted here as "variant-quality." For analyses of variants at particular genome positions, we included samples with > 200x mean coverage in a 50-bp window containing the site of interest.

Signatures of selection

We characterized within-host diversity in 101 variant-quality samples from mOPV2 recipients. Minor iSNV (< 50% frequency) were dispersed across the genome in both the 5' noncoding region and the polyprotein, with greater numbers of variants at lower frequencies (Figure 3.2A). Each sample contained a median of 9 minor iSNV (IQR 6-15). There were more minor iSNV per sample with greater time since vaccination (Figure 3.2B). This association remained significant even after we controlled for the time-varying factor of viral load, which can affect iSNV identification ($p < 0.001$, multiple linear model). Our estimates of minor iSNV frequency were consistent when compared between technical replicates of 11 variant-quality samples (adjusted r-squared = 0.763, Figure 3.8A). While previous work has shown that the measured frequency of a variant can be affected by mutations in the primer binding sites (Grubaugh et al., 2019a), we were unable to distinguish this effect from the overall error in our frequency measurements. We identified a greater proportion of synonymous iSNV relative to nonsynonymous iSNV (ratio 0.43, Figure 3.2C). However, in the VP1 capsid subunit, there was an enrichment of nonsynonymous minor iSNV compared to other protein coding regions (Figure 3.2D). We calculated the dN/dS ratio with samples from mOPV2 recipients that had full consensus sequences across the polyprotein (n = 157). VP1 had the highest dN/dS ratio compared to the rest of the protein coding regions (Table 3.4).

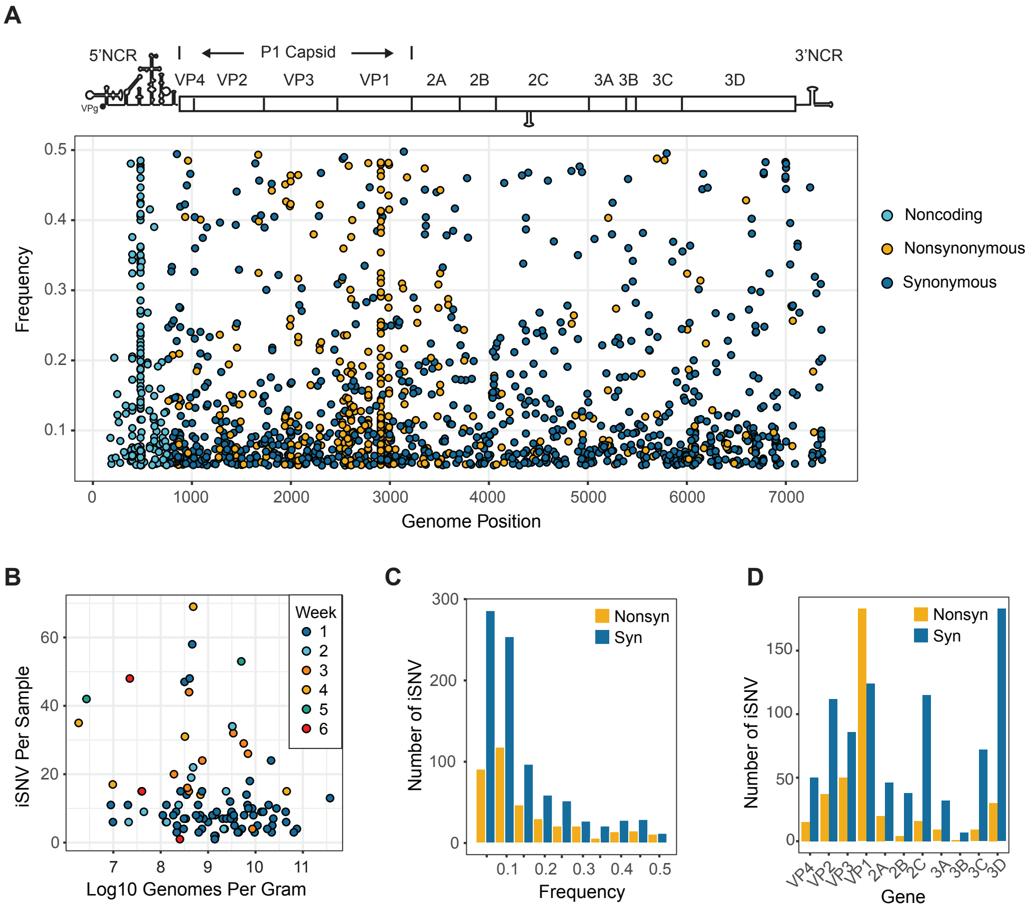


Figure 3.2. Within-host diversity in 101 variant-quality samples from mOPV2 vaccine recipients. (A) Minor iSNV shown as points, with frequency on the y-axis and genome position on the x-axis. Non-coding iSNV are shown in light blue, non-synonymous iSNV in yellow, and synonymous iSNV in dark blue. (B) Number of minor iSNV (y-axis) versus $\log_{10}$ of genome copies per gram of stool (x-axis). Color of each point is shown by the week post-vaccination of sample collection. (C) Histogram of minor iSNV in polyprotein by frequency with bin width of 0.05. Non-synonymous iSNV are shown in yellow, and synonymous iSNV in dark blue. (D) Histogram of minor iSNV by protein coding region in the polyprotein. Non-synonymous iSNV are shown in yellow, and synonymous iSNV in dark blue.

Positive selection of gatekeeper mutations

The dN/dS ratio is an imperfect metric for detecting selection, particularly within hosts, and it is unable to identify positive selection of mutations in noncoding regions (Kryazhimskiy and Plotkin, 2008). While changes in frequency of viral variants can be caused by multiple evolutionary forces, observing the same mutation arise in independent viral populations is suggestive of positive selection (Dolan et al., 2018; Gutierrez et al., 2019). We therefore analyzed the mutational dynamics at three positions that are major attenuating sites – positions 481 and 398 in the 5' noncoding region and codon 143 of VP1 (nucleotide positions 2908-2910). We used our time-series data to directly measure the frequency changes of the gatekeeper mutations in vaccine recipients. All three were present in several individuals within the first week of vaccination. A cross-sectional analysis of mutation frequency as a function of time demonstrated fixation of A481G within 2-3 weeks and U2909C in about 5 weeks, although there was substantial interindividual variability (Figure 3.3A and 3.3B). While A481G reached consensus in 11 of 14 samples by week 2, U2909C reached consensus in only 5 of 22 samples by week 3. VP1-143 most frequently reverted from isoleucine to threonine, but several other alternative residues were present (Figure 3.3C). Data from individuals with more than one sequenced sample demonstrated a rapid increase in frequency. Although mutation frequency occasionally decreased, presumably due to stochastic effects, these mutations increased in nearly all individuals (Figure 3.3B and 3.3D). We applied a beta regression model to estimate the time to fixation in the population. For mutations A481G, U2909C, and U398C, the model predicts a frequency of > 0.5 at weeks 2, 5, and 12, respectively, and > 0.95 by weeks 6, 13, and 46, respectively (Figure 3.9B). While we had more data from individuals in the bOPV/IPV study arms, each mutation rose in frequency over the same time interval regardless of vaccination history. This suggests that the selection for these mutations is not substantially driven by the presence of mucosal immunity generated by tOPV (Figure 3.9A). Overall, our time-series data on mOPV2 recipients shows rapid fixation of mutations at key attenuating sites in the first several weeks post-vaccination.

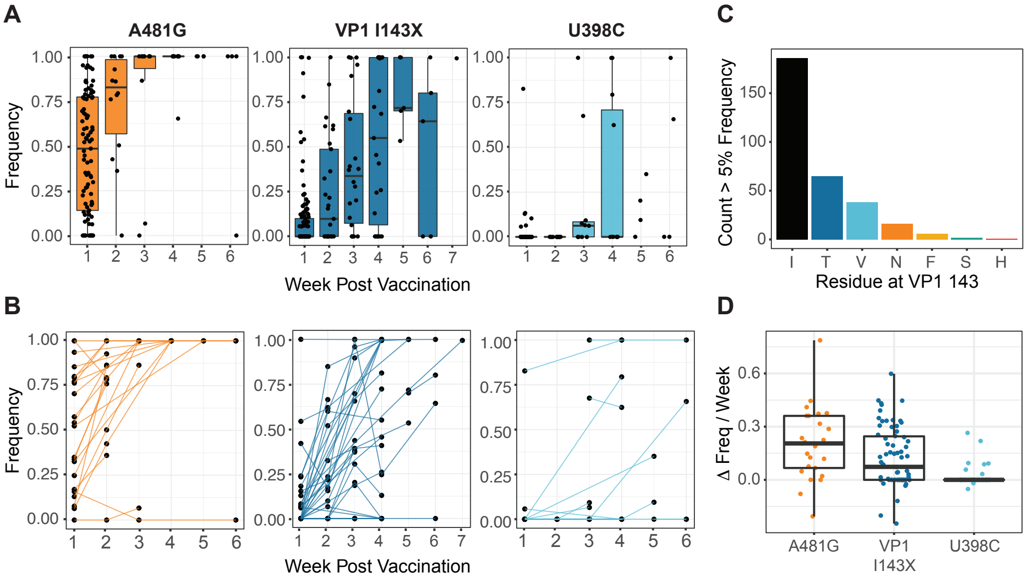


Figure 3.3. Selection of gatekeeper mutations in vaccine recipients. (A) Frequency of A481G, VP1-143X, and U398C by time from vaccination. Each point represents one sample, and boxplots are shown for weeks with five or more data points. Boxplots represent the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range. (B) Frequency of A481G, VP1-143X, and U398C by time from vaccination. Each point represents one sample, with lines connecting samples from the same individual. (C) Barplot showing number of samples with the indicated residues present at a frequency of 5% or above at VP1-143. (D) Change in frequency per week of three gatekeeper mutations prior to reaching fixation. Boxplots represent the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range.

Additional sites with positive selection

We next identified non-gatekeeper mutations that occurred independently across mOPV2 recipients. We restricted our analysis to 83 individuals who received mOPV2 and for whom we had at least one variant-quality sample. While most mutations were unique to a given viral population, a large number of mutations were present at ≥ 5% frequency in ≥ 2 individuals (Figure 3.4A). We performed a permutation test to quantitatively assess whether this distribution could occur due to chance alone. We drew random sites across the genome and tallied the number of sites shared across multiple individuals (see Methods). We observed more shared mutations than would be expected by chance for mutations in ≥ 3 individuals (Figure 3.4A). This result is robust to the assumption that all genome sites can be mutated; reducing the fraction of sites available for mutation did not affect statistical significance until the fraction dropped below 50% (Figure 3.10).

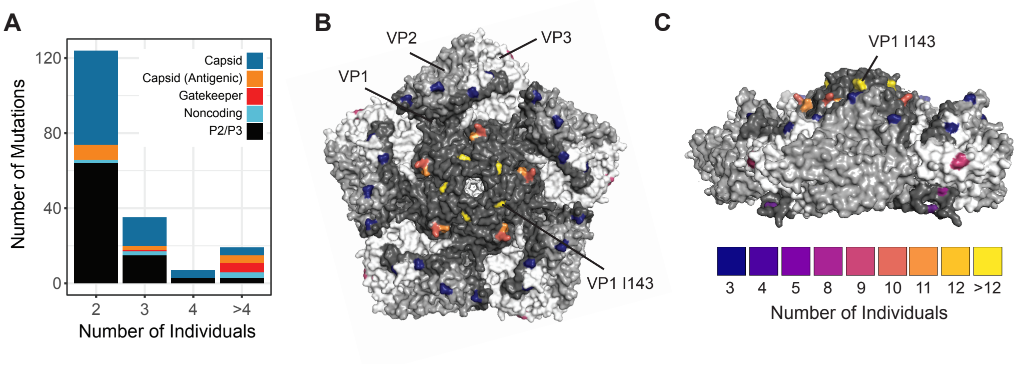


Figure 3.4. Mutations arising in multiple mOPV2 vaccine recipients. (A) Stacked barplot of the number of mutations identified (y-axis) by the number of individuals with each mutation (x-axis). Mutations in ≥ 3 individuals were statistically significant by permutation test, see text. Colors show the category of each mutation. (B) Structure of type 2 poliovirus capsid pentamer (PDB: 1EAH) and side view (C). Highlighted residues are color-coded by number of mOPV2 vaccine recipients with non-synonymous substitutions at that amino acid site.

Excluding the gatekeeper mutations, we found 19 mutations that were present in ≥ 4 individuals (Table 3.1). Two mutations, G491A and G619U, were located in the IRES. Five mutations were located outside the capsid (P1) in polyprotein regions P2 and P3. The capsid proteins (VP1-4) were highly represented (12 of 19 mutations), with four mutations encoding nonsynonymous mutations in known antigenic sites. Mutation A2986G encodes a nonsynonymous substitution, VP1-K169E, in antigenic region NAg1. Two more mutations in NAg1, G2782A and C2783A, encode nonsynonymous changes at VP1-101 (A101T and A101D, respectively). Mutation A1997G, found in 8 of 83 individuals, encodes VP3-H77R in antigenic region NAg3b. This mutation was identified in a previous phylogenetic study as having intermediate evidence for positive selection across cVDPV lineages (Stern et al., 2017). Our host-level data show that this mutation is indeed under strong positive selection. We did not detect U2523C, C2006A, U1376A, and U3320A, which are predicted to occur > 2 months after vaccination (Stern et al., 2017).

Table 3.1. Mutations identified in multiple individuals.

Mutation[a]	Individuals[b]	Group	Type[c]	Region	Fraction of cVDPV[d]
A481G	72	Gatekeeper	Noncoding	5' UTR	1
U2909C	25	Gatekeeper	NS	VP1	0.75
A2908G	19	Gatekeeper	NS	VP1	0.03
U398C	16	Gatekeeper	Noncoding	5' UTR	0.94
A2074G	12	Capsid	NS	VP3	0.01
A2992G	11	Capsid	NS	VP1	0.03
A2986G	10	Antigenic	NS	VP1	0.01
G6084U	10	3D	S	3D	0.03
A1997G	8	Antigenic	NS	VP3	0.05
U2909A	8	Gatekeeper	NS	VP1	0.03
U882C	8	Capsid	S	VP4	0.02
G2782A	7	Antigenic	NS	VP1	0.01
G491A	7	Noncoding	Noncoding	5' UTR	0.14
G619U	7	Noncoding	Noncoding	5' UTR	0
C2609U	6	Capsid	NS	VP1	0.01
C2783A	5	Antigenic	NS	VP1	0
U4374C	5	2C	S	2C	0.55
A3490G	4	2A	NS	2A	0.02
C2291U	4	Capsid	NS	VP3	0.01
C2580U	4	Capsid	S	VP1	0.1
G1282A	4	Capsid	NS	VP2	0
U1641C	4	Capsid	S	VP2	0.68
U5811A	4	3C	S	3C	0.14
U6693A	4	3D	S	3D	0.31

[a] Mutation presented as base in OPV2, position in OPV2 reference genome, and base in samples.
[b] Number of individuals with each mutation present at a frequency of 5% or greater. The total number of individuals analyzed is 83.
[c] Nonsynonymous (NS) or synonymous (S) mutations relative to the OPV2 reference genome.
[d] Fraction of cVDPV genomes with each mutation.

We also examined independent capsid mutations at the codon level, such that distinct mutations at the same amino acid site were included. We identified 19 amino acid sites at which three or more individuals exhibited nonsynonymous substitutions at a frequency of > 5%. Many of these sites mapped to the surface of the type 2 capsid (Figure 3.4B). In addition to the amino acid sites specified above, this analysis revealed four more antigenic sites with parallel non-synonymous substitutions: VP3-58 (T58I and T58A) and VP1-T291A in NAg3a, VP2-N72D in NAg3b, and VP1-S222P in NAg2. Although dN/dS analysis is often not sensitive enough to identify positive selection at specific sites over short time scales, amino acid sites VP1-143 and VP1-101 had a dN/dS ratio greater than 1 based on analysis of consensus genomes (Pr(ω > 1) > 0.95, Bayes empirical Bayes method, Table 3.4). Together, our results demonstrate rapid positive selection of mutations in the IRES and exposed sites in the OPV2 capsid.

We sought to determine whether mutations selected early in OPV2 evolution persist in genomes of neurovirulent cVDPV. We queried alignments of cVDPV from cases of acute flaccid paralysis for the presence of the mutations identified here (Table 3.1). As expected, the gatekeeper mutations were reliably detected in nearly all cVDPV genomes. Of the other 19 mutations, we found 16 in at least 1% of cVDPV genomes queried. We detected some mutations, including G491A, C2580U, and U1641C, in at least 10% of cVDPV genomes. While it is unknown how these mutations may contribute to the development of cVDPV outbreaks, our data show that mutations that recur in divergent cVDPV lineages can be identified very early in OPV2 evolution.

Estimation of the transmission bottleneck

Transmission bottlenecks influence the rate of adaptation and are important for understanding viral evolution in host populations (Elena et al., 2001). The OPV vaccine trial included stool samples collected from the household contacts of mOPV2 recipients. Shedding in these individuals allowed us to analyze the extent of viral diversity that is transmitted to new hosts. We identified four transmission pairs for which we had sequence data from both the donor and recipient collected within one week of each other (Table 3.5); sequencing of more putative household pairs was thwarted by low viral loads, especially in household contacts. In each case, transmission occurred within the first three weeks after vaccination, consistent with the known magnitude and duration of

OPV shedding. We obtained variant-quality samples from donor and recipient of one pair; the rest had either consensus or partial genomes with regions of variant-quality coverage.

We compared within-host diversity across donors and recipients using only the genomic regions that had a depth of coverage sufficient for identification of within-host variants in both samples. There were few polymorphic sites shared across hosts in these household pairs (Figure 3.5A). While major variants (> 50% frequency) in the donor were usually found in the recipient, most minor variants were not found in the recipient, suggestive of a narrow transmission bottleneck.

We applied two models to quantify the effective genetic bottleneck at transmission (Leonard et al., 2017; McCrone et al., 2018). We use the term "effective" bottleneck to clarify that we are capturing mutations that transmit and persist in the population of the recipient host. The presence-absence model asks whether a polymorphism in the donor is present or absent in the recipient, assuming perfect detection. Here, transmission is modeled as a random sampling process in which the probability of transmission is a function of the mutation frequency in the donor and the size of the transmission bottleneck. We used maximum likelihood optimization to find the bottleneck size distribution that best fit the data, assuming that the bottlenecks follow a zero-truncated Poisson distribution. Under the presence-absence model, the mean bottleneck size was 1.98 (lambda = 1.57, 95% confidence interval 0.43 – 3.63), indicating that most bottlenecks are 2 and 95% of bottlenecks are less than 4 (Figure 3.5B). We also applied a beta-binomial model, which incorporates the sensitivity of detecting variants in the recipient and allows for time-dependent stochastic loss of variants. The beta-binomial model yielded a mean bottleneck size of 2.11 (lambda = 1.74). The model fit was not significantly better than the presence-absence model (AIC 42.1 for presence-absence vs. 39.7 for beta-binomial), indicating that the loss of sensitivity might not be an important factor in these models. We also estimated bottlenecks for each pair individually with both models (Table 3.6). The two models produced the same estimates for each pair, although the beta-binomial model resulted in slightly larger confidence intervals. These results suggest that few genetically distinct OPV2 genomes transmit and persist in new hosts.

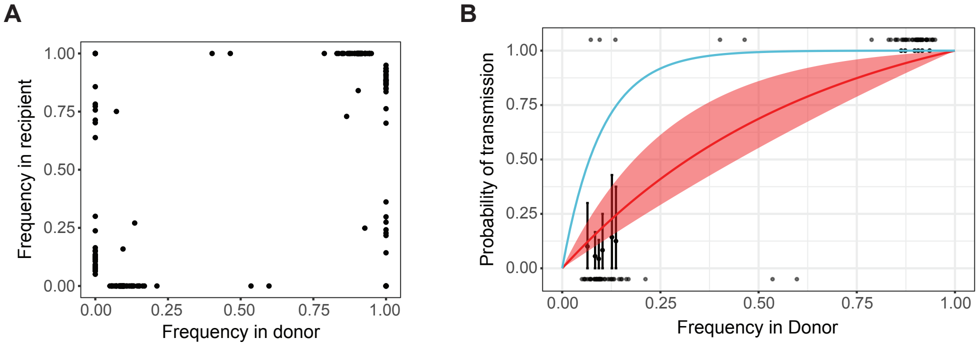


Figure 3.5. Shared viral diversity across transmission pairs and transmission bottleneck. (A) iSNV for four pairs of mOPV2 recipients and their household contacts. Each iSNV is plotted as a point with its frequency in the recipient (y-axis) versus its frequency in the donor (x-axis). (B) Presence-absence bottleneck model fit compared with data. Frequency of donor iSNV on the x-axis and probability of transmission on the y-axis. Transmitted iSNV are shown along the top of the plot and non-transmitted iSNV are shown along the bottom. The red line shows the probability of transmission as a function of donor frequency given the mean bottleneck estimate, with a 95% confidence interval shown by the shaded area. The blue line shows the probability of transmission given a bottleneck size of 10 unique genomes. The black points on the graph represent the probability of transmission from the measured iSNV using a sliding window of 3% width and a step size of 1.5%.

These models assume that minor variants are transmitted independently, but variants can potentially be linked within hosts, which could result in an inflated bottleneck estimate. Because we used amplicon sequencing with short reads, we are unable to evaluate and exclude mutation linkage across the entire genome. However, we were able to evaluate linkage among pairs of intrahost variants that were near enough to be spanned by individual reads. These variant pairs were largely independent (Figure 3.11). Given these results and the fact that our bottleneck estimate was already low, the assumption of independent variant transmission is likely reasonable in this context.

A tight transmission bottleneck limits the spread of gatekeeper mutations

We sought to investigate whether a narrow transmission bottleneck would impact the transmission of mutations that are positively selected within vaccine recipients. Based on the estimated bottleneck size, we calculated the probability of transmission of each of the three gatekeeper mutations as a function of time, using the median frequency from mOPV2 recipients (Figure 3.6A). We then calculated the fraction of samples from transmission recipients that had each mutation present, regardless of whether we had sequence data from a donor population (Figure 3.6B). The fraction of transmission samples with each mutation is consistent with the calculated probability of transmission over time given the size of the bottleneck. We identified A481G in most transmission samples, consistent with its rapid fixation. However, few samples contained U2909C and no samples contained U398C. The majority of transmission events occurred within the first two weeks after the vaccination campaign, prior to when most vaccine recipients acquired U2909C and U398C (Taniuchi et al., 2017). These data suggest that the three gatekeeper mutations were not preferentially transmitted; instead, they suggest that mutations must rise to an appreciable frequency early enough within a donor population to be frequently transmitted through a narrow bottleneck.

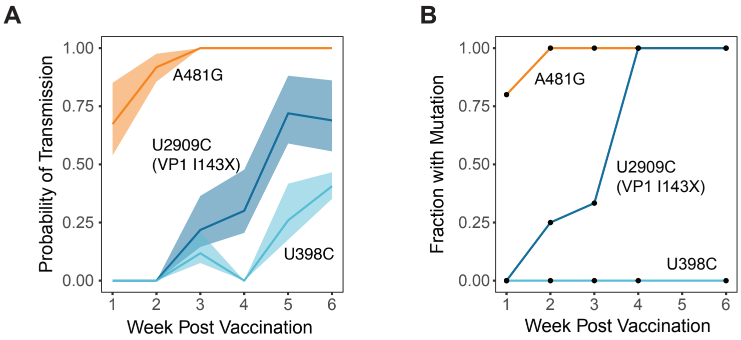


Figure 3.6. Impact of a tight bottleneck on transmission of gatekeeper mutations. (A) The probability of transmission of each gatekeeper mutation calculated from the median frequency over time in the mOPV2 recipients given the estimated bottleneck. The shaded areas represent 95% confidence intervals based on the model fit. (B) The fraction of each gatekeeper mutation present above a frequency of 5% in samples from household contacts as a function of time since the vaccination campaign.

Discussion

We used whole genome deep sequencing to define the within-host evolutionary dynamics of OPV2 in a clinical trial in Matlab, Bangladesh (Taniuchi et al., 2017). The trial enabled analyses of longitudinal samples from a defined and synchronized point of mOPV2 vaccination and household transmission in a community with high enteric pathogen burden and vaccine coverage. These results provide a rare window into the evolutionary dynamics that occur in the first weeks following vaccination. Similar to other RNA viruses, we identified strong purifying selection across the poliovirus genome within hosts (McCrone et al., 2018). However, in stark contrast to other viruses, we found evidence for strong within-host positive selection at multiple sites. Although high population immunity in Bangladesh limited the number of transmission samples available from the trial, we were able to quantify the transmission of key reversion mutations and estimate a tight bottleneck in this setting. Our findings enhance our knowledge on the within-host and transmission dynamics of polioviruses in relation to the development of cVDPV.

We find that positive selection is remarkably strong within vaccine recipients, with a magnitude that is seldom found in the within-host evolution of acute RNA viruses. We and others have rarely identified strong selection for mutations at the within-host level, even for mutations that should have beneficial effects (Debbink et al., 2017; Dinis et al., 2016). The within-host evolution of several arboviruses is characterized by purifying selection and a large effect of stochastic genetic drift (Lequime et al., 2016; Parameswaran et al., 2012). In household cohort studies of influenza virus infection, we have found little evidence for positive selection within the span of a single infection (McCrone et al., 2018; Valesano et al., 2020a), and iSNV are rarely observed in more than one individual. In contrast, in this cohort we identified 24 mutations that were identified in ≥ 4 individuals. Although comparisons to other viruses are complicated by differences in duration of infection, genome structure, and other factors, the extent of parallel evolution in OPV at this scale is remarkable. Whereas wild polioviruses and other endemic RNA viruses may already exist near local fitness peaks, OPV is significantly attenuated and is under intense pressure to climb the fitness landscape by accessing available high-impact mutations (Stern et al., 2017). OPV is also unique in that each population starts from the same founder genetic sequence, making parallel trajectories more likely to occur (Gutierrez et al., 2019).

Outside of the three gatekeeper mutations, we find that there are multiple additional sites under selection early in OPV2 evolution that reflect re-adaptation to the human host. There are several potential reasons why our study revealed mutations that have not been previously identified in cell culture or phylogenetic studies. While cell culture and animal models can be a helpful proxy for inferring selective pressures (Geoghegan and Holmes, 2018), they do not always capture the direction and magnitude of evolutionary forces in natural hosts. Similarly, phylogenetic studies have yielded important insights into the evolution and epidemiology of cVDPV, but may not be able to infer selective advantage of mutations with weaker effects due to limitations in sampling or statistical power. Whereas, previous phylogenetic work on cVDPV2 found only limited evidence for positive selection at VP3-77 and VP1-222, they appear to be strongly selected within hosts (Shaw et al., 2018; Stern et al., 2017). Finally, it is important to recognize that phylogenetic studies may differ in the fitness effects they reveal due to differences in time and scale. Our results primarily reflect within-host selection observed through parallel evolution among vaccine recipients sampled longitudinally whereas previous phylogenetic analyses are based on shared variation among surveillance samples collected from many people over time and linked by sustained transmission. While we found that 16 of the 19 positively-selected, non-gatekeeper variants recur in cVDPV lineages, they do not routinely fix. Some mutations might be repeatedly selected within hosts and prove to be detrimental between them.

The underlying selective pressures and consequences of these mutations are unclear, but their locations suggest functional significance. Mutations in the IRES have been shown to affect protein translation and replicative capacity (Avanzino et al., 2018). Mutations in the capsid enable adaptation to replication at physiologic temperatures, and could provide increased structural stability or modulate receptor binding and viral entry (Macadam et al., 1991; Robinson et al., 2014). Although some of the capsid sites identified here are recognized by neutralizing antibodies (Patel et al., 1993; Shaw et al., 2018), there is little evidence that these mutations lead to antigenic escape, as they do in influenza virus or HIV (Hedestam et al., 2008). Even highly diverged cVDPV strains with significant antigenic evolution are still neutralized by serum from vaccinated individuals (Shaw et al., 2018), and vaccination with OPV is used to control outbreaks of cVDPV (Kew and Pallansch, 2018). Rather than antigenic escape, we suggest that these mutations lead to improved within-host replication, and therefore greater shedding and transmission. Epidemiologic data suggest that at some unknown point in cVDPV evolution, OPV achieves a level of transmissibility that is similar to that of wild polioviruses (Famulare et al., 2018; Jenkins et al., 2010; Tebbens et al., 2013). Selection for phenotypes related to this increase in transmissibility, like enteric replication and shedding, are likely the earliest pressures the virus faces (Bull et al., 2018). Not all capsid antigenic sites may be involved in this process. There can be frequent amino acid substitution at many antigenic sites in the OPV2 capsid, often reverting back to previous replacements, suggesting that some sites are more tolerant to mutation and evolve more by genetic drift than selection (Shaw et al., 2018). However, our results indicate that a subset of these capsid sites experience positive selection and likely have functional effects related to improved replication and transmission within the human host.

Our identification of specific sites under positive selection has implications for genetic surveillance of VDPV. In VDPV isolates, the time since vaccine administration is estimated by molecular clock methods on VP1 sequence data (Jorba et al., 2008). For OPV2, the threshold for calling a strain a VDPV – as opposed to OPV-like – is 0.6% divergence, or ≥ 6 nucleotide substitutions (Wassilak et al., 2011). Prior work has integrated the fixation rates of gatekeeper mutations into molecular clock models to refine estimates of the time between VDPV detection and initial vaccination (Famulare et al., 2016). By accounting for these rapidly selected mutations, the authors inferred that type 2 VDPVs are younger than estimated based on neutral evolution alone. Here, we used our longitudinal data to determine the fine-scale dynamics of these three gatekeeper mutations and to characterize the variability that can manifest at the individual scale. Fixation rate estimates that are grounded in direct measurements are relevant for modeling efforts that rely on these parameters. In addition, we suggest that molecular clock models of VDPV might benefit from incorporation of the sites identified in this work in two ways. First, the rates of selection at these sites could be integrated into existing models of time since vaccine dosing for VDPVs. Second, these sites under putative selection could be excluded from neutral molecular clocks for VDPV divergence time estimation.

A surprising and important finding is that a tight bottleneck (1-4 distinct genomes) limited the transmission of within-host variants to new hosts. It is certainly possible that loss of variants in the recipient population could result in an under-estimation of the bottleneck. In this study, we were limited by the week-long interval between sample collection, which means that transmission could have occurred several days prior to sampling from the household contact. We also used a conservative frequency threshold of 5%, which may miss transmission of variants that remain at low frequencies across both hosts. However, it is unlikely that variants below 5% would be consistently transmitted while variants from 5-50% are not. Furthermore, the results of the beta-binomial model suggest that imperfect detection and stochastic loss did not have a large influence on the bottleneck estimate.

In support of the finding of a small bottleneck in this transmission setting, we have no evidence for a between-host advantage of the gatekeeper mutations despite strong evidence of within-host selection. If genomes with gatekeeper mutations are transmitted preferentially, we would have expected to see them present in more household contacts and at higher frequencies than observed here. The sample size in this study limits our ability to detect small effects, but the data generally do not support substantial preferential transmission in this study population. A narrow, non-selective bottleneck can explain the pattern of transmission to household contacts. In this scenario, mutations selected within vaccine recipients must rise above a threshold frequency prior to transmission in order to transit a narrow bottleneck. This would limit spread of minor iSNV that have not been selected quickly enough before finding a new host.

Of course, the gatekeeper mutations eventually fix in nearly all vaccine-derived lineages and can arise *de novo* in each subsequent host (Famulare et al., 2016; Stern et al., 2017). Population immunity and differences in fecal-oral exposure between Matlab, which has never experienced a cVDPV outbreak, and other settings where cVDPV outbreaks are more common may lead to important selective differences not observed here. In contrast to the lack of between-host selective effects in this study, late transmission events at the end of the duration of shedding when variant fractions are high, may have a larger impact on the spread of positively selected mutations. Transmission later in infections when viral load is low and to more distant community contacts was uncommon in this highly immune population (Taniuchi et al., 2017). Furthermore, bottlenecks may be larger in populations with lower background immunity and higher fecal-oral exposure where naturally acquired doses are likely higher (Famulare et al., 2018), which would allow positive selection among minor variants to act.

It is likely that newly developed live-attenuated polio vaccines will face the same underlying selection pressures as mOPV2. Our results suggest that once a beneficial mutation occurs on a highly attenuated OPV background, there is strong selection to drive it to fixation. Strategies that decrease the fitness benefit of any single mutation, like modifications to IRES domain V and codon deoptimization, may be effective at sufficiently prolonging the time to reversion (Konopka-Anstadt et al., 2020; Yeh et al., 2020). However, modest decreases in mutation rate by introduction of high-fidelity RNA-dependent RNA polymerase modifications might have lesser impact, as the mutation rate is still orders of magnitude higher than in other organisms (Sanjuán et al., 2010). In the setting of a virus starting from low fitness with a high mutation rate, whether a mutation achieves fixation or not may be more dependent on size of the fitness benefit rather than the waiting time for *de novo* generation of the mutation. This effect is illustrated by one next generation OPV2 (nOPV2) design, which prevents A481G by modification of IRES domain V. In individuals receiving this nOPV2, VP1-143 and U398C still readily revert despite a high-fidelity 3D^{pol} (Yeh et al., 2020). High-fidelity polymerase modifications themselves may not be stable, as seen by the reversion and compensation of a type 1 poliovirus fidelity mutant due to a fitness defect in cell culture (Fitzsimmons et al., 2018). Our results also suggest that there are mutations other than the three gatekeepers that increase fitness and contribute to reversion. Monitoring the genetic changes of new vaccine designs in sufficiently large cohorts will be important for evaluation of the genetic stability at these additional sites.

Acknowledgements

We thank the participants and field workers in the original vaccine trial for their many contributions. We acknowledge the following individuals for their roles in implementation of the vaccine trial and sharing of samples: Dr. Mohammed Yunus, M.B.B.S. (Division of Infectious Diseases, icddr,b, Dhaka, Bangladesh), and Dr. William A. Petri, M.D./Ph.D (Division of Infectious Diseases and International Health, University of Virginia, Charlottesville, VA). We thank Suzanne Stroup for technical support. We thank Andrew Macadam, PhD and Cara Burns, PhD for providing the OPV1 and OPV2 plasmids, respectively. We thank JT McCrone for helpful discussion on data analysis. This work was funded by a grant from the Bill and Melinda Gates Foundation (to MF and MT) and a Burroughs Wellcome Fund Investigator in the Pathogenesis of Infectious Diseases Award (to ASL).

Supplemental Figures and Tables

Table 3.2. Genome amplification primers used in this study.

Name	Sequence
PanSabin_Seg1_Fwd	5'-CCCGYAACTTAGAMGCA-3'
PanSabin_Seg1_Rev	5'-CTGACACAAAMCCMAGSATG-3'
PanSabin_Seg2_Fwd	5'-TCTGCCCRGTKGATTAYCTC-3'
PanSabin_Seg2_Rev	5'-TCAGTRAATTTYTTCAACCAACT-3'
PanSabin_Seg3_Fwd	5'-GTMAATGATCACAACCC-3'
PanSabin_Seg3_Rev	5'-GTTGGAAAGTTGTACATTAG-3'
PanSabin_Seg4_Fwd	5'-TGTCCTTTAGTGTGTGG-3'
PanSabin_Seg4_Rev	5'-CCCAATCCAATTCGACTG-3'

Table 3.3. Validation of within-host variant identification by sequencing mock populations.

		(1) replicate, 4.5 x 10^4 copies/µL[a]			(2) replicates, 9 x 10^3 copies/µL[a]		
Coverage	Frequency	Sensitivity	Specificity	FP[b]	Sensitivity	Specificity	FP[b]
200x	10%	1	1	0	1	1	0
	5%	1	1	0	0.94	1	0
	2%	0.6	1	0	0.49	1	0
	1%	0.17	1	0	0.06	1	0
500x	10%	1	1	0	1	1	0
	5%	1	1	0	1	1	0
	2%	0.91	1	0	0.74	1	0
	1%	0.54	1	0	0.4	1	0
1000x	10%	1	0.9999	1	1	1	0
	5%	1	1	0	1	1	0
	2%	1	1	0	0.97	0.9999	1
	1%	0.91	1	0	0.69	1	0

		(2) replicates, 9 x 10^2 copies/µL[a]			(1) replicate, 9 x 10^2 copies/µL[a]		
Coverage	Frequency	Sensitivity	Specificity	FP[b]	Sensitivity	Specificity	FP[b]
200x	10%	0.89	1	0	1	0.9995	7
	5%	0.83	1	0	0.80	0.9994	9
	2%	0.4	1	0	0.46	0.9990	14
	1%	0	1	0	0.31	0.9997	4
500x	10%	0.97	1	0	1	0.9991	13
	5%	0.97	1	0	0.97	0.9985	21
	2%	0.49	1	0	0.51	0.9984	23
	1%	0.03	1	0	0.51	0.9992	12
1000x	10%	1	1	0	1	0.9987	19
	5%	1	1	0	0.97	0.9974	37
	2%	0.63	1	0	0.91	0.9985	22
	1%	0.03	0.9999	2	0.03	0.9989	16

[a] Copies/µL is 1000-fold lower than copies/gram of stool.

[b] Number of identified false positives.

Table 3.4. Gene-wise estimates of dN/dS ratio.

Gene	Omega (dN/dS)
VP4	0.26326
VP2	0.05951
VP3	0.44309
VP1	1.20709
2A	0.25478
2B	0.00010
2C	0.05137
3A	0.09244
3B	0.00010
3C	0.10622
3D	0.04747

Table 3.5. Samples from transmission pairs used in bottleneck analysis.

Donor ID[a]	Recipient ID[a]	Donor Vaccination Date[b]	Donor Sample Date[c]	Recipient Sample Date[d]	Time Difference (days)
115	10115	2016-01-26	2016-02-02	2016-02-01	1
171	10171	2016-01-25	2016-01-31	2016-01-31	0
702	20702	2016-01-25	2016-02-08	2016-02-08	0
927	10927	2016-01-28	2016-02-24	2016-02-17	7

[a] Anonymous IDs per individual.
[b] Date of mOPV2 administration in trial vaccination campaign.
[c] Date of sample collection from mOPV2 recipient used in the bottleneck analysis.
[b] Date of sample collection from the household contact used in the bottleneck analysis. For each recipient, this is the first longitudinal sample positive for OPV2 by RT-PCR.

Table 3.6. Transmission bottleneck estimates for two models.

Pair ID	Presence-absence model estimate[a]	Beta-binomial model estimate[a]
115	1 (1 – 2)	1 (1 – 3)
171	2 (2 – 4)	2 (2 – 7)
702	2 (2 – 2)	2 (2 – 3)
927	2 (2 – 5)	2 (2 – 4)

[a] 95% confidence interval shown in parentheses.

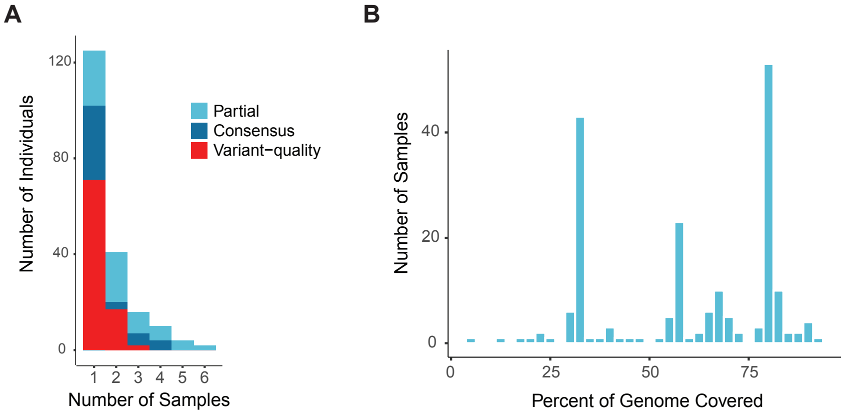


Figure 3.7. Sequencing coverage. (A) Overlapping bar chart of the number of individuals (y-axis) by the number of samples sequenced from a given individual (x-axis). Colors represent the genome coverage groups shown in Figure 1. (B) Composition of the partial genome samples. Number of samples (y-axis) by the percent of the genome covered above a 10x threshold (x-axis).

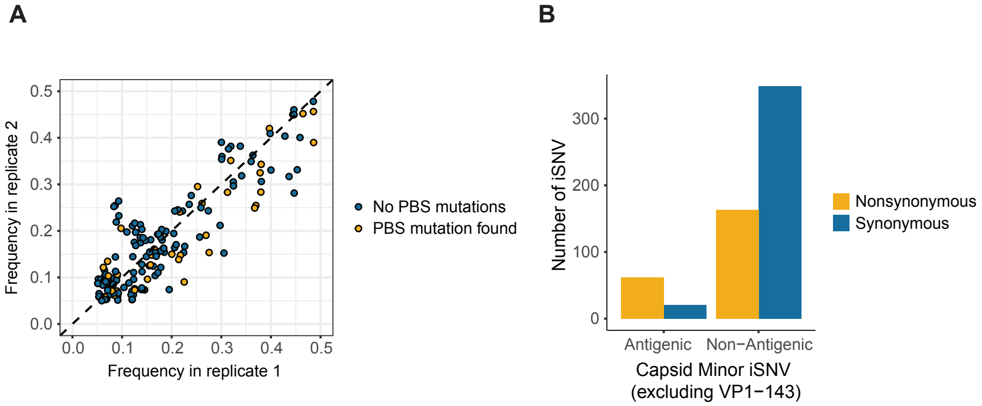


Figure 3.8. Minority SNV. (A) Concordance of iSNV frequency measurements across 11 samples sequenced in duplicate. Frequency of iSNV in replicate 2 (y-axis) is shown by the frequency of an iSNV in replicate 1 (x-axis), with colors showing iSNV on amplicon(s) with or without mutations in primer binding sites. (B) Histogram of minor iSNV in the capsid region by antigenic status, excluding VP1-143. Nonsynonymous iSNV are shown in yellow, and synonymous iSNV in dark blue.

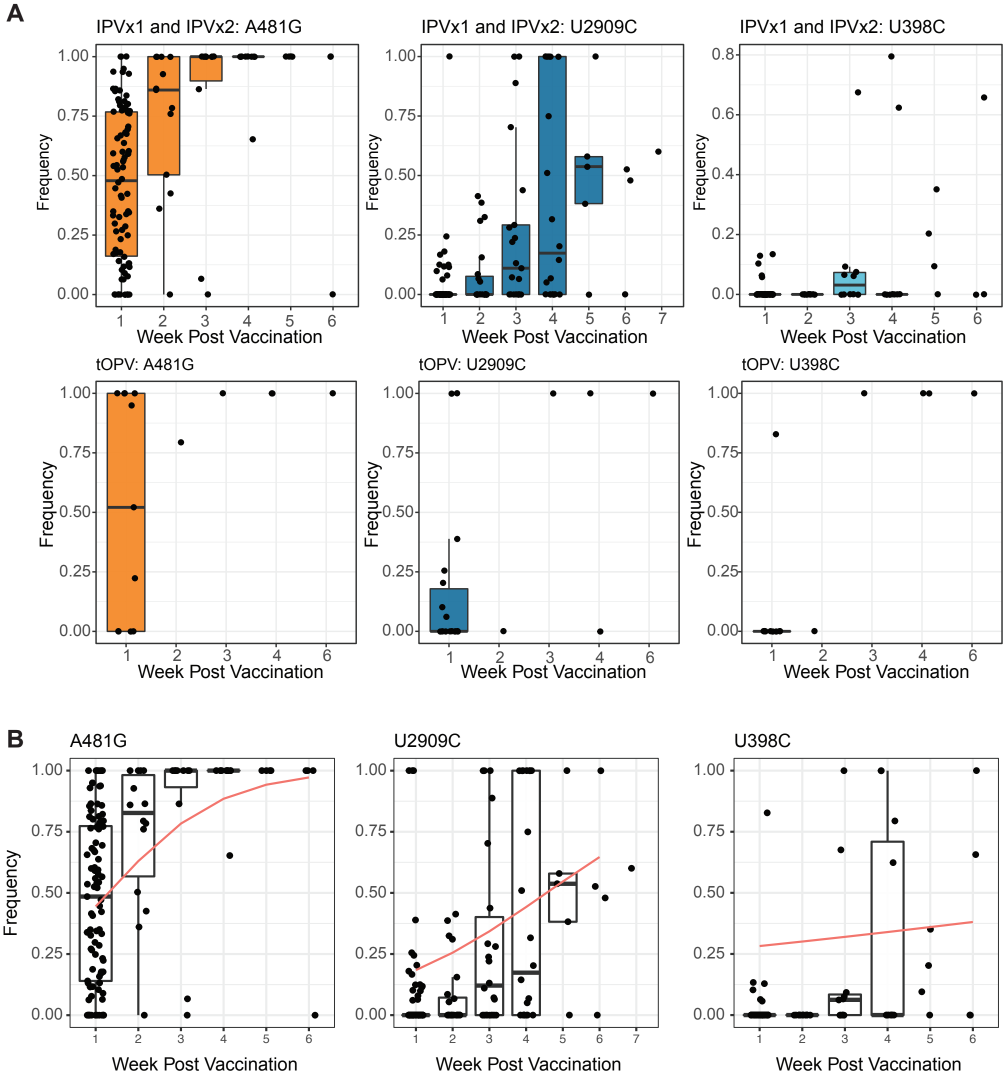


Figure 3.9. Gatekeeper mutations. (A) Frequency of A481G, VP1-143X, and U398C by time from vaccination across arms of the vaccine trial. Samples from IPV arms are shown on the top, and samples from tOPV arms are shown on the bottom. Each point represents one sample, and boxplots are shown for weeks with five or more data points. Boxplots represent the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range. (B) Frequency of A481G, VP1-143X, and U398C by time from vaccination across with the beta regression model fits for each mutation (red lines). The underlying data are the same as in Figure 3A. Each point represents one sample, and boxplots are shown for weeks with five or more data points. Boxplots represent the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range.

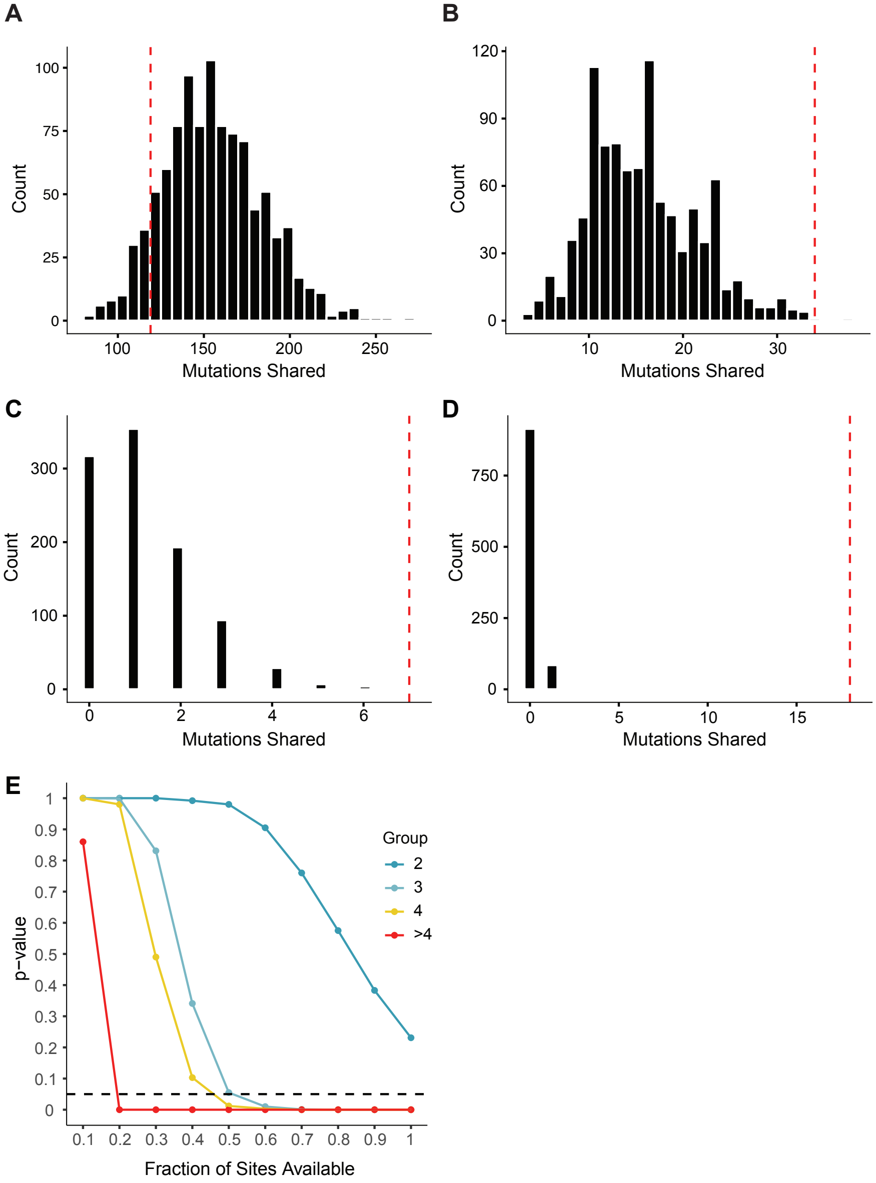


Figure 3.10. Permutation tests.. In A - D, the dotted red line is the observed number of mutations shared across a given number of individuals, and the bars show the results of 1000 random permutations. Results are shown for mutations shared across two individuals (A), three individuals (B), four individuals (C), and greater than four individuals (D). P-values are designated as the proportion of random permutations equal or greater than the observed number of shared mutations. In E, the p-values for each group are shown as a function of the genome fraction available for mutations. The horizontal dotted line represents $\alpha = 0.05$.

A

Frequency[1]	0.072	0.135	0.121	0.132	0.152	0.081	0.402	0.097	0.088	0.169	0.064	0.116	0.109	0.126	0.058	0.102	0.13	0.125
Position[2]	396	481	888	1035	1510	1593	1641	2006	2115	3184	3352	3579	4143	4207	4665	4692	4707	4716
396	2421	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
481	1457	2636	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
888	0	0	1804	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1035	0	0	819	2369	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1510	0	0	0	0	3199	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1593	0	0	0	0	2032	3621	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1641	0	0	0	0	1447	2968	3931	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2006	0	0	0	0	0	0	0	4383	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2115	0	0	0	0	0	0	0	2121	4555	NA	NA	NA	NA	NA	NA	NA	NA	NA
3184	0	0	0	0	0	0	0	0	0	1580	NA	NA	NA	NA	NA	NA	NA	NA
3352	0	0	0	0	0	0	0	0	0	439	6563	NA	NA	NA	NA	NA	NA	NA
3579	0	0	0	0	0	0	0	0	0	0	637	9323	NA	NA	NA	NA	NA	NA
4143	0	0	0	0	0	0	0	0	0	0	0	0	7996	NA	NA	NA	NA	NA
4207	0	0	0	0	0	0	0	0	0	0	0	0	5423	8279	NA	NA	NA	NA
4665	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7860	NA	NA	NA
4692	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7308	8118	NA	NA
4707	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6908	7579	8361	NA
4716	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6405	7089	7664	8326

[1] Frequency of minor variants found in donor of pair 702 at the 19 positions listed.

[2] Positions of 19 minor variants found in donor of pair 702. Table values are the total number of reads overlapping both positions.

B

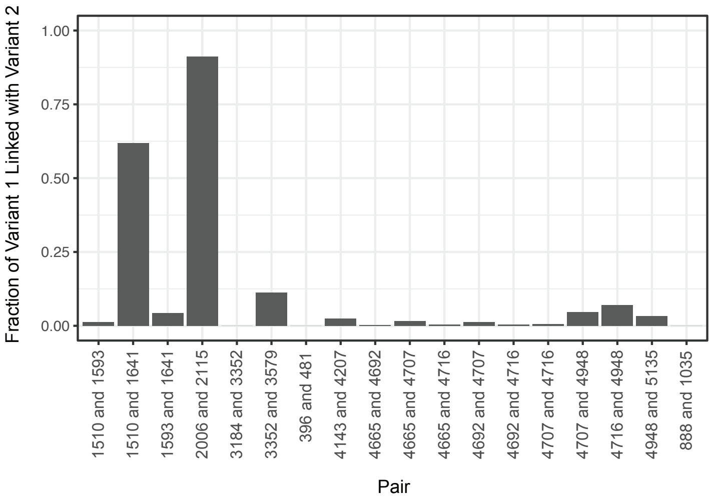


Figure 3.11. Linkage of mutations. (A) The frequency of 20 minor variants present in the donor for pair 702 (top). The table values show the number of sequence reads overlapping each pair of minor variants. (B) Bar chart showing the fraction of minor variant 1 found linked to minor variant 2 in overlapping sequence reads. The 18 pairs of minor variants are shown here by their genome position.

CHAPTER IV

Temporal Dynamics of SARS-CoV-2 Mutation Accumulation Within and Across Infected Hosts

Note: This chapter is a modified version of the published article:

Valesano AL, Rumfelt KE, Dimcheff DE, Blair CN, Fitzsimmons WJ, Petrie JG, Martin ET, Lauring AS. 2021. "Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts." PLoS Pathog 17(4):e1009499.

PLOS grants all authors full rights to reuse their articles in dissertations.

Introduction

Over the course of the SARS-CoV-2 pandemic, whole genome sequencing has been widely used to characterize patterns of broad geographic spread, transmission in local clusters, and the spread of specific viral variants (Fauver et al., 2020; Geoghegan et al., 2020; Meredith et al., 2020; Miller et al., 2020; Munnink et al., 2020; Sekizuka et al., 2020). Early reports demonstrated that SARS-CoV-2 exhibits genetic diversity within infected hosts, but this has been less studied than consensus-level genomic diversity (Shen et al., 2020). Intrahost diversity is an important complement to consensus sequencing. Patterns of viral intrahost diversity throughout individual infections can suggest the relative importance of natural selection and stochastic genetic drift (Lauring, 2020). Shared intrahost variants between individuals can reveal loci under convergent evolution and enable measurement of the transmission bottleneck, a critical determining factor in the spread of new genetic variants (Gutierrez et al., 2019; Lythgoe et al., 2020a). Studies of SARS-CoV-2 intrahost diversity may shed light on selective pressures applied at the individual level, such as antivirals and antibody-based therapeutics. While a clear understanding of within-host evolution can inform how SARS-CoV-2 spreads on broader scales, there have been relatively few comprehensive studies of intrahost dynamics (Lythgoe et al., 2020a; Popa et al., 2020; Tonkin-Hill et al., 2020).

Sequencing of intrahost populations can also potentially be applied to genomic epidemiology (Villabona-Arenas et al., 2020). A common goal in sequencing specimens from case clusters is to infer transmission linkage, which can guide future public health and infection control interventions. However, the relatively low substitution rate and genetic diversity of SARS-CoV-2 present challenges to inference of individual transmission pairs (Sikkema et al., 2020; Villabona-Arenas et al., 2020). In the pandemic setting, there is a non-negligible chance that two individuals who are epidemiologically unrelated could be infected with nearly identical viral genomes. Viruses from a single local outbreak may have few differentiating substitutions, limiting the ability of sequencing to resolve exact transmission chains. Identification of shared intrahost variants between individuals has been explored in other pathogens to overcome this obstacle (Maio et al., 2018; Martin et al., 2018; Skums et al., 2018; Worby et al., 2014, 2017). However, use of this approach for SARS-CoV-2 will depend on a solid understanding of the forces that shape the generation and spread of genetic variants.

There are several unresolved questions that will dictate the utility of intrahost diversity for genomic epidemiology. First, there must be sufficient intrahost diversity generated during acute infection prior to a transmission event. How much intrahost diversity is accumulated over time from infection onset is currently unknown. Second, the population bottleneck during transmission must be sufficiently wide to allow minor variants to be transmitted to recipient hosts (McCrone and Lauring, 2018; Zwart and Elena, 2015). Third, *de novo* generation of the same minor variants across multiple infections must be sufficiently rare. Independent generation of shared minor variants by recurrent positive selection or genetic drift in unlinked hosts could confound transmission inference (Worby et al., 2017). Finally, measurements of intrahost diversity must be accurate and account for several potential sources of error (Grubaugh et al., 2019a; McCrone and Lauring, 2016). Although previous studies have described within-host variation of SARS-CoV-2 (James et al., 2020; Lythgoe et al., 2020a; Moreno et al., 2020; Popa et al., 2020; Shen et al., 2020; Tonkin-Hill et al., 2020; Wang et al., 2020b), few have addressed the sources of systematic errors and batch effects in variant identification. To assess the utility of SARS-CoV-2 intrahost diversity for transmission inference, we need a clearer understanding of its temporal variation throughout infection and the extent of convergent evolution across individuals. Addressing these questions will also be valuable for understanding SARS-CoV-2 evolution.

Here, we sequenced SARS-CoV-2 genomes from 325 residual upper respiratory samples from hospitalized patients and employees at the University of Michigan. To validate our sequencing approach, we sequenced defined mixtures of two synthetic RNA controls and found that low input viral load decreases the specificity of variant calling. We find that observed intrahost diversity does not vary significantly by day since symptom onset. Intrahost variants can be shared between individuals that are unlikely to be related by transmission, suggesting that variants can arise by parallel evolution. These results inform our understanding of SARS-CoV-2 diversification in human hosts and highlight important considerations for sequence-based inference in the virus's genomic epidemiology.

Methods

We collected clinical metadata and residual diagnostic specimens positive for SARS-CoV-2 from hospitalized patients enrolled in the CDC HAIVEN (Hospitalized Adult Influenza Vaccine Effectiveness Network) study and infected employees enrolled in the HARVI (hospital associated respiratory virus infection) study. These studies and the use of residual specimens were approved by the University of Michigan Institutional Review Board.

Date of illness onset for hospitalized patients was collected individually via medical chart abstraction from physician notes. Michigan Medicine employees with any suspected COVID-19 symptoms were asked to call a COVID-19 healthcare worker hotline before reporting to work. Date of symptom onset, a list of symptoms, close contacts, travel history, and work location and description were recorded. After testing, employee clusters were determined by illness onset date, positive test status, and work location.

Genome amplification and sequencing

Residual samples from nasopharyngeal swabs and sputum specimens were centrifuged at 1200 x g. and 200 microliters were aliquoted. RNA was extracted with the Invitrogen PureLink Pro 96 Viral RNA/DNA Purification Kit and eluted in volumes of 100 microliters. Complementary DNA was reverse transcribed with SuperScript IV (ThermoFisher). The SARS-CoV-2 genome was amplified in two multiplex PCR reactions using the ARTIC Network V3 primer sets. Sequencing libraries were prepared with the NEBNext Ultra II kit and pooled in equal volumes after barcoding. The pooled sequencing library was gel extracted to remove adapter dimers. Libraries were sequenced on an Illumina MiSeq at the University of Michigan Microbiome Core facility (v2 chemistry, 2x250 cycles). To validate this approach, we used two synthetic RNA controls that differ by seven single nucleotide mutations, Wuhan-Hu-1 and EPI_ISL_418227 (Twist Bioscience, San Francisco, CA). We mixed the two RNAs at various copy numbers (10^5, 10^4, 10^3, 10^2 genome copies/μL) and frequencies (0%, 0.25%, 0.5%, 1%, 2%, 5%, 10%, and 100%). We amplified and sequenced each RNA mixture as described above. We defined true positive iSNV as mutations at the seven sites that differ between the two synthetic RNA controls (C241U, C335U, C2416U, C3037U, C14408U, A23403G, G25563U). We defined false positives as any iSNV other than the seven true-positive mutations.

Viral load measurements

We measured SARS-CoV-2 genome copy concentration for each sample by qPCR using conditions outlined in the CDC 2019-Novel Coronavirus EUA protocol (https://www.fda.gov/media/134922/download). The nucleocapsid gene was amplified using the CDC N1 primer and probe set as follows: 2019-nCoV_N1 Forward Primer GACCCCAAAATCAGCGAAAT; 2019-nCoV_N1 Reverse Primer TCTGGTTACTGCCAGTTGAATCTG; 2019-nCoV_N1 Probe ACCCCGCATTACGTTTGGTGGACC. Probe sequences were FAM labeled with Iowa Black quencher (Integrated DNA Technologies, Coralville, IA). Reactions were performed using TaqPath 1-step RT-qPCR master mix (Thermofisher, Waltham, MA) with 500 nM of each primer and 250 nM of each probe in a total reaction volume of 20 μl. Cycling conditions were as follows: 2 min at 25 °C, 15 min at 50 °C, 2 min at 95 °C, and 45 cycles of 3 seconds at 95 °C, 30 seconds at 55 °C. Samples were run on an Applied Biosystems 7500 FAST real-time PCR system. Cycle threshold (Ct) was designated uniformly across PCR runs.

Standard curves based on serial dilutions of a plasmid containing the nucleocapsid sequence were used to determine copy number for each plate of samples. Copy number is expressed in genome copies per microliter of extracted viral RNA.

Analysis of sequence reads

We aligned reads to the MN908947.3 reference genome with BWA-MEM version 0.7.15 (Li, 2013). We removed sequencing adaptors and trimmed ARTIC primer sequences with iVar 1.2.1 (Grubaugh et al., 2019a). We determined the consensus sequences with iVar 1.2.1, taking the most common base as the consensus (>50% frequency). We placed an N at positions along the MN908947.3 reference with fewer than 10 reads. We manually inspected insertions and deletions by visualizing alignments with IGV (version 2.8.0) (Robinson et al., 2011). We identified intrahost single nucleotide variants relative to the MN908947.3 reference genome with iVar 1.2.1 using the following parameters: sample with viral load ≥ 10^3 copies/μL; sample with consensus genome length of ≥ 29000; sample with ≥ 80% of genome sites above 200x coverage; iSNV frequency of 2 - 50%; read depth of ≥ 100 at iSNV sites; ≥ 10 reads with average Phred score of > 35 supporting a given iSNV; iVar p-value of < 0.0001. All samples on which we called variants had > 50,000 mapped reads. We accounted for strand bias by performing a two-sided Fisher's exact test for hypothesis that the forward/reverse strand counts supporting the variant base are derived from the same distribution as the consensus base. We then applied a Bonferroni multiple test correction and excluded variants with an adjusted p-value < 0.05. We used a multiple linear model to evaluate the correlation of sample iSNV richness to day post symptom onset and viral load (base 10 log). To generate a phylogenetic tree, we aligned consensus genomes with MUSCLE 3.8.31 and masked positions that are known to commonly exhibit homoplasies or sequencing errors (2020a). We generated a maximum likelihood phylogeny with IQ-TREE, using a GTR model and 1000 ultrafast bootstrap replicates (Edgar, 2004; Nguyen et al., 2015). Evolutionary lineages (Pango lineages) were assigned with PANGOLIN (Rambaut et al., 2020).

Data and code availability

Raw sequence reads are available as FASTQ files from the Sequence Read Archive at accession number PRJNA682212, with human-mapping reads removed. Analysis code is available at https://github.com/lauringlab/SARSCov2_Intrahost. Consensus genome sequences are publicly available at the GitHub link and on GISAID.

Results

We retrieved respiratory specimens collected through diagnostic testing from March – May 2020. We sequenced samples from two groups: inpatients who were part of an observational study of COVID-19 in hospitalized individuals (n = 190), and symptomatic employees who presented to occupational health services (n = 135). All employees were diagnosed and treated in outpatient settings, except for one who was admitted as an inpatient. Basic demographic information is described in a separate work (Dimcheff et al., 2021). Genome copy number determined by qPCR of the nucleocapsid gene was highly variable and decreased by day from symptom onset ($p < 0.001$, linear model, Figure 4.1A). We obtained 212 complete genomes (Figure 4.1B), mostly from samples with higher viral loads (Figure 4.1B). Consensus genomes had a median of 7 substitutions relative to the Wuhan-Hu-1/2019 reference sequence (range 4 – 12). Phylogenetic analysis of whole consensus genomes identified 10 unique evolutionary lineages in our cohort (lineages determined by the PANGOLIN system, see Methods; Figure 4.1C). Most sequenced genomes fell in lineage B.1. We evaluated whether any employees were part of an epidemiologically linked cluster based on illness onset date, positive test status, and work location. We found that some employees were part of epidemiologically linked clusters (Figure 4.1C). The genomes from clusters 2, 10, 19, 20, and one pair in cluster 29 had ≤ 1 consensus difference, while the rest had 2 – 7 differences. Many employees in different clusters also had identical or nearly identical consensus genomes, which reflects the low genetic diversity of SARS-CoV-2 at this stage of the pandemic. We have no information on epidemiologic linkage for the remaining sequenced individuals. It is highly unlikely that there are direct transmission pairs in our dataset, but we cannot conclusively rule out coincident transmission linkage. Therefore, this population largely reflects a cross-section of infected individuals who are epidemiologically unlinked.

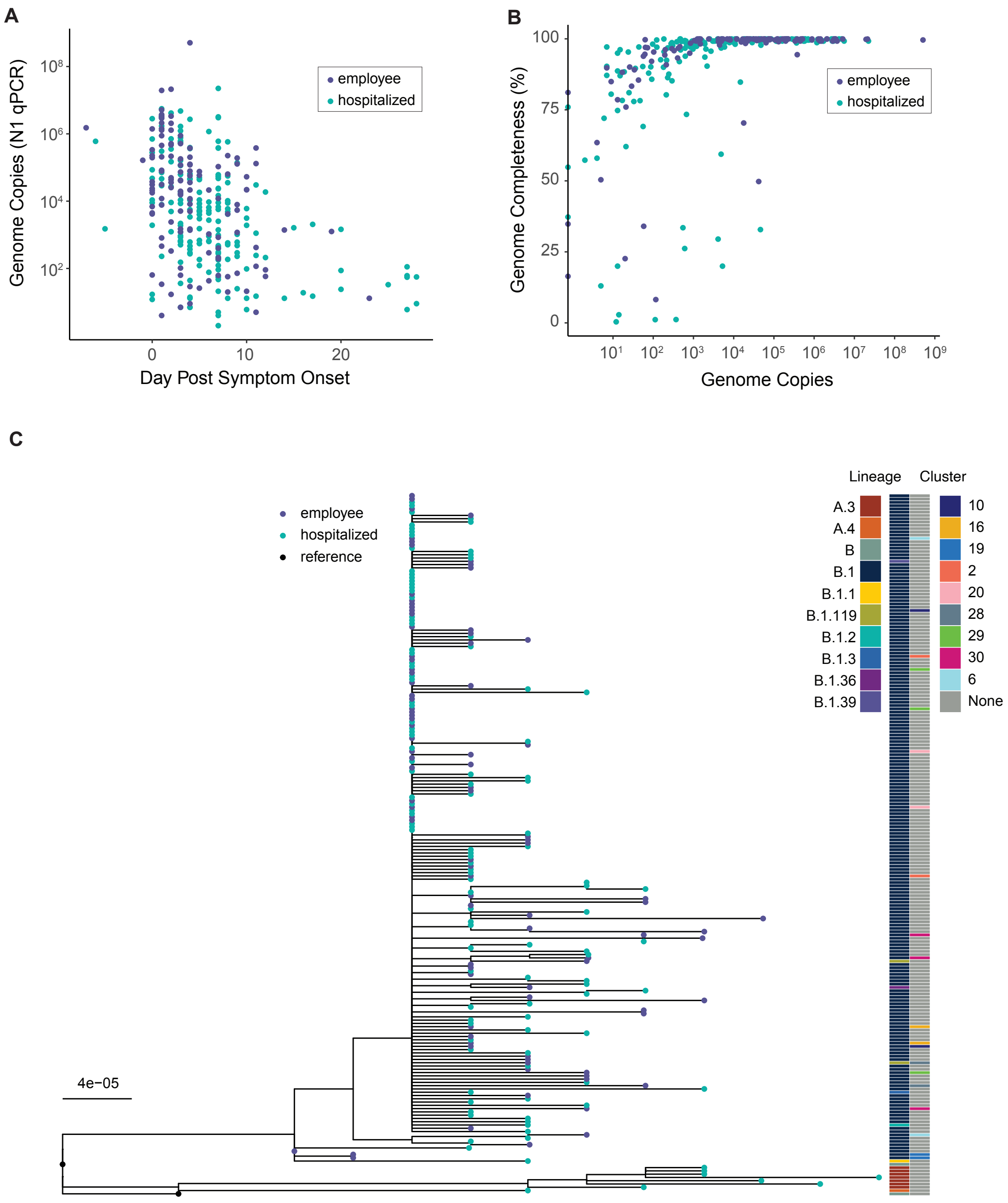


Figure 4.1. Viral shedding and overview of genome sequencing data. (A) Viral load by day of infection in hospitalized patients (teal) and employees (violet). Viral load, measured by qPCR of the N gene in units of genome copies per microliter of extracted RNA, is on the y-axis and day post symptom onset is on the x-axis. (B) Genome completeness by viral load in hospitalized patients (teal) and employees (violet). Viral load as shown in (A) is on the x-axis and the fraction of the genome covered above 10x read depth is shown on the y-axis. (C) Maximum-likelihood phylogenetic tree. Tips represent complete consensus genomes from hospitalized patients (teal) and employees (violet). The axis shows divergence from the root (Wuhan-Hu-1/2019). The second genome displayed as "reference" is Wuhan/WH01/2019. Heatmaps show PANGOLIN evolutionary linage (left) and epidemiologic cluster (right).

Benchmarking accuracy of variant identification

Identification of viral within-host variants can be prone to errors (Grubaugh et al., 2019a; McCrone and Lauring, 2016). Therefore, we performed a mixing study to evaluate the accuracy of our pipeline for identifying intrahost single nucleotide variants (iSNV). We mixed two synthetic RNA controls that differ by seven single nucleotide substitutions at defined frequencies and input concentrations (Figure 4.2A). These mixtures were sequenced using the same approach as the clinical samples. We identified true iSNV at the expected frequencies at ≥ 10^3 copies/μL (Figure 4.2B). There was greater variance in the observed variant frequencies at 10^2 copies/μL compared to higher input concentrations. We obtained high sensitivity for iSNV at ≥ 2% frequency and ≥ 10^3 copies/μL with sufficient genome coverage. Many false positive iSNV remained at ≥ 2% frequency and 10^2 copies/μL despite multiple quality filters (Figure 4.6). However, false positive iSNV per sample drastically decreased with input concentrations ≥ 10^3 copies/μL. Three false positive iSNV were identified in multiple samples above 10^4 copies/μL: A3350U, G6669A, and U13248A. Mutation U13914G recurred in multiple samples at input concentrations of 10^3 copies/μL and below. We suspect that they represent low-frequency variants present in the synthetic RNA controls, as has been observed in other studies with synthetic controls from the same manufacturer (Lythgoe et al., 2020a). Excluding these recurrent sites, there were few false positive iSNV per sample with input concentrations above 10^3 copies/μL (Figure 4.2C). Together, these data indicate that sufficient input viral load is a critical factor for accurate identification of iSNV.

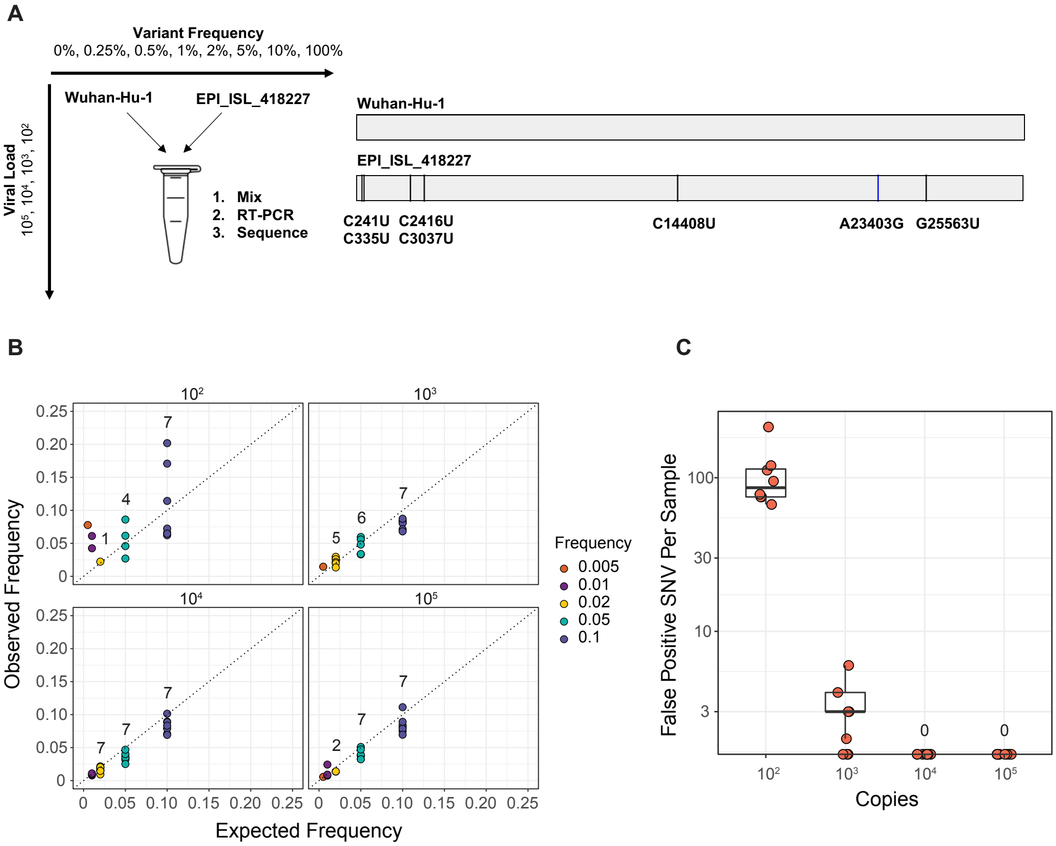


Figure 4.2. Assessing accuracy of intrahost variant detection by sequencing defined viral mixtures. (A) Schematic of the experiment. Wuhan-Hu-1 (reference) and EPI_ISL_418227 (variant) RNA were mixed at the given frequencies and viral loads (units of genome copies per microliter, representing the resulting mixture). Mixtures of RNA were amplified and sequenced in the same fashion as the clinical specimens. Reference and variant genomes differ by seven single nucleotide substitutions. (B) Observed frequency by expected frequency. Observed frequency of the true positive intrahost single nucleotide variants (iSNV) is on the y-axis and expected iSNV frequency is on the x-axis. Synthetic RNA copy number in units of genome copies per microliter of RNA is shown above each facet. Values above the points indicate the number of variants detected in that group (maximum of seven per group). (C) False positive iSNV. Number of false positive iSNV per sample is shown on the y-axis (base 10 log scale) and viral load as shown in (B) is on the x-axis, excluding iSNV at positions 3350, 6669, 13248, and 13419. Each point represents a unique sample and the boxplots represent the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range.

Low within-host diversity of SARS-CoV-2

Based on our benchmarking experiment, we identified iSNV in 178 specimens with viral loads $\geq 10^3$ copies/μL (Figure 4.3A). We excluded position 11083, which is near a natural poly-U site and prone to sequencing errors (2020b), as well as the four sites with recurrent false positives (nucleotide positions 3350, 6669, 13248, and 13914). Most specimens exhibited fewer than ten minor iSNV (median 1, IQR 0 – 2, Figure 4.3B). There were four outlier specimens with greater than 15 iSNV. In these samples, iSNV were dispersed throughout the genome at various frequencies, so it is difficult to determine whether they represent mixed infections (Tonkin-Hill et al., 2020). The locations of these samples on sequencing plates were not suggestive of cross-contamination. There was no difference in minor iSNV richness between hospitalized patients and employees treated as outpatients ($p = 0.25$, Mann-Whitney U test, Figure 4.7). We identified more minor iSNV encoding non-synonymous changes than synonymous ones across most open reading frames (Figure 4.3C) and identified more iSNV at lower frequencies (Figure 4.3D), which together is suggestive of mild within-host purifying selection. Sample iSNV richness decreased with higher viral loads by about 1 iSNV per 10-fold increase in viral load ($p = 0.01$, multiple linear model, Figure 4.8). Sample iSNV richness did not correlate with day from symptom onset ($p = 0.79$, multiple linear model, Figure 4.3E). This result was robust to exclusion of the four outlier samples and exclusion of viral load from the model. These results show that within-host diversity is low and remains that way over the duration of most SARS-CoV-2 infections.

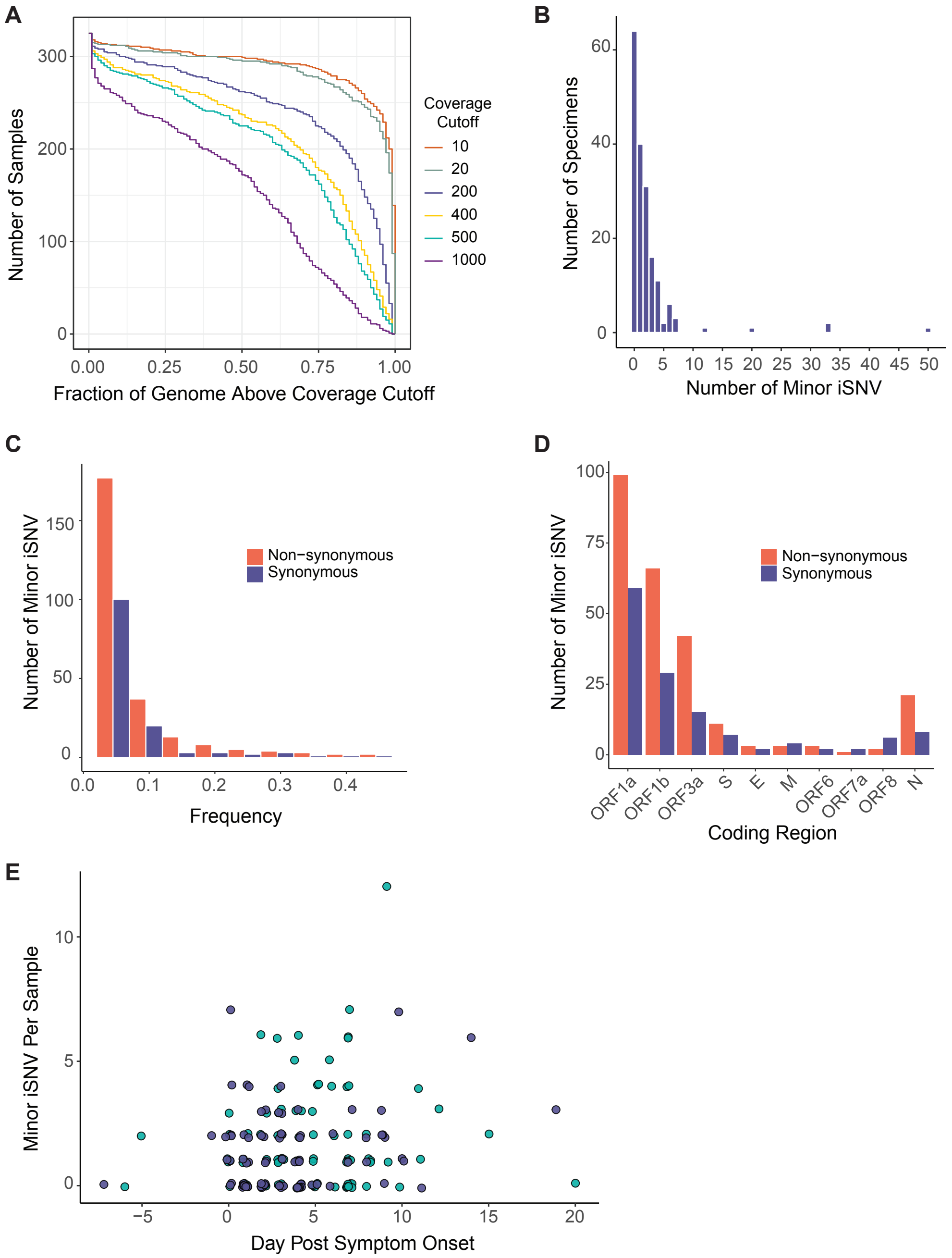


Figure 4.3. SARS-CoV-2 intrahost single nucleotide variant diversity. (A) Sequencing coverage for clinical samples. The number of clinical samples (y-axis) is shown by the fraction of the genome above a given read depth threshold (x-axis). The different lines show the data evaluated with six read depth thresholds. (B) Histogram of the number of specimens (y-axis) by the number of minor iSNV per sample (x-axis), n = 178. (C) Number of minor iSNV by frequency with a bin width of 0.05. Non-synonymous iSNV are shown in orange and synonymous iSNV are shown in violet. (D) Number of minor iSNV by coding region. Non-synonymous iSNV are shown in orange and synonymous iSNV are shown in violet. (E) Scatterplot of the number of minor iSNV per sample (y-axis) by the day post symptom onset (x-axis). Hospitalized patients are shown in teal and employees shown in violet. The four samples with > 15 iSNV shown in (B) are excluded from the plot for visualization.

Variants shared across individuals

Next, we investigated patterns of shared intrahost diversity between individuals. Most iSNV were unique to a single individual. However, 18 iSNV were present in multiple specimens (Figure 4.4A). None of these mutations were located at sites known to commonly produce errors or homoplasies (van Dorp et al., 2020, 2020b). Two iSNV were present in three individuals (G12331A and A11782G, both synonymous changes in ORF1a). There was no clear phylogenetic clustering of genomes exhibiting these shared iSNV (Figure 4.9), and G12331A was shared between samples from different viral lineages (13 substitutions). These two mutations were first detected in our samples in late March 2020 (Figure 4.4B). None reached > 1% frequency per week in consensus sequences submitted to GISAID through mid-November 2020. These results suggest that iSNV that arise convergently across viral lineages are not necessarily predictive of subsequent global spread of those mutations.

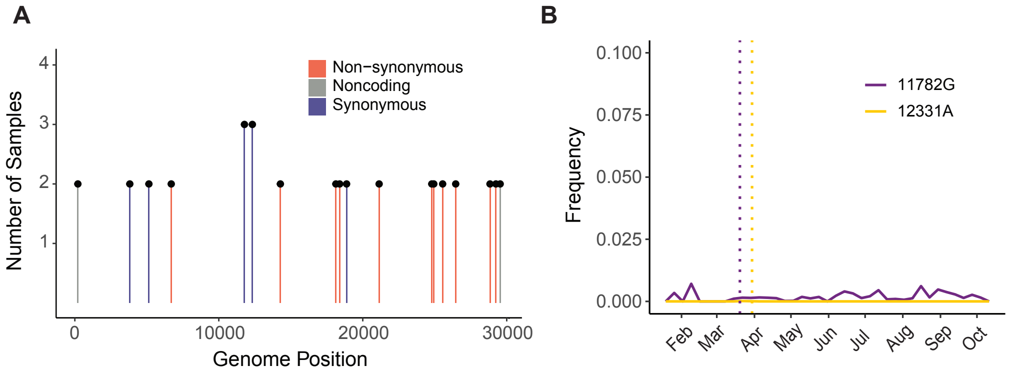


Figure 4.4. Shared iSNV across samples and their frequency in global consensus genomes. (A) Shared iSNV across samples, with the number of samples sharing the iSNV (y-axis) by the genome position (x-axis). Colors indicate the iSNV coding change relative to the reference. (B) The frequency (y-axis) of three iSNV shared by three or more samples over time (x-axis). The consensus genomes are from GISAID, as available on 2020-11-11. The vertical dotted lines represent the earliest time we detected each iSNV in our samples.

Transmission inference based on shared iSNV integrates information such as consensus genome sequences, sample dates, and shared iSNV (Worby et al., 2017). Therefore, we compared shared iSNV across all unique pairs of specimens used for variant calling (n = 15753, Figure 4.5). Because most iSNV were unique to an individual, most pairs did not share iSNV and only 0.14% of pairs shared one iSNV. Many pairs with shared iSNV were sequenced in separate batches, which reduces the likelihood that shared iSNV are due to cross-contamination. No employee pairs in the same epidemiologic cluster shared iSNV, which are the only pairs in our dataset who are likely part of the same transmission network. The rest of the pairs of individuals are most likely not directly linked by transmission and probably share iSNV by chance. We identified nine unique pairs with shared iSNV between genomes that were near-identical (0 – 1 consensus differences), three of which were collected within one week of each other. We also identified shared iSNV between 13 pairs separated by ≥ 2 consensus substitutions (Figure 4.5A and 4.5B) and 15 pairs with collection dates 7 – 28 days apart (Figure 4.5B). Due to differences in viral lineage and time of collection, these are very unlikely to be transmission pairs. Together, these data indicate that iSNV can arise convergently between individuals who are unlikely to be related by transmission.

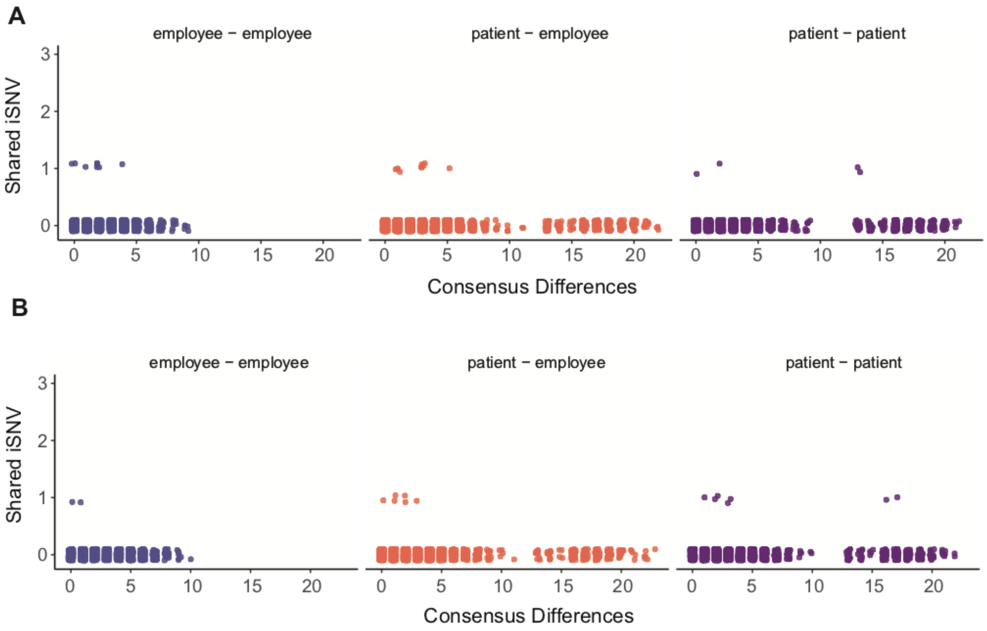


Figure 4.5. Pairwise comparisons of shared iSNV. Each unique pair is shown as a single point, with employee-employee pairs in violet (left), patient-employee pairs in orange (middle), and patient-patient pairs in purple (right). The number of iSNV shared by each pair is shown on the y-axis with the number of consensus differences between the pair of genomes on the x-axis. Pairs of samples collected within seven days of each other are displayed in (A), and pairs of samples collected greater than seven days apart are shown in (B).

Discussion

Accurate characterization of SARS-CoV-2 intrahost diversity is important for understanding the spread of new genetic variants and its potential use in transmission inference. In this study, we sequenced upper respiratory specimens from a cohort of hospitalized COVID-19 patients and infected employees. We found that intrahost diversity is low and its distribution does not vary by time since symptom onset. We identified iSNV shared across viral genomes separated by time and disparate evolutionary lineages, indicating that iSNV can arise convergently. Because variants may be shared through parallel mutation rather than transmission, caution is warranted in the use of shared iSNV alone for inferring transmission chains. Intrahost variants shared across multiple individuals did not precede an increase in frequency in global consensus genomes, which suggests that identifying convergent iSNV may have limited utility in tracking broader SARS-CoV-2 evolution.

Specimen viral load is important when measuring intrahost diversity. We and others have shown that samples with low viral loads are prone to false positive iSNV and lower sensitivity (Grubaugh et al., 2019a; McCrone and Lauring, 2016; Valesano et al., 2020b). A strength of our study is that we experimentally validated the accuracy of our variant calling by sequencing defined populations. Based on these results, we excluded samples with low viral load from subsequent analyses. Future studies of SARS-CoV-2 intrahost diversity should report and account for specimen viral loads to avoid this common source of error. We did not benchmark our sequencing approach for detecting insertions and deletions (indels) and therefore did not report these for the clinical specimens.

Intrahost indels could conceivably provide useful information about within-host evolution, but accurate detection is also subject to similar issues of sample quality and viral load.

The low level of intrahost diversity that we found here is consistent with a recent preprint by Lythgoe et al. (Lythgoe et al., 2020a). The fact that our work and the study by Lythgoe et al. were performed with different geographical areas, sequencing approaches (ARTIC Network amplicons vs. veSEQ metagenomic sequencing), and analysis methods lends credence to the results. Lythgoe et al. reported more shared variation than seen here, but this is most likely due to sequencing a greater number of samples among individuals within known epidemiologic clusters. We and Lythgoe et al. measure a lower level of intrahost diversity at the 2% frequency threshold compared to a recent study in Austria (Popa et al., 2020). The reasons for this are not clear, but it is likely due to differences in sample viral loads and variant calling methods. We did not find a difference in intrahost diversity between hospitalized COVID-19 patients and those treated as outpatients, which suggests that viral diversity may not be a reliable marker for disease severity.

Measuring viral diversity over the course of infection is relevant for understanding how variants are transmitted to new hosts. Only genetic variants present at the time of a transmission event will have the opportunity to spread. Because SARS-CoV-2 usually transmits just before or several days after symptom onset (He et al., 2020; Rhee et al.), it is important to define viral diversity in this window. Our cross-sectional analysis of diversity by time since symptom onset indicates that diversity does not significantly increase over the course of infection. A significant fraction of samples may not exhibit any iSNV at the time of transmission, which could limit the utility of iSNV for linking transmission pairs. Only a large bottleneck would lead to onward spread of most iSNV present during early infection. However, it is important to recognize that although the absolute level of diversity may not change over time, different variants may arise or go extinct during a given infection. This phenomenon was observed in a recent study by Tonkin-Hill et al. (Tonkin-Hill et al., 2020). Serial samples from individuals could address this issue with higher resolution. Low diversity within hosts also shapes our expectations for emergence of resistance to drugs and monoclonal antibodies. With such limited substrate for selection to act upon, the short window of time between treatment and transmission could limit the spread of a variant selected within a host. Even during prolonged infections in immunocompromised hosts, there is only limited evidence of resistance to various COVID-19 therapeutics (Baang et al., 2021; Buckland et al., 2020; Kemp et al., 2020).

Parallel evolution is a critical factor to consider in the interpretation of shared intrahost variation (Worby et al., 2017). Even if iSNV identification were perfectly specific, iSNV can arise in parallel due to biological processes such as natural selection and genetic drift. A key finding of this work is that iSNV can arise in genomes that are unrelated by local transmission, specifically those across large time intervals and lineages. Shared iSNV between individuals with identical genomes collected the same week may also have arisen in parallel. These pairs are most likely not epidemiologically linked, but we are unable to rule out coincident local transmission in the community. Because iSNV can arise in parallel in genomes that are not linked by transmission, caution is needed when relying entirely on shared iSNV for transmission inference (Tonkin-Hill et al., 2020; Villabona-Arenas et al., 2020).

We also found that identifying iSNV across multiple individuals did not precede an increase of those mutations in frequency in global consensus genomes. It is unclear whether these mutations arose due to positive selection, chance, or mutational "hotspots" (Tonkin-Hill et al., 2020). It is possible that these mutations were lost due to purifying selection within hosts or during transmission

(Lauring, 2020; Xue et al., 2018). These results suggest that iSNV may have lower utility for tracking broader SARS-CoV-2 evolution, but larger sample sizes in more geographic areas are necessary to evaluate this.

One of the most important variables for transmission inferences is the size of the transmission bottleneck (Worby et al., 2017). If parallel evolution of iSNV occurs regularly and the transmission bottleneck is very small, that would increase the likelihood that shared iSNV are due to convergence rather than transmission. However, if the bottleneck is large, then iSNV may become more valuable for detecting transmission networks when consensus genomes are limited. There are currently conflicting results on the SARS-CoV-2 bottleneck size. Popa et al. estimated a bottleneck size of greater than 1000 unique genomes (Popa et al., 2020). In contrast, Lythgoe et al. estimated a bottleneck size range from 1 – 8 unique genomes based on 14 household pairs (Lythgoe et al., 2020a). Lythgoe et al. in particular used extensive controls and validation for preventing contamination and identifying sequencing errors. Other studies both in humans and in domestic cats have estimated small bottlenecks (Braun et al., 2020; Wang et al., 2020a). It is difficult to interpret these contrasting results because each study used different sequencing and analysis methodologies. In recent work on influenza A virus, a study of methodological differences was key for resolving different conclusions about the bottleneck size (Xue and Bloom, 2019a). One factor that has not yet been clearly defined is how the time interval between donor-recipient pairs affects SARS-CoV-2 bottleneck estimates. We expect that further work will clarify the reasons behind these conflicting estimates.

Because of the high incidence and low mutation rate of SARS-CoV-2, genomic epidemiology is necessarily constrained in its ability to determine exact transmission chains in an outbreak. Using minor genetic variation to increase the resolution of genomic epidemiology requires attention to the underlying processes of within-host viral evolution and awareness of possible confounders. Unified statistical frameworks that incorporate sequences, metadata, and epidemiological models are likely the most robust approaches for integrating intrahost variants, but these models also must account for parallel evolution (Maio et al., 2018; Skums et al., 2018; Worby et al., 2017). As others have recently suggested (Tonkin-Hill et al., 2020), we caution against assigning transmission pairs solely by virtue of shared iSNV in the absence of clear epidemiologic information.

Acknowledgements

We thank the University of Michigan Clinical Microbiology Laboratory and the University of Michigan Central Biorepository for their assistance in providing samples. We thank Christina Cartaciano and the University of Michigan Microbiome Core for their assistance in sequencing. We thank Emily Stoneman from Michigan Medicine Occupational Health Services for assistance with employee data. This work was supported by a University of Michigan COVID-19 Response Innovation Grant (to ASL), K01AI141579 (to JGP) and CDC U01 IP000974 (to ETM).

Supplemental Figures

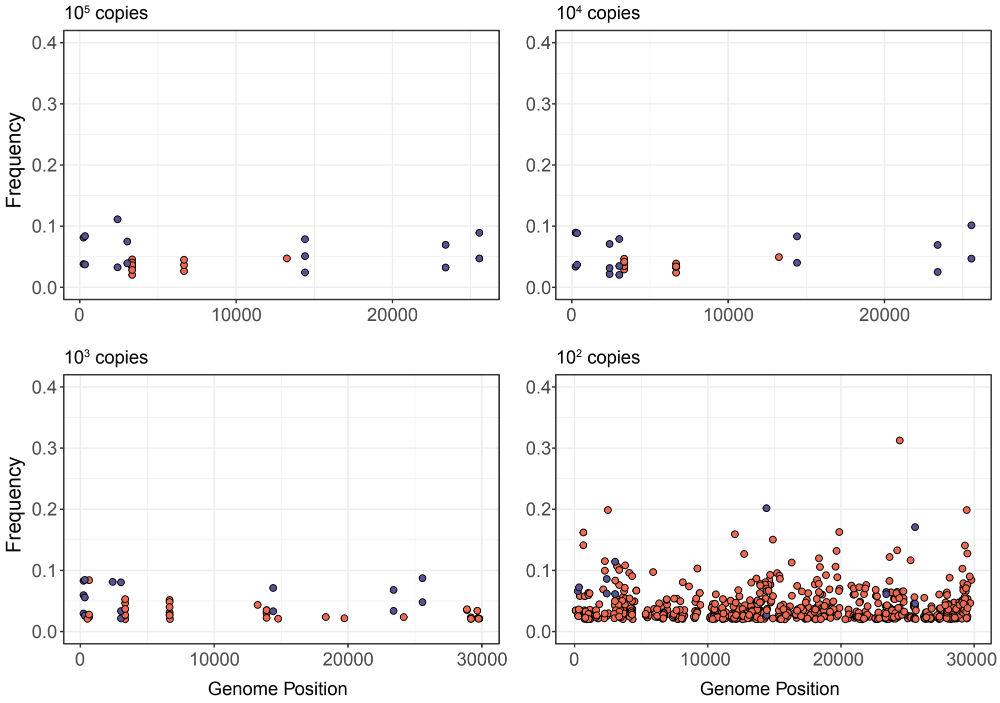


Figure 4.6. True and false positive iSNV in RNA mixture validation experiment. Each iSNV is shown as a point, with the frequency on the y-axis and genome position on the x-axis. True positive iSNV are shown in violet and false positive iSNV are shown in orange. All iSNV displayed have a frequency of 2% or greater. Viral loads are shown above each facet, in units of genome copies per microliter of RNA.

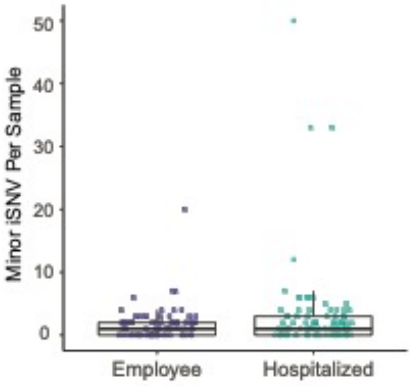


Figure 4.7. Number of minor iSNV per sample across groups. Hospitalized patients are shown by teal points and employees shown by violet points. Boxplots for each group represent the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range.

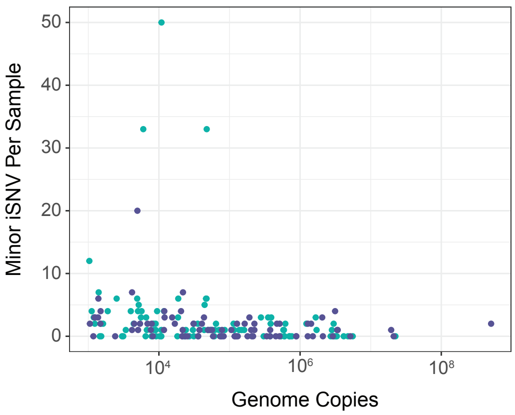


Figure 4.8. Number of minor iSNV per sample by genome copies. Hospitalized patients are shown by teal points and employees shown by violet points.

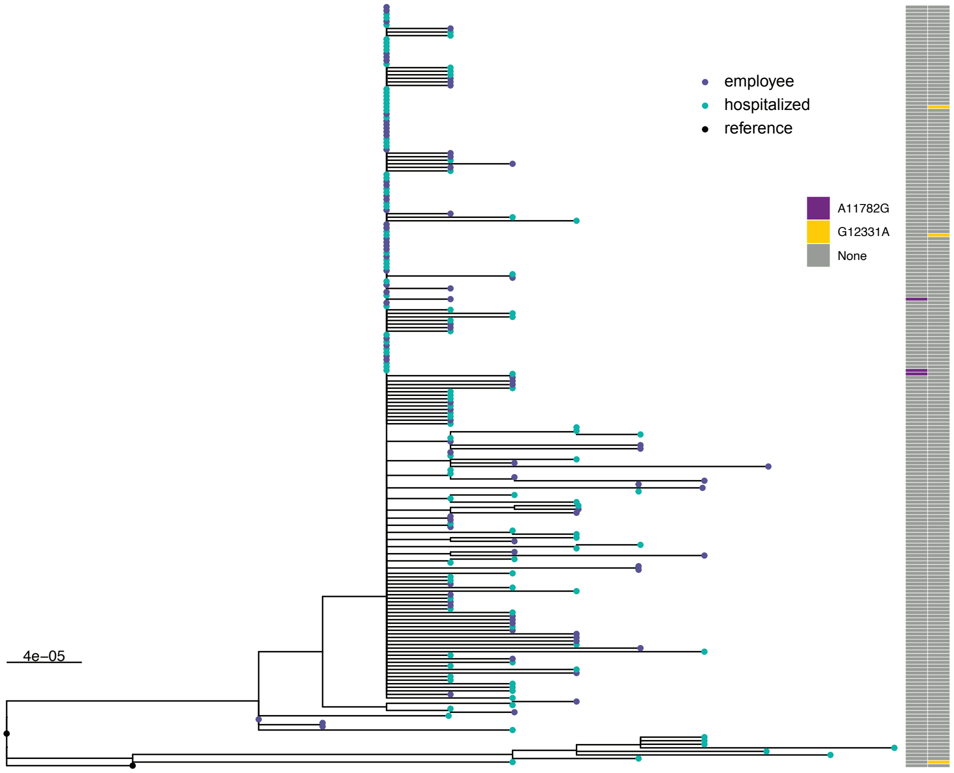


Figure 4.9. Maximum likelihood phylogenetic tree. Tips represent complete consensus genomes from hospitalized patients (teal) and employees (violet). The x-axis shows divergence from the root (Wuhan-Hu-1/2019). Heatmaps show samples that contain each mutation as an iSNV.

CHAPTER V

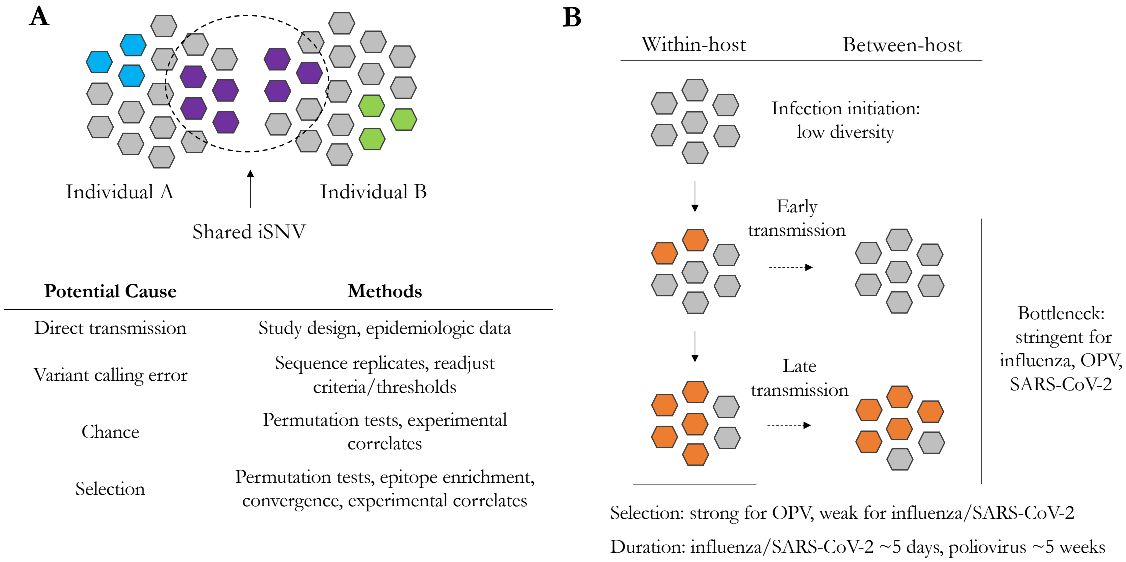
Discussion

Although it has been known for decades that RNA viruses establish genetically diverse populations, only recently has technology enabled comprehensive, genome-scale measurement of viral populations within human hosts. Such studies have revealed important aspects of viral evolution, transmission, and epidemiology that would not be revealed by consensus sequences. However, there are still challenges for accurate characterization of these populations and the level of biological inference that can be extracted from the results. In my thesis, I have applied careful methods for variant identification with rigorous benchmarking and validation. These experiments laid the primary foundation for making evolutionary inferences from the results. Each chapter demonstrates the value of moving beyond convenience samples and sequencing specimens from observational and clinical studies. In Chapter II, I used samples from a longitudinal household cohort to determine that influenza B virus accumulates less intrahost diversity than influenza A viruses. Combined with the measurement of a narrow transmission bottleneck, this points to a relationship between the mutation rate of a virus and its intrahost diversity. In Chapter III, I used a similar household-based field trial to identify mutations that are involved in the early reversion of the oral polio vaccine and to understand how mutations spread to new hosts. These data generated new hypotheses about the molecular basis of OPV reversion, informed poliovirus genetic surveillance, and suggested strategies for novel vaccine designs. In Chapter IV, I quantified the intrahost diversity of SARS-CoV-2 and the extent of parallel mutation among unrelated individuals. Contrary to several high-profile reports, I found little intrahost diversity over acute infections. My findings changed our expectations of how frequent of antiviral resistance will arise and how much minor variants contribute to global dynamics. Finally, I integrated what we have learned about viral within-host dynamics to assess its utility for enhancing genomic epidemiology and transmission reconstruction.

Challenges of interpreting shared variants

A common feature across viruses is that most intrahost variants are unique to a specific population. This is not surprising given the random nature of errors in genome replication. However, there is usually some fraction of variants that is shared by two or more individuals in a study population. The biological interpretation of these shared variants is central to evolutionary inferences. There are several reasons that minor variants may be shared: random or systematic sequencing error, chance (e.g. genetic drift), selection, or direct transmission (Figure 5.1A). Distinguishing these causes is not trivial and has been a major point of conflict in previous studies.

Prior to sequencing, proper study design is essential for interpretation of shared variants, especially for evaluating direct transmission. In order to assess whether variants are shared due to direct transmission, there must be documentation of exposure or epidemiologic link. In Chapter IV, we had epidemiologic data for no individuals except a few employee infection clusters, which limited our ability to compare diversity across transmission pairs. In Chapter II and III, we employed household-based cohorts. Household studies have long been used to estimate secondary attack rates and assess vaccine effectiveness (Petrie et al., 2013; Tsang et al., 2016). In the HIVE study, we identified putative transmission pairs by identifying households with multiple individuals who tested positive by RT-qPCR within one week of each other. We applied the sequence data to assess whether the genetic distance between the viruses was consistent with direct transmission. In all cases, household pairs had viruses with low genetic distance. However, we did identify identical viruses in people from different households, illustrating that sequence data alone is insufficient to assess transmission linkage. In the polio vaccine trial, we took advantage of RT-qPCR monitoring of household contacts in the weeks following vaccination of one household individual. Because the original field trial included household and community transmission as a primary endpoint, we were able to compare variants across transmission pairs with high confidence. One drawback of these cohorts was a relatively small number of transmission pairs. In the OPV trial, high population immunity enabled the study to occur safely and ethically but limited the extent of transmission. In the HIVE study, the number of pairs depended on the severity of a given influenza season, and some study households had a positive influenza case. Different study designs may have better enrichment for transmission pairs. Case-ascertainment studies identify index cases in clinic settings and then enroll the appropriate household contacts (Iyengar et al., 2015).


A
Individual A
Individual B
Shared iSNV
B
Within-host
Between-host
Infection initiation: low diversity
Early transmission
Late transmission
Bottleneck: stringent for influenza, OPV, SARS-CoV-2
Selection: strong for OPV, weak for influenza/SARS-CoV-2
Duration: influenza/SARS-CoV-2 ~5 days, poliovirus ~5 weeks

Potential Cause	Methods
Direct transmission	Study design, epidemiologic data
Variant calling error	Sequence replicates, readjust criteria/thresholds
Chance	Permutation tests, experimental correlates
Selection	Permutation tests, epitope enrichment, convergence, experimental correlates

Figure 5.1. Summary of findings. (A) Depiction of iSNV shared between two individuals. The potential causes for shared iSNV and ways to investigate them are shown in the table. (B) Summary of findings. Each virus had relatively limited intrahost diversity. Selection was strong for OPV and weaker for influenza virus and SARS-CoV-2. All three viruses have stringent transmission bottlenecks. If transmission occurs later, iSNV may be more likely to pass through the narrow bottleneck.

In addition to study design, accurate identification of minor variants is critical for interpreting shared variants. My approach to limit sequence errors in each chapter was tailored to each specific context and viral system. Common to each study were experiments sequencing known viral mixtures to estimate the sensitivity and specificity of variant calling. We attempted to recreate the conditions and characteristics of the clinical specimens to strengthen the inference from actual samples. In Chapter III, I mixed two viruses passaged in cell culture and diluted in human stool-derived nucleic acids to simulate the template-rich environment in stool extracts. In Chapter IV, I mixed two synthetic RNA with known sequence. Across the three viruses studied, there was a consistent decrease in sensitivity for variants at lower frequencies. Although we could detect some variants below 1-2% frequency, the sensitivity was poor. To minimize the effect of low sensitivity and specificity on our results, we used fairly conservative frequency thresholds of 2-5% in each study. In future work, inclusion of defined mixtures and negative controls on each sequencing batch could help account for bias from sequencing run-specific artifacts and differences in read depth.

An important variable in these experiments was the input viral load. We and others have shown that low viral loads can dramatically decrease variant sensitivity and specificity for influenza virus and flaviviruses (Grubaugh et al., 2019a; McCrone and Lauring, 2016). Accounting for the variable viral loads and RNA quality in clinical specimens is crucial to avoid over-estimating the accuracy of a sequencing workflow. In my thesis, I extended this principle to polioviruses and SARS-CoV-2. The optimal cutoff for specimen viral load differs by the virus and sequencing approach, but other studies have suggested using a cutoff of 1000 genome copies of input RNA. The empirical cutoffs that I used in each chapter are all above this suggested threshold.

The wide variety of strategies for template enrichment, sequencing method, and bioinformatic analysis precludes simple adoption of the same variant calling criteria in every study. While it is impossible to obtain perfect sensitivity and specificity, estimates of these metrics across ranges of viral load, read depth, and technical replicates can generate expectations for the degree of shared variation due to sequence errors.

Estimation of transmission bottlenecks particularly depends on careful and accurate identification of shared minor variants. There are two notable cases when sequencing accuracy and interpretation of shared variation was a major fulcrum point for the overall conclusions. The first study to estimate the IAV transmission bottleneck in humans reported values of about 200 unique genomes passed between individuals (Poon et al., 2016). However, other studies estimated bottleneck sizes of 1-2 unique genomes (McCrone et al., 2018). In the first study, there was a large extent of shared variation between individuals, including individuals from different households. Between household pairs, variants at intermediate frequencies were rarely shared, while variants at low frequencies were shared. A careful re-analysis of the raw sequence reads revealed that paired end reads were improperly assigned to samples from different individuals, creating the appearance of greater shared variation than actually present (Xue and Bloom, 2019a). Bottleneck analysis with properly demultiplexed reads was consistent with other estimates of 1-2 unique genomes. Although this is an extreme example of how sequence errors can affect downstream results, it highlights the critical role that rigorous benchmarking and validation play in intrahost studies.

There have also been conflicting results between findings for SARS-CoV-2. Two high-profile preprints posted early in the pandemic found a large degree of shared variation and wide transmission bottlenecks. The first study reported as many as 400 intrahost variants per sample and bottleneck sizes of about 1000 unique genomes across transmission pairs in twelve households in Austria (Popa et al., 2020). The second study inferred the presence of wide transmission bottlenecks by virtue of extensive shared variation among samples in two geographic regions in the United Kingdom (Lythgoe et al., 2020b). However, later studies of household pairs as well as mammalian animal models suggested low intrahost diversity and narrow transmission bottlenecks of 1 – 10 unique genomes (Braun et al., 2021; Wang et al., 2020b). In late 2020, the preprint from groups in the United Kingdom was revised to use more conservative variant calling methods which resulted in opposite conclusions: low per-sample diversity of 0 – 5 minor variants and tight transmission bottlenecks across direct household pairs (Lythgoe et al., 2021). A recent re-analysis of the data from the household pairs in Austria revealed a strong dependence of the transmission bottleneck estimates on the frequency threshold used (Martin and Koelle, 2021). The most parsimonious explanation for these results is the presence of low-frequency variants across transmission pairs that resulted from errors in sequencing, due to variation in viral load or some other factor related to variant identification. Chapter IV of my thesis is consistent with the later findings of low intrahost diversity and narrow transmission bottlenecks. These cases demonstrate the difficulty of accurate intrahost variant detection and the kinds of controls that are required to avoid spurious conclusions.

Although the first studies on RNA virus transmission bottlenecks have provided important insights, there are many biological questions that remain. Transmission bottlenecks are not static values and probably exist as distributions that vary across different biological contexts. Transmission bottlenecks for a given virus may vary by transmission route, immune profile of the donor and recipient, and viral subtype. Understanding the molecular genetics of viral transmission in different populations could help uncover settings in which novel variants are more likely to emerge. Identification of these populations would have implications for infection prevention, genetic surveillance, and selection of influenza vaccine strains. Experiments in ferret models indicate that the soft palate may be an important anatomical site of influenza virus adaptation (Lakdawala et al., 2015). Further research in humans might explore the evolution of influenza viruses and SARS-CoV-

2 in the upper vs. lower respiratory tract and its relationship to the diversity of the transmitted population. Transmission dynamics may also differ by droplet vs. aerosol transmission, across gradients of ambient temperature and humidity, or in the setting of mixed infections. Future work should also evaluate the impact of host immune status on transmission bottlenecks. For example, vaccination of the donor or recipient host might constrain the replication of antigenic variants and decrease the effective transmission bottleneck size. More sophisticated methods of assessing immune function, such as mutational scanning in the presence of polyclonal sera or B cell receptor profiling, might offer mechanistic insight into how individual-level immunity shapes transmission dynamics. Various virus subtypes or lineages may also differ in their transmission dynamics. Influenza A subtype H3N2 evolves more quickly on a global scale compared to pdmH1N1 and influenza B viruses, but it is unclear whether new H3N2 variants emerge more easily on the within-host level or on larger population scales (Bedford et al., 2015; Morris et al., 2020). There is significant evidence that SARS-CoV-2 lineage B.1.1.7 is more transmissible than ancestral variants and sheds more viral particles on average (Volz et al., 2021). This is an important opportunity to evaluate the transmission bottleneck for two viral lineages that are phenotypically distinct but are otherwise similar. Lastly, it is unclear how the transmission bottleneck size, a population genetic measurement, relates to infectious dose, a functional measurement. These values are not necessarily correlated but are often conflated in the literature (Popa et al., 2020). For example, a virus may have a large infectious dose and few genetically distinct particles that establish the new infection. To parse out these distinctions would require technically difficult experiments in animal models and may be less valuable than defining transmission dynamics in humans.

Transmission bottlenecks are often assumed to act neutrally. Because bottlenecks involve a reduction of the effective population size, there will always be an increase in the strength of genetic drift in this process. However, it is possible that transmission bottlenecks between humans may also exert selective pressure. There is evidence for positive and purifying selection of genetic variants in studies of HIV and models of influenza virus (Carlson et al., 2014; Lakdawala et al., 2015). How and under what circumstances transmission bottlenecks act selectively are open questions. Distinguishing a neutral from selective bottleneck will probably be challenging and require analysis of many pairs. A simple approach might be to design a permutation test to identify sites or mutations that are overrepresented in recipients, accounting for their frequency in the donor. Minor variants that are selected during transmission might be observed more often in recipients than expected given their frequency and the bottleneck distribution. Other population genetic models have also been recently developed to address similar questions (Ghafari et al., 2020; Lumby et al., 2018). If evidence arises for selective bottlenecks during transmission between human hosts, it will be interesting to compare the relative importance of within and between-host selection on emergence of new variants in populations. Simple models suggest that a 1% selective advantage compared to wildtype virus has a much more powerful effect within hosts than between them (Bedford and Malik, 2016). However, a recent modeling study suggests that selection at the initial point of infection in a new host better explains the relatively frequent emergence of influenza virus antigenic variants despite weak within-host selection (Morris et al., 2020). I expect that further work combining human observational studies, genetic and epidemiologic models, and experimental systems will help untangle these complex dynamics.

Detecting natural selection with within-host sequencing data

In addition to understanding transmission dynamics, intrahost genetic diversity can also be leveraged to identify natural selection acting within hosts. Observing "real-world" evolution within hosts provides context for laboratory experiments that cannot always capture the direction and magnitude of selective pressures in humans. However, the time scale of intrahost studies is short by definition, which can obfuscate negative selection that has not yet had sufficient time to act. This creates barriers for standard evolutionary analyses such as dN/dS (Crandall et al., 1999), though similar approaches have been recently used to link within-host evolution and global evolutionary patterns (Xue and Bloom, 2020).

A preponderance of shared intrahost variation may suggest the presence of positive selection pressure. However, whether shared variation indicates selection or some other process depends on the size of the study population. In large datasets, it is conceivable that many variants shared across individuals could occur simply due to chance. On the other hand, detecting the same intrahost variant in a study consisting of only a handful of individuals suggests some degree of parallel selection, as was seen in a recent study of IAV in four immunocompromised individuals (Xue et al., 2017). It is not obvious how to determine which variants are biologically relevant and worthy of further investigation. My thesis demonstrates that there are no universally applicable criteria, but there are several potential strategies depending on the evolutionary context and study design.

The simplest method is to link variants enriched in human populations with previous experimental data, but there are limited data on fitness effects of single nucleotide mutations in RNA viruses. Massively multiplex assessments of variant effects are becoming more tractable for phenotypes like replication speed, neutralization by monoclonal and polyclonal antibodies, and drug resistance (Dingens et al., 2019; Doud et al., 2018; Lee et al., 2018). However, these data are laborious to generate, and their interpretation may be confounded by epistasis on varying genetic backgrounds.

In the absence of prior experimental data, we must rely on computational and statistical methods for determining which variants are influenced by positive selection. The primary method I used in my thesis was to compare the number of times a variant appeared in a study cohort to a null distribution. In Chapter II, there was only one iSNV shared across one individual, and it encoded a synonymous amino acid change. While synonymous variation can have fitness effects, it is more likely that this variant appeared twice merely by chance. Conversely, in Chapter III, I identified hundreds of shared mutations across the 83 people analyzed. There was a clear gradient in the strength of selection. Mutation A481G was found in nearly all individuals, followed by U2909C and U398C, all of which have known replication advantages. However, there were also many synonymous variants shared by only two people. To estimate a count cutoff for variants that could be shared due to chance alone, I performed permutation tests based on the observed level of intrahost diversity and the size of the study population. This analysis suggested that variants shared between two people could easily occur by chance, but variants shared by 3-4 or more people may have arisen due to positive selection pressure. Some of the mutations above this cutoff have weak associations with OPV genetic reversions in other phylogenetic studies, such as the non-synonymous capsid mutation A1997G (Famulare et al., 2016; Stern et al., 2017). Other mutations were found in many individuals but otherwise have no supporting experimental data. For example, G6084U is a synonymous variant in the 3D RdRP gene that is not present in any known RNA secondary structures. However, this variant is found in 3% of circulating VDPV genomes, making it conceivable that it has some fitness benefit.

Better models may help determine which mutations most likely have positive fitness effects. Factors like position in RNA secondary structures or type of amino acid change may suggest underlying biological mechanisms for an enriched mutation. More complicated permutation tests could also assess whether there is a pairwise association between two enriched mutations, indicative of an epistatic interaction. However, statistical tests of parallel mutation in phylogenetic analyses and permutation tests at the within-host level are ultimately limited in their statistical power and rely heavily on sampling density. With a smaller dataset, we might only have been able to reliably detect the gatekeeper mutations in multiple people. Permutation tests also depend on the degree of intrahost diversity in each sample. If there are few or no variants per sample, such as in influenza viruses or SARS-CoV-2, shared variants may reach statistical significance in permutation tests even if they are unlikely to be under positive selection.

Longitudinal samples can strengthen inference of positive selection on intrahost variants (Illingworth et al., 2020). If a mutation consistently rises in frequency within individuals, this is powerful evidence of selection. I observed this effect for A481G and U2909C within hosts that had 2 – 3 successfully sequenced samples. However, these mutations confer massive fitness benefits and equally strong selective sweeps are unlikely to occur during acute infections of endemic RNA viruses (Stern et al., 2017). For mutations with smaller fitness benefits, it becomes difficult to distinguish whether a rise in variant frequency is due to selection or genetic drift. This is an important limitation of the studies in my thesis. One aspect of future work will be to integrate estimates of effective population size in inferences of natural selection. For influenza viruses, effective population size may differ by host immune status. A study of influenza A virus in immunocompetent hosts estimated small population sizes and a dominant effect of genetic drift (McCrone et al., 2020). A recent study of influenza B virus in an immunocompromised host suggested larger population sizes (Lumby et al., 2020). These disparate estimates may be due to differences in methodology or acute vs. chronic infection.

Identification of genetic variants within hosts can provide grounds for subsequent experimental work. Specifically, investigation of the mutations in discussed Chapter III may shed light on the molecular mechanisms of OPV phenotypic reversion. I found six mutations that arose in > 10% of vaccine recipients, four of which encode non-synonymous substitutions in the capsid. These mutations likely confer a fitness benefit, but it might not be strong enough to be detected in phylogenetic analyses on limited cVDPV genomes (Famulare et al., 2016; Stern et al., 2017). An initial experiment could compare the change in replicative capacity in cell culture or organoid models with the level of enrichment in the human cohort. If these mutations enable faster within-host replication, we expect them to show intermediate fitness in cell culture compared to the gatekeeper mutations. Similar experiments could also assess epistatic interactions between these mutations and the gatekeepers. Mouse models could be used to quantify shedding and infectivity relative to ancestral OPV2 and the gatekeeper mutations. However, in the current phase of global polio eradication, wet lab experiments with type 2 poliovirus can only be performed in a small number of closely monitored laboratories. Nevertheless, my findings in Chapter III provide new avenues for investigators who are designing novel OPV2 vaccines (Konopka-Anstadt et al., 2020; Yeh et al., 2020).

More broadly, it will be beneficial to pair descriptive within-host analyses with "wet-lab" experiments in various viral systems. These joint approaches could clarify how often antiviral resistance or antibody escape occurs within acute infections for viruses like influenza and SARS-CoV-2. For example, sequencing approaches that use unique molecular identifiers (UMI) could be used to accurately identify intrahost variants down to very low frequencies (0.1%) in specific genome areas of interest, such as antigenic epitopes on influenza virus HA/NA or the SARS-CoV-2 spike protein. If this was applied across a large cohort, there might be more statistical power for permutation tests to identify sites that are recurrently mutated within-hosts. Then these mutations could be assayed for escape from antibody neutralization by patient sera or shedding duration in animal models. Mutations identified in viral polymerases could be evaluated for effects on replication speed or mutation rate. Even if these low-frequency mutations do not spread onwards due to transmission bottlenecks or other constraints, they could still identify sites that experience selection at the within-host level. This kind of data could inform genomic surveillance of respiratory viruses and inform the design of yearly influenza vaccines.

Impact of infection duration on variant spread

So far, I have discussed considerations for accurately detecting intrahost variants and their downstream effects on measuring transmission bottlenecks and inferring within-host natural selection. However, to understand how within-host dynamics translate into global-scale evolutionary patterns, we need to consider these processes in the context of infection kinetics. Narrow bottleneck estimates suggest a model where beneficial variants can only be transmitted if they arise early enough during an acute infection to reach intermediate or high frequencies before a transmission event occurs (Figure 5.1B). If they arise too late, they will not be present at sufficient frequency to be reliably transmitted. This model suggests that infections with longer duration could enable more viral evolutionary change and play a larger role in global population dynamics.

Chapter III of my thesis illustrates this concept in detail. In primary OPV2 recipients, there was strong selective pressure to drive the gatekeeper mutations to fixation in most people within several weeks after vaccination. Poliovirus generally sheds from susceptible hosts for 4-6 weeks, providing ample time for natural selection to act at the within-host level (Famulare et al., 2018). However, most household transmission events occurred within 1 – 2 weeks after vaccination, before mutations U2909C and U398C had reached high frequencies (Taniuchi et al., 2017). Therefore, the gatekeeper mutations were transmitted to household contacts in a time-dependent manner: only A481G was consistently identified in household contacts, while U2909C was more likely to be detected three or more weeks after vaccination. There is also some evidence for a similar dynamic in influenza virus, which can shed longer in young children (Han et al., 2021).

These dynamics can also help us form expectations for how often SARS-CoV-2 variants will emerge at the individual level. To evaluate the risk of within-host variants spreading to new hosts, we must consider the interval of infectious viral shedding. It is important to remember that this interval is different than the duration of detecting viral RNA by RT-qPCR. For SARS-CoV-2, infected individuals often transmit before developing symptoms, and then are infectious for about five days thereafter (He et al., 2020). In Chapter IV, I showed that intrahost diversity remains low over the first several days after symptom onset. Although we detected genetic variants in individuals who were three weeks into their infections, variants present at this time are unlikely to transmit. Within the span of plausible transmissibility, there were very few mutations present at levels that would survive a narrow bottleneck. A recent study highlights how viral kinetics provides important context for interpreting within-host variation. In this study, investigators used unique molecular identifiers to measure diversity in the spike protein at high resolution from SARS-CoV-2 clinical specimens (Ko et al., 2021). They detected a cluster of mutations in the spike protein receptor binding domain in one individual. This mutation cluster was detected at low frequency on illness day 11, with a transient increase to 25% on illness day 15 before ultimately being lost from the population. While this study suggests that antigenic variants can arise within hosts after seroconversion, these mutations did not arise until several days after the standard window of transmission. Even variants at a frequency of 25% are not guaranteed to transmit through narrow bottlenecks (Braun et al., 2021; Lythgoe et al., 2021; Martin and Koelle, 2021).

The conflict between selection, infectious shedding duration, and transmission bottlenecks results in inefficient spread of variants from any single individual. However, when the number of infections is large, even rare events can happen. For most of the SARS-CoV-2 pandemic, there have been few instances of viral variants arising that have a measurable fitness advantage (Lauring and Hodcroft, 2021; Volz et al., 2021). As more infections have occurred over time, the global population of SARS-CoV-2 viruses has had more opportunities to generate novel combinations of variants within hosts in the timeframe of potential transmission. Therefore, the recent emergence of variants of concern (VOC, i.e. B.1.1.7, B.1.351, and P.1) probably involved a convergence of rare events.

A related dynamic is at play in the emergence of cVDPV. Compared to the number of OPV doses administered every year, the number of cVDPV emergence events is actually rather small (Famulare et al., 2018). There are probably both epidemiologic and evolutionary factors that constrain cVDPV emergence. A narrow transmission bottleneck in OPV could introduce heterogeneity into the spread of gatekeeper mutations in populations. At a population scale, this would effectively reduce the number of OPV viruses that acquire all three gatekeeper mutations and other mutations that are involved in its phenotypic reversion. However, it is unknown to what degree these evolutionary dynamics influence cVDPV emergence compared to epidemiologic variables, such as availability of susceptible host networks (Famulare et al., 2020). Recent modeling studies suggest that the structure of community social networks and population immunity may play a larger role than viral evolution in cVDPV emergence (Wong et al., 2020). My work in Chapter III provides key insights into the evolutionary dynamics of OPV that could be integrated into such models of cVDPV emergence that combine viral genetics and host epidemiology.

Infections in immunocompromised hosts are the most extreme cases of prolonged shedding, but it is not clear how often they lead to subsequent outbreaks. Infections of SARS-CoV-2 in immunocompromised hosts have been noted in case reports to shed infectious particles for months (Avanzato et al., 2020; Baang et al., 2021; Choi et al., 2020; Kemp et al., 2020). Some, but not all, of these cases have exhibited within-host evolution at genomic loci implicated in VOC B.1.1.7. It is therefore a reasonable suggestion that the origins of B.1.1.7, have some association with an immunocompromised host(s). This will necessarily remain a point of reasonable speculation. However, other RNA viruses like norovirus and poliovirus also establish long-term infections in immunocompromised hosts, and there is no evidence yet that they shed transmissible particles and contribute to viral epidemics or emergence of new strains (Bok et al., 2016; Dunn et al., 2015). Therefore, it is presently unclear how much infections in immunocompromised hosts contribute to global viral evolution. It is important to define the clinical characteristics of immunocompromised hosts that associate with prolonged infectious shedding, the phenotypic changes in viruses during long-term infections, and the extent to which these infections contribute to epidemics in immunocompetent populations.

Implications for sequence-based transmission inference

My thesis has implications for the feasibility of using within-host variants to enhance current efforts of sequence-based transmission inference and network reconstruction. There is enthusiasm for this approach in the SARS-CoV-2 literature, but many studies do not consider how within-host evolutionary dynamics impact these analyses (Lau et al., 2020; Sapoval et al., 2020). Chapter IV displays multiple obstacles for the productive use of intrahost variants for making transmission inferences. First, I showed that detecting viral variants is highly error prone. If studies fail to account for these sources of error, it could generate many false positive transmission links. Second, I showed that most infections exhibit little genetic variation during the interval of highest transmissibility, and those variants that are present have generally very low frequencies. Third, I showed that the same genetic variants can arise in parallel across unrelated hosts which could also cause false positive transmission links. Lastly, I discussed these findings in the context of a narrow transmission bottleneck. If the bottleneck is small, then it is conceivably more likely for variants to be shared by individuals not linked by transmission than direct transmission pairs (Martin and Koelle, 2021; Worby et al., 2017).

Therefore, in my view it is not advisable to compare intrahost variants by hand to assign transmission linkages without strong epidemiologic evidence (Tonkin-Hill et al., 2020). However, there are comprehensive statistical frameworks that have shown intrahost variation to sometimes be of use, even when limited by tight bottlenecks (Maio et al., 2018). However, these models have not fully accounted for parallel evolution. An important focus of future work will be to quantitatively estimate the rate of parallel evolution of variants in similar genetic backgrounds and compare that to the likelihood of direct minor variant transmission. Over time, as SARS-CoV-2 lineages continue to diversify, it may become more difficult to obtain large enough datasets with genetically similar viruses to accomplish robust estimates. However, greater consensus diversity in circulating viruses will also make it easier to separate unrelated case clusters based on genetic distance. Recent work has also suggested that there are mutational "hotspots" in the SARS-CoV-2 genome (Tonkin-Hill et al., 2020). High-throughput mutational profiling methods could help determine whether these sites are more tolerant to mutations than other sites, and therefore more likely to arise within hosts in parallel.

Given the enthusiasm around using advanced genetic technologies for "precision epidemiology," it is important to be aware of common pitfalls and areas where technology can have larger impact. I suggest that the potential gains of using intrahost variation in transmission inference are outweighed by the logistical and biological obstacles. In many scenarios, it is not necessary or feasible to establish exactly "who infected whom." Consensus genome sequencing, while relatively coarse, can still provide important insights into broad patterns of viral transmission. Intrahost sequencing has more value when applied on a limited basis, focused on well-designed observational studies with rich clinical metadata, clearly defined viral kinetics, and simpler contact networks. These studies are the most likely to reveal the subtleties of viral evolutionary dynamics across different populations.

BIBLIOGRAPHY

Acevedo, A., Brodsky, L., and Andino, R. (2014). Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature *505*, 686–690.

Alam, N., Ali, T., Razzaque, A., Rahman, M., Zahirul Haq, M., Saha, S.K., Ahmed, A., Sarder, A.M., Moinuddin Haider, M., Yunus, M., et al. (2017). Health and Demographic Surveillance System (HDSS) in Matlab, Bangladesh. Int. J. Epidemiol. *46*, 809–816.

Avanzato, V.A., Matson, M.J., Seifert, S.N., Pryce, R., Williamson, B.N., Anzick, S.L., Barbian, K., Judson, S.D., Fischer, E.R., Martens, C., et al. (2020). Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. Cell *183*, 1901-1912.e9.

Avanzino, B.C., Jue, H., Miller, C.M., Cheung, E., Fuchs, G., and Fraser, C.S. (2018). Molecular mechanism of poliovirus Sabin vaccine strain attenuation. J. Biol. Chem. jbc.RA118.004913.

Baang, J.H., Smith, C., Mirabelli, C., Valesano, A.L., Manthei, D.M., Bachman, M.A., Wobus, C.E., Adams, M., Washer, L., Martin, E.T., et al. (2021). Prolonged Severe Acute Respiratory Syndrome Coronavirus 2 Replication in an Immunocompromised Patient. J. Infect. Dis. *223*, 23–27.

Bedford, T., and Malik, H.S. (2016). Did a Single Amino Acid Change Make Ebola Virus More Virulent? Cell *167*, 892–894.

Bedford, T., Riley, S., Barr, I.G., Broor, S., Chadha, M., Cox, N.J., Daniels, R.S., Gunasekaran, C.P., Hurt, A.C., Kelso, A., et al. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature *523*, 217–220.

Blake, I.M., Pons-Salort, M., Molodecky, N.A., Diop, O.M., Chenoweth, P., Bandyopadhyay, A.S., Zaffran, M., Sutter, R.W., and Grassly, N.C. (2018). Type 2 Poliovirus Detection after Global Withdrawal of Trivalent Oral Vaccine. N. Engl. J. Med.

Bok, K., Prevots, D.R., Binder, A.M., Parra, G.I., Strollo, S., Fahle, G.A., Behrle-Yardley, A., Johnson, J.A., Levenson, E.A., Sosnovtsev, S.V., et al. (2016). Epidemiology of Norovirus Infection Among Immunocompromised Patients at a Tertiary Care Research Hospital, 2010–2013. Open Forum Infect. Dis. *3*.

Boot, H.J., Sonsma, J., van Nunen, F., Abbink, F., Kimman, T.G., and Buisman, A.-M. (2007). Determinants of monovalent oral polio vaccine mutagenesis in vaccinated elderly people. Vaccine *25*, 4706–4714.

Bowers, R.M., Clum, A., Tice, H., Lim, J., Singh, K., Ciobanu, D., Ngan, C.Y., Cheng, J.-F., Tringe, S.G., and Woyke, T. (2015). Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics *16*, 1–12.

Braun, K.M., Moreno, G.K., Halfmann, P.J., Baker, D.A., Boehm, E.C., Weiler, A.M., Haj, A.K., Hatta, M., Chiba, S., Maemura, T., et al. (2020). Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck. BioRxiv 2020.11.16.384917.

Braun, K.M., Moreno, G.K., Halfmann, P.J., Hodcroft, E.B., Baker, D.A., Boehm, E.C., Weiler, A.M., Haj, A.K., Hatta, M., Chiba, S., et al. (2021). Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck. PLOS Pathog. *17*, e1009373.

Buckland, M.S., Galloway, J.B., Fhogartaigh, C.N., Meredith, L., Provine, N.M., Bloor, S., Ogbe, A., Zelek, W.M., Smielewska, A., Yakovleva, A., et al. (2020). Treatment of COVID-19 with remdesivir in the absence of humoral immunity: a case report. Nat. Commun. *11*, 6385.

Bull, J.J., Smithson, M.W., and Nuismer, S.L. (2018). Transmissible Viral Vaccines. Trends Microbiol. *26*, 6–15.

Burns, C.C., Shaw, J., Jorba, J., Bukbuk, D., Adu, F., Gumede, N., Pate, M.A., Abanida, E.A., Gasasira, A., Iber, J., et al. (2013). Multiple Independent Emergences of Type 2 Vaccine-Derived Polioviruses during a Large Outbreak in Northern Nigeria. J. Virol. *87*, 4907–4922.

Burns, C.C., Diop, O.M., Sutter, R.W., and Kew, O.M. (2014). Vaccine-Derived Polioviruses. J. Infect. Dis. *210*, S283–S293.

Carlson, J.M., Schaefer, M., Monaco, D.C., Batorsky, R., Claiborne, D.T., Prince, J., Deymier, M.J., Ende, Z.S., Klatt, N.R., DeZiel, C.E., et al. (2014). Selection bias at the heterosexual HIV-1 transmission bottleneck. Science *345*, 1254031.

Chen, R., and Holmes, E.C. (2008). The Evolutionary Dynamics of Human Influenza B Virus. J. Mol. Evol. *66*, 655.

Choi, B., Choudhary, M.C., Regan, J., Sparks, J.A., Padera, R.F., Qiu, X., Solomon, I.H., Kuo, H.-H., Boucau, J., Bowman, K., et al. (2020). Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. N. Engl. J. Med. *383*, 2291–2293.

Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, H.C., and Salzman, N.P. (1999). Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. *16*, 372–382.

Cudini, J., Roy, S., Houldcroft, C.J., Bryant, J.M., Depledge, D.P., Tutill, H., Veys, P., Williams, R., Worth, A.J.J., Tamuri, A.U., et al. (2019). Human cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and evidence of within-host recombination. Proc. Natl. Acad. Sci. *116*, 5693–5698.

Debbink, K., McCrone, J.T., Petrie, J.G., Truscon, R., Johnson, E., Mantlo, E.K., Monto, A.S., and Lauring, A.S. (2017). Vaccination has minimal impact on the intrahost diversity of H3N2 influenza viruses. PLOS Pathog. *13*, e1006194.

Dedepsidis, E., Karakasiliotis, I., Paximadi, E., Kyriakopoulou, Z., Komiotis, D., and Markoulatos, P. (2006). Detection of unusual mutation within the VP1 region of different re-isolates of poliovirus Sabin vaccine. Virus Genes *33*, 183–191.

Di Paola, N., Sanchez-Lockhart, M., Zeng, X., Kuhn, J.H., and Palacios, G. (2020). Viral genomics in Ebola virus research. Nat. Rev. Microbiol. *18*, 365–378.

Dimcheff, D.E., Valesano, A.L., Rumfelt, K.E., Fitzsimmons, W.J., Blair, C., Mirabelli, C., Petrie, J.G., Martin, E.T., Bhambhani, C., Tewari, M., et al. (2021). SARS-CoV-2 Total and Subgenomic RNA Viral Load in Hospitalized Patients. MedRxiv 2021.02.25.21252493.

Dingens, A.S., Arenz, D., Overbaugh, J., and Bloom, J.D. (2019). Massively Parallel Profiling of HIV-1 Resistance to the Fusion Inhibitor Enfuvirtide. Viruses *11*, 439.

Dinis, J.M., Florek, N.W., Fatola, O.O., Moncla, L.H., Mutschler, J.P., Charlier, O.K., Meece, J.K., Belongia, E.A., and Friedrich, T.C. (2016). Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans. J. Virol. *90*, 3355–3365.

Dolan, P.T., Whitfield, Z.J., and Andino, R. (2016). Mechanisms and Concepts in RNA Virus Population Dynamics and Evolution. Annu. Rev. Virol.

Dolan, P.T., Whitfield, Z.J., and Andino, R. (2018). Mapping the Evolutionary Potential of RNA Viruses. Cell Host Microbe *23*, 435–446.

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. *83*, 104351.

Doud, M.B., Lee, J.M., and Bloom, J.D. (2018). How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin. Nat. Commun. *9*, 1386.

Dudas, G., Bedford, T., Lycett, S., and Rambaut, A. (2015). Reassortment between Influenza B Lineages and the Emergence of a Coadapted PB1–PB2–HA Gene Complex. Mol. Biol. Evol. *32*, 162–172.

Dunn, G., Klapsa, D., Wilton, T., Stone, L., Minor, P.D., and Martin, J. (2015). Twenty-Eight Years of Poliovirus Replication in an Immunodeficient Individual: Impact on the Global Polio Eradication Initiative. PLOS Pathog. *11*, e1005114.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Elena, S.F., Sanjuán, R., Bordería, A.V., and Turner, P.E. (2001). Transmission bottlenecks and the evolution of fitness in rapidly evolving RNA viruses. Infect. Genet. Evol. *1*, 41–48.

Escalera-Zamudio, M., Golden, M., Gutiérrez, B., Thézé, J., Keown, J.R., Carrique, L., Bowden, T.A., and Pybus, O.G. (2020). Parallel evolution in the emergence of highly pathogenic avian influenza A viruses. Nat. Commun. *11*, 5511.

Famulare, M., Chang, S., Iber, J., Zhao, K., Adeniji, J.A., Bukbuk, D., Baba, M., Behrend, M., Burns, C.C., and Oberste, M.S. (2016). Sabin Vaccine Reversion in the Field: a Comprehensive Analysis of Sabin-Like Poliovirus Isolates in Nigeria. J. Virol. *90*, 317–331.

Famulare, M., Selinger, C., McCarthy, K.A., Eckhoff, P.A., and Chabot-Couture, G. (2018). Assessing the stability of polio eradication after the withdrawal of oral polio vaccine. PLOS Biol. *16*, e2002468.

Famulare, M., Wong, W., Haque, R., Platts-Mills, J.A., Saha, P., Aziz, A.B., Ahmed, T., Islam, M.O., Uddin, M.J., Bandyopadhyay, A.S., et al. (2020). Community structure mediates Sabin 2 polio vaccine virus transmission. MedRxiv 2020.07.01.20144501.

Fauver, J.R., Petrone, M.E., Hodcroft, E.B., Shioda, K., Ehrlich, H.Y., Watts, A.G., Vogels, C.B.F., Brito, A.F., Alpert, T., Muyombwe, A., et al. (2020). Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. Cell *181*, 990-996.e5.

Fitzsimmons, W.J., Woods, R.J., McCrone, J.T., Woodman, A., Arnold, J.J., Yennawar, M., Evans, R., Cameron, C.E., and Lauring, A.S. (2018). A speed–fidelity trade-off determines the mutation rate and virulence of an RNA virus. PLOS Biol. *16*, e2006459.

Geoghegan, J.L., and Holmes, E.C. (2018). Evolutionary Virology at 40. Genetics *210*, 1151–1162.

Geoghegan, J.L., Ren, X., Storey, M., Hadfield, J., Jelley, L., Jefferies, S., Sherwood, J., Paine, S., Huang, S., Douglas, J., et al. (2020). Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. Nat. Commun. *11*, 6351.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nat. Commun. *3*, 811.

Ghafari, M., Lumby, C.K., Weissman, D.B., and Illingworth, C.J.R. (2020). Inferring Transmission Bottleneck Size from Viral Sequence Data Using a Novel Haplotype Reconstruction Method. J. Virol. *94*.

Grubaugh, N.D., Ladner, J.T., Kraemer, M.U.G., Dudas, G., Tan, A.L., Gangavarapu, K., Wiley, M.R., White, S., Thézé, J., Magnani, D.M., et al. (2017). Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature *546*, 401–405.

Grubaugh, N.D., Gangavarapu, K., Quick, J., Matteson, N.L., De Jesus, J.G., Main, B.J., Tan, A.L., Paul, L.M., Brackney, D.E., Grewal, S., et al. (2019a). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. *20*, 8.

Grubaugh, N.D., Ladner, J.T., Lemey, P., Pybus, O.G., Rambaut, A., Holmes, E.C., and Andersen, K.G. (2019b). Tracking virus outbreaks in the twenty-first century. Nat. Microbiol. *4*, 10.

Gutierrez, B., Escalera-Zamudio, M., and Pybus, O.G. (2019). Parallel molecular evolution and adaptation in viruses. Curr. Opin. Virol. *34*, 90–96.

Han, A.X., Maurer-Stroh, S., and Russell, C.A. (2018). Individual immune selection pressure has limited impact on seasonal influenza virus evolution. Nat. Ecol. Evol. 1.

Han, A.X., Garza, Z.C.F., Welkers, M.R.A., Vigeveno, R.M., Duong, T.N., Mai, L.T.Q., Thai, P.Q., Thoang, D.D., Anh, T.T.N., Tuan, H.M., et al. (2021). Within-host evolutionary dynamics of seasonal and pandemic human influenza A viruses in young children. BioRxiv 2021.03.20.436248.

He, X., Lau, E.H.Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y.C., Wong, J.Y., Guan, Y., Tan, X., et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat. Med. 1–4.

Hedestam, G.B.K., Fouchier, R.A.M., Phogat, S., Burton, D.R., Sodroski, J., and Wyatt, R.T. (2008). The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. Nat. Rev. Microbiol. *6*, 143–155.

Hodcroft, E.B., Domman, D.B., Snyder, D.J., Oguntuyo, K.Y., Diest, M.V., Densmore, K.H., Schwalm, K.C., Femling, J., Carroll, J.L., Scott, R.S., et al. (2021). Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting amino acid position 677. MedRxiv 2021.02.12.21251658.

Holubar, M., Sahoo, M.K., Huang, C., Mohamed-Hadley, A., Liu, Y., Waggoner, J.J., Troy, S.B., García-García, L., Ferreyra-Reyes, L., Maldonado, Y., et al. (2019). Deep sequencing prompts the modification of a real-time RT-PCR for the serotype-specific detection of polioviruses. J. Virol. Methods *264*, 38–43.

Illingworth, C.J.R., Raghwani, J., Serwadda, D., Sewankambo, N.K., Robb, M.L., Eller, M.A., Redd, A.R., Quinn, T.C., and Lythgoe, K.A. (2020). A de novo approach to inferring within-host fitness effects during untreated HIV-1 infection. PLOS Pathog. *16*, e1008171.

Iyengar, P., von Mollendorf, C., Tempia, S., Moerdyk, A., Valley-Omar, Z., Hellferscee, O., Martinson, N., Chhagan, M., McMorrow, M., Gambhir, M., et al. (2015). Case-ascertained study of household transmission of seasonal influenza — South Africa, 2013. J. Infect. *71*, 578–586.

Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A., and Swanstrom, R. (2011). Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc. Natl. Acad. Sci. *108*, 20166–20171.

James, S.E., Ngcapu, S., Kanzi, A.M., Tegally, H., Fonseca, V., Giandhari, J., Wilkinson, E., Chimukangara, B., Pillay, S., Singh, L., et al. (2020). High Resolution analysis of Transmission Dynamics of Sars-Cov-2 in Two Major Hospital Outbreaks in South Africa Leveraging Intrahost Diversity. MedRxiv 2020.11.15.20231993.

Jenkins, H.E., Aylward, R.B., Gasasira, A., Donnelly, C.A., Mwanza, M., Corander, J., Garnier, S., Chauvin, C., Abanida, E., Pate, M.A., et al. (2010). Implications of a Circulating Vaccine-Derived Poliovirus in Nigeria. N. Engl. J. Med. *362*, 2360–2369.

Jorba, J., Campagnoli, R., De, L., and Kew, O. (2008). Calibration of Multiple Poliovirus Molecular Clocks Covering an Extended Evolutionary Range. J. Virol. *82*, 4429–4440.

Kariuki, S.M., Selhorst, P., Ariën, K.K., and Dorfman, J.R. (2017). The HIV-1 transmission bottleneck. Retrovirology *14*, 22.

Kemp, S., Harvey, W., Datir, R., Collier, D., Ferreira, I., Carabelii, A., Robertson, D.L., and Gupta, R.K. (2020). Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70. BioRxiv 2020.12.14.422555.

Kew, O., and Pallansch, M. (2018). Breaking the Last Chains of Poliovirus Transmission: Progress and Challenges in Global Polio Eradication. Annu. Rev. Virol. *5*, null.

Kew, O., Morris-Glasgow, V., Landaverde, M., Burns, C., Shaw, J., Garib, Z., André, J., Blackman, E., Freeman, C.J., Jorba, J., et al. (2002). Outbreak of Poliomyelitis in Hispaniola Associated with Circulating Type 1 Vaccine-Derived Poliovirus. Science *296*, 356–359.

Kew, O.M., Sutter, R.W., de Gourville, E.M., Dowdle, W.R., and Pallansch, M.A. (2005). Vaccine-Derived Polioviruses and the Endgame Strategy for Global Polio Eradication. Annu. Rev. Microbiol. *59*, 587–635.

Kilpatrick, D.R., Iber, J.C., Chen, Q., Ching, K., Yang, S.-J., De, L., Mandelbaum, M.D., Emery, B., Campagnoli, R., Burns, C.C., et al. (2011). Poliovirus serotype-specific VP1 sequencing primers. J. Virol. Methods *174*, 128–130.

Ko, S.H., Mokhtari, E.B., Mudvari, P., Stein, S., Stringham, C.D., Wagner, D., Ramelli, S., Ramos-Benitez, M.J., Strich, J.R., Jr, R.T.D., et al. (2021). High-throughput, single-copy sequencing reveals SARS-CoV-2 spike variants coincident with mounting humoral immunity during acute COVID-19. PLOS Pathog. *17*, e1009431.

Konopka-Anstadt, J.L., Campagnoli, R., Vincent, A., Shaw, J., Wei, L., Wynn, N.T., Smithee, S.E., Bujaki, E., Te Yeh, M., Laassri, M., et al. (2020). Development of a new oral poliovirus vaccine for the eradication end game using codon deoptimization. Npj Vaccines *5*, 1–9.

Kryazhimskiy, S., and Plotkin, J.B. (2008). The Population Genetics of dN/dS. PLOS Genet. *4*, e1000304.

Lakdawala, S.S., Jayaraman, A., Halpin, R.A., Lamirande, E.W., Shih, A.R., Stockwell, T.B., Lin, X., Simenauer, A., Hanson, C.T., Vogel, L., et al. (2015). The soft palate is an important site of adaptation for transmissible influenza viruses. Nature *526*, 122–125.

Langat, P., Raghwani, J., Dudas, G., Bowden, T.A., Edwards, S., Gall, A., Bedford, T., Rambaut, A., Daniels, R.S., Russell, C.A., et al. (2017). Genome-wide evolutionary dynamics of influenza B viruses on a global scale. PLOS Pathog. *13*, e1006749.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lau, B.T., Pavlichin, D., Hooker, A.C., Almeda, A., Shin, G., Chen, J., Sahoo, M.K., Huang, C., Pinsky, B.A., Lee, H., et al. (2020). Profiling SARS-CoV-2 mutation fingerprints that range from the viral pangenome to individual infection quasispecies. MedRxiv 2020.11.02.20224816.

Lauring, A.S. (2020). Within-Host Viral Diversity: A Window into Viral Evolution. Annu. Rev. Virol.

Lauring, A.S., and Hodcroft, E.B. (2021). Genetic Variants of SARS-CoV-2—What Do They Mean? JAMA *325*, 529.

Lee, J.M., Huddleston, J., Doud, M.B., Hooper, K.A., Wu, N.C., Bedford, T., and Bloom, J.D. (2018). Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proc. Natl. Acad. Sci. *115*, E8276–E8285.

Lee, J.M., Eguia, R., Zost, S.J., Choudhary, S., Wilson, P.C., Bedford, T., Stevens-Ayers, T., Boeckh, M., Hurt, A.C., Lakdawala, S.S., et al. (2019). Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. ELife *8*, e49324.

Leonard, A.S., McClain, M.T., Smith, G.J.D., Wentworth, D.E., Halpin, R.A., Lin, X., Ransier, A., Stockwell, T.B., Das, S.R., Gilbert, A.S., et al. (2016). Deep Sequencing of Influenza A Virus from a Human Challenge Study Reveals a Selective Bottleneck and Only Limited Intrahost Genetic Diversification. J. Virol. *90*, 11247–11258.

Leonard, A.S., Weissman, D.B., Greenbaum, B., Ghedin, E., and Koelle, K. (2017). Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. J. Virol. *91*, e00171-17.

Lequime, S., Fontaine, A., Ar Gouilh, M., Moltini-Conclois, I., and Lambrechts, L. (2016). Genetic Drift, Purifying Selection and Vector Genotype Shape Dengue Virus Intra-host Genetic Diversity in Mosquitoes. PLoS Genet. *12*.

Lévêque, N., and Semler, B.L. (2015). A 21st Century Perspective of Poliovirus Replication. PLOS Pathog. *11*, e1004825.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lumby, C.K., Nene, N.R., and Illingworth, C.J.R. (2018). A novel framework for inferring parameters of transmission from viral sequence data. PLOS Genet. *14*, e1007718.

Lumby, C.K., Zhao, L., Breuer, J., and Illingworth, C.J. (2020). A large effective population size for established within-host influenza virus infection. ELife *9*, e56915.

Lythgoe, K.A., Hall, M., Ferretti, L., Cesare, M. de, MacIntyre-Cockett, G., Trebes, A., Andersson, M., Otecko, N., Wise, E.L., Moore, N., et al. (2020a). Within-host genomics of SARS-CoV-2. BioRxiv 2020.05.28.118992.

Lythgoe, K.A., Hall, M., Ferretti, L., Cesare, M. de, MacIntyre-Cockett, G., Trebes, A., Andersson, M., Otecko, N., Wise, E.L., Moore, N., et al. (2020b). Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. BioRxiv 2020.05.28.118992.

Lythgoe, K.A., Hall, M., Ferretti, L., Cesare, M. de, MacIntyre-Cockett, G., Trebes, A., Andersson, M., Otecko, N., Wise, E.L., Moore, N., et al. (2021). SARS-CoV-2 within-host diversity and transmission. Science *372*.

Macadam, A.I., Pollard, S.R., Ferguson, G., Dunn, G., Skuce, R., Almond, J.W., and Minor, P.D. (1991). The 5' noncoding region of the type 2 poliovirus vaccine strain contains determinants of attenuation and temperature sensitivity. Virology *181*, 451–458.

Macadam, A.J., Pollard, S.R., Ferguson, G., Skuce, R., Wood, D., Almond, J.W., and Minor, P.D. (1993). Genetic Basis of Attenuation of the Sabin Type 2 Vaccine Strain of Poliovirus in Primates. Virology *192*, 18–26.

Macklin, G.R., O'Reilly, K.M., Grassly, N.C., Edmunds, W.J., Mach, O., Krishnan, R.S.G., Voorman, A., Vertefeuille, J.F., Abdelwahab, J., Gumede, N., et al. (2020). Evolving epidemiology of poliovirus serotype 2 following withdrawal of the type 2 oral poliovirus vaccine. Science.

Maio, N.D., Worby, C.J., Wilson, D.J., and Stoesser, N. (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. PLOS Comput. Biol. *14*, e1006117.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12.

Martin, M.A., and Koelle, K. (2021). Reanalysis of deep-sequencing data from Austria points towards a small SARS-COV-2 transmission bottleneck on the order of one to three virions. BioRxiv 2021.02.22.432096.

Martin, M.A., Lee, R.S., Cowley, L.A., Gardy, J.L., and Hanage, W.P. (2018). Within-host Mycobacterium tuberculosis diversity and its utility for inferences of transmission. Microb. Genomics *4*, e000217.

McCrone, J.T., and Lauring, A.S. (2016). Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. J. Virol. *90*, 6884–6895.

McCrone, J.T., and Lauring, A.S. (2018). Genetic bottlenecks in intraspecies virus transmission. Curr. Opin. Virol. *28*, 20–25.

McCrone, J.T., Woods, R.J., Martin, E.T., Malosh, R.E., Monto, A.S., and Lauring, A.S. (2018). Stochastic processes constrain the within and between host evolution of influenza virus. ELife *7*, e35962.

McCrone, J.T., Woods, R.J., Monto, A.S., Martin, E.T., and Lauring, A.S. (2020). The effective population size and mutation rate of influenza A virus in acutely infected individuals. BioRxiv 2020.10.24.353748.

McWhite, C.D., Meyer, A.G., and Wilke, C.O. (2016). Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin. Virus Evol. *2*.

Meredith, L.W., Hamilton, W.L., Warne, B., Houldcroft, C.J., Hosmillo, M., Jahun, A.S., Curran, M.D., Parmar, S., Caller, L.G., Caddy, S.L., et al. (2020). Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect. Dis. *0*.

Miller, D., Martin, M.A., Harel, N., Tirosh, O., Kustin, T., Meir, M., Sorek, N., Gefen-Halevi, S., Amit, S., Vorontsov, O., et al. (2020). Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nat. Commun. *11*, 5518.

Moncla, L.H., Bedford, T., Dussart, P., Horm, S.V., Rith, S., Buchy, P., Karlsson, E.A., Li, L., Liu, Y., Zhu, H., et al. (2020). Quantifying within-host diversity of H5N1 influenza viruses in humans and poultry in Cambodia. PLOS Pathog. *16*, e1008191.

Montmayeur, A.M., Ng, T.F.F., Schmidt, A., Zhao, K., Magaña, L., Iber, J., Castro, C.J., Chen, Q., Henderson, E., Ramos, E., et al. (2017). High-Throughput Next-Generation Sequencing of Polioviruses. J. Clin. Microbiol. *55*, 606–615.

Monto, A.S., Malosh, R.E., Petrie, J.G., Thompson, M.G., and Ohmit, S.E. (2014). Frequency of Acute Respiratory Illnesses and Circulation of Respiratory Viruses in Households With Children Over 3 Surveillance Seasons. J. Infect. Dis. *210*, 1792–1799.

Monto, A.S., Malosh, R.E., Evans, R., Lauring, A.S., Gordon, A., Thompson, M.G., Fry, A.M., Flannery, B., Ohmit, S.E., Petrie, J.G., et al. (2019). Data resource profile: Household Influenza Vaccine Evaluation (HIVE) Study. Int. J. Epidemiol. *48*, 1040–1040g.

Moreno, G.K., Braun, K.M., Halfmann, P.J., Prall, T.M., Riemersma, K.K., Haj, A.K., Lalli, J., Florek, K.R., Kawaoka, Y., Friedrich, T.C., et al. (2020). Limited SARS-CoV-2 diversity within hosts and following passage in cell culture. BioRxiv 2020.04.20.051011.

Morris, D.H., Petrova, V.N., Rossine, F.W., Parker, E., Grenfell, B.T., Neher, R.A., Levin, S.A., and Russell, C.A. (2020). Asynchrony between virus diversity and antibody selection limits influenza virus evolution. ELife *9*.

Moya, A., Holmes, E.C., and González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. *2*, 279–288.

Munnink, B.B.O., Nieuwenhuijse, D.F., Stein, M., O'Toole, Á., Haverkate, M., Mollers, M., Kamga, S.K., Schapendonk, C., Pronk, M., Lexmond, P., et al. (2020). Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat. Med. 1–6.

Muzychenko, A.R., Lipskaya, G.Yu., Maslova, S.V., Svitkin, Y.V., Pilipenko, E.V., Nottay, B.K., Kew, O.M., and Agol, V.I. (1991). Coupled mutations in the 5'-untranslated region of the Sabin poliovirus strains during in vivo passages: structural and functional implications. Virus Res. *21*, 111–122.

Nelson, M.I., and Holmes, E.C. (2007). The evolution of epidemic influenza. Nat. Rev. Genet. *8*, 196–205.

Nelson, M.I., Simonsen, L., Viboud, C., Miller, M.A., Taylor, J., George, K.S., Griesemer, S.B., Ghedin, E., Sengamalay, N.A., Spiro, D.J., et al. (2006). Stochastic Processes Are Key Determinants of Short-Term Evolution in Influenza A Virus. PLOS Pathog. *2*, e125.

Neverov, A., and Chumakov, K. (2010). Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines. Proc. Natl. Acad. Sci. *107*, 20063–20068.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. *32*, 268–274.

Nobusawa, E., and Sato, K. (2006). Comparison of the Mutation Rates of Human Influenza A and B Viruses. J. Virol. *80*, 3675–3678.

Ohmit, S.E., Petrie, J.G., Malosh, R.E., Cowling, B.J., Thompson, M.G., Shay, D.K., and Monto, A.S. (2013). Influenza Vaccine Effectiveness in the Community and the Household. Clin. Infect. Dis. *56*, 1363–1369.

Ohmit, S.E., Petrie, J.G., Malosh, R.E., Johnson, E., Truscon, R., Aaron, B., Martens, C., Cheng, C., Fry, A.M., and Monto, A.S. (2016). Substantial Influenza Vaccine Effectiveness in Households With Children During the 2013–2014 Influenza Season, When 2009 Pandemic Influenza A(H1N1) Virus Predominated. J. Infect. Dis. *213*, 1229–1236.

Palmenberg, A.C., Spiro, D., Kuzmickas, R., Wang, S., Djikeng, A., Rathe, J.A., Fraser-Liggett, C.M., and Liggett, S.B. (2009). Sequencing and Analyses of All Known Human Rhinovirus Genomes Reveal Structure and Evolution. Science *324*, 55–59.

Parameswaran, P., Charlebois, P., Tellez, Y., Nunez, A., Ryan, E.M., Malboeuf, C.M., Levin, J.Z., Lennon, N.J., Balmaseda, A., Harris, E., et al. (2012). Genome-Wide Patterns of Intrahuman Dengue Virus Diversity Reveal Associations with Viral Phylogenetic Clade and Interhost Diversity. J. Virol. *86*, 8546–8558.

Patel, V., Ferguson, M., and Minor, P.D. (1993). Antigenic Sites on Type 2 Poliovirus. Virology *192*, 361–364.

Paul Glezen, W., Schmier, J.K., Kuehn, C.M., Ryan, K.J., and Oxford, J. (2013). The Burden of Influenza B: A Structured Literature Review. Am. J. Public Health *103*, e43–e51.

Petrie, J.G., Ohmit, S.E., Cowling, B.J., Johnson, E., Cross, R.T., Malosh, R.E., Thompson, M.G., and Monto, A.S. (2013). Influenza Transmission in a Cohort of Households with Children: 2010-2011. PLOS ONE *8*, e75339.

Petrie, J.G., Malosh, R.E., Cheng, C.K., Ohmit, S.E., Martin, E.T., Johnson, E., Truscon, R., Eichelberger, M.C., Gubareva, L.V., Fry, A.M., et al. (2017). The Household Influenza Vaccine Effectiveness Study: Lack of Antibody Response and Protection Following Receipt of 2014–2015 Influenza Vaccine. Clin. Infect. Dis. *65*, 1644–1651.

Petrova, V.N., and Russell, C.A. (2018). The evolution of seasonal influenza viruses. Nat. Rev. Microbiol. *16*, 47–60.

Pons-Salort, M., Molodecky, N.A., O'Reilly, K.M., Wadood, M.Z., Safdar, R.M., Etsano, A., Vaz, R.G., Jafari, H., Grassly, N.C., and Blake, I.M. (2016). Population Immunity against Serotype-2 Poliomyelitis Leading up to the Global Withdrawal of the Oral Poliovirus Vaccine: Spatio-temporal Modelling of Surveillance Data. PLOS Med. *13*, e1002140.

Poon, L.L.M., Song, T., Rosenfeld, R., Lin, X., Rogers, M.B., Zhou, B., Sebra, R., Halpin, R.A., Guan, Y., Twaddle, A., et al. (2016). Quantifying influenza virus diversity and transmission in humans. Nat. Genet. *48*, 195–200.

Popa, A., Genger, J.-W., Nicholson, M.D., Penz, T., Schmid, D., Aberle, S.W., Agerer, B., Lercher, A., Endler, L., Colaço, H., et al. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci. Transl. Med. *12*.

Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K., and Holmes, E.C. (2008). The genomic and epidemiological dynamics of human influenza A virus. Nature *453*, 615–619.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 1–5.

Ren, R.B., Moss, E.G., and Racaniello, V.R. (1991). Identification of two determinants that attenuate vaccine-related type 2 poliovirus. J. Virol. *65*, 1377–1382.

Rhee, C., Kanjilal, S., Baker, M., and Klompas, M. Duration of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infectivity: When Is It Safe to Discontinue Isolation? Clin. Infect. Dis.

Robinson, C.M., Jesudhasan, P.R., and Pfeiffer, J.K. (2014). Bacterial Lipopolysaccharide Binding Enhances Virion Stability and Promotes Environmental Fitness of an Enteric Virus. Cell Host Microbe *15*, 36–46.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Rota, P.A., Wallis, T.R., Harmon, M.W., Rota, J.S., Kendal, A.P., and Nerome, K. (1990). Cocirculation of two distinct evolutionary lineages of influenza type B virus since 1983. Virology *175*, 59–68.

Sahoo, M.K., Holubar, M., Huang, C., Mohamed-Hadley, A., Liu, Y., Waggoner, J.J., Troy, S.B., Garcia-Garcia, L., Ferreyra-Reyes, L., Maldonado, Y., et al. (2017). Detection of Emerging Vaccine-Related Polioviruses by Deep Sequencing. J. Clin. Microbiol. *55*, 2162–2171.

Sanden, S. van der, Pallansch, M.A., Kassteele, J. van de, El-Sayed, N., Sutter, R.W., Koopmans, M., and Avoort, H. van der (2009). Shedding of Vaccine Viruses with Increased Antigenic and Genetic Divergence after Vaccination of Newborns with Monovalent Type 1 Oral Poliovirus Vaccine. J. Virol. *83*, 8693–8704.

Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., and Belshaw, R. (2010). Viral Mutation Rates. J. Virol. *84*, 9733–9748.

Sapoval, N., Mahmoud, M., Jochum, M.D., Liu, Y., Elworth, R.A.L., Wang, Q., Albin, D., Ogilvie, H., Lee, M.D., Villapol, S., et al. (2020). Hidden genomic diversity of SARS-CoV-2: implications for qRT-PCR diagnostics and transmission. BioRxiv 2020.07.02.184481.

Sarcey, E., Serres, A., Tindy, F., Chareyre, A., Ng, S., Nicolas, M., Vetter, E., Bonnevay, T., Abachin, E., and Mallet, L. (2017). Quantifying low-frequency revertants in oral poliovirus vaccine using next generation sequencing. J. Virol. Methods *246*, 75–80.

Sekizuka, T., Itokawa, K., Kageyama, T., Saito, S., Takayama, I., Asanuma, H., Nao, N., Tanaka, R., Hashino, M., Takahashi, T., et al. (2020). Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. Proc. Natl. Acad. Sci.

Shaw, J., Jorba, J., Zhao, K., Iber, J., Chen, Q., Adu, F., Adeniji, A., Bukbuk, D., Baba, M., Henderson, E., et al. (2018). Dynamics of Evolution of Poliovirus Neutralizing Antigenic Sites and Other Capsid Functional Domains during a Large and Prolonged Outbreak. J. Virol. *92*.

Shen, J., Kirk, B.D., Ma, J., and Wang, Q. (2009). Diversifying selective pressure on influenza B virus hemagglutinin. J. Med. Virol. *81*, 114–124.

Shen, Z., Xiao, Y., Kang, L., Ma, W., Shi, L., Zhang, L., Zhou, Z., Yang, J., Zhong, J., Yang, D., et al. (2020). Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. Clin. Infect. Dis. *71*, 713–720.

Sikkema, R.S., Pas, S.D., Nieuwenhuijse, D.F., O'Toole, Á., Verweij, J., Linden, A. van der, Chestakova, I., Schapendonk, C., Pronk, M., Lexmond, P., et al. (2020). COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. Lancet Infect. Dis. *0*.

Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D., et al. (2018). QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics *34*, 163–170.

Stern, A., Yeh, M.T., Zinger, T., Smith, M., Wright, C., Ling, G., Nielsen, R., Macadam, A., and Andino, R. (2017). The Evolutionary Pathway to Virulence of an RNA Virus. Cell *169*, 35-46.e19.

Taniuchi, M., Famulare, M., Zaman, K., Uddin, M.J., Upfill-Brown, A.M., Ahmed, T., Saha, P., Haque, R., Bandyopadhyay, A.S., Modlin, J.F., et al. (2017). Community transmission of type 2 poliovirus after cessation of trivalent oral polio vaccine in Bangladesh: an open-label cluster-randomised trial and modelling study. Lancet Infect. Dis. *17*, 1069–1079.

Tebbens, R.J.D., Pallansch, M.A., Kim, J.-H., Burns, C.C., Kew, O.M., Oberste, M.S., Diop, O.M., Wassilak, S.G.F., Cochi, S.L., and Thompson, K.M. (2013). Oral Poliovirus Vaccine Evolution and Insights Relevant to Modeling the Risks of Circulating Vaccine-Derived Polioviruses (cVDPVs). Risk Anal. *33*, 680–702.

Thompson, W.W., Shay, D.K., Weintraub, E., Brammer, L., Cox, N., Anderson, L.J., and Fukuda, K. (2003). Mortality Associated With Influenza and Respiratory Syncytial Virus in the United States. JAMA *289*, 179–186.

Tonkin-Hill, G., Martincorena, I., Amato, R., Lawson, A.R.J., Gerstung, M., Johnston, I., Jackson, D.K., Park, N.R., Lensing, S.V., Quail, M.A., et al. (2020). Patterns of within-host genetic diversity in SARS-CoV-2. BioRxiv 2020.12.23.424229.

Tsang, T.K., Lau, L.L.H., Cauchemez, S., and Cowling, B.J. (2016). Household Transmission of Influenza Virus. Trends Microbiol. *24*, 123–133.

Valesano, A.L., Fitzsimmons, W.J., McCrone, J.T., Petrie, J.G., Monto, A.S., Martin, E.T., and Lauring, A.S. (2020a). Influenza B Viruses Exhibit Lower Within-Host Diversity than Influenza A Viruses in Human Hosts. J. Virol. *94*.

Valesano, A.L., Taniuchi, M., Fitzsimmons, W.J., Islam, M.O., Ahmed, T., Zaman, K., Haque, R., Wong, W., Famulare, M., and Lauring, A.S. (2020b). The Early Evolution of Oral Poliovirus Vaccine Is Shaped by Strong Positive Selection and Tight Transmission Bottlenecks. Cell Host Microbe *0*.

Vignuzzi, M., and López, C.B. (2019). Defective viral genomes are key drivers of the virus–host interaction. Nat. Microbiol. 1.

Vijaykrishna, D., Holmes, E.C., Joseph, U., Fourment, M., Su, Y.C., Halpin, R., Lee, R.T., Deng, Y.-M., Gunalan, V., Lin, X., et al. (2015). The contrasting phylodynamics of human influenza B viruses. ELife *4*, e05055.

Villabona-Arenas, C.J., Hanage, W.P., and Tully, D.C. (2020). Phylogenetic interpretation during outbreaks requires caution. Nat. Microbiol. *5*, 876–877.

Visher, E., Whitefield, S.E., McCrone, J.T., Fitzsimmons, W., and Lauring, A.S. (2016). The Mutational Robustness of Influenza A Virus. PLOS Pathog. *12*, e1005856.

Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., O'Toole, Á., et al. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 1–17.

Wang, D., Wang, Y., Sun, W., Zhang, L., Ji, J., Zhang, Z., Cheng, X., Li, Y., Xiao, F., Zhu, A., et al. (2020a). Population Bottlenecks and Intra-host Evolution during Human-to-Human Transmission of SARS-CoV-2. BioRxiv 2020.06.26.173203.

Wang, G.P., Sherrill-Mix, S.A., Chang, K.-M., Quince, C., and Bushman, F.D. (2010). Hepatitis C Virus Transmission Bottlenecks Analyzed by Deep Sequencing. J. Virol. *84*, 6218–6228.

Wang, Y., Wang, D., Zhang, L., Sun, W., Zhang, Z., Chen, W., Zhu, A., Huang, Y., Xiao, F., Yao, J., et al. (2020b). Intra-host Variation and Evolutionary Dynamics of SARS-CoV-2 Population in COVID-19 Patients. BioRxiv 2020.05.20.103549.

Wassilak, S., Pate, M.A., Wannemuehler, K., Jenks, J., Burns, C., Chenoweth, P., Abanida, E.A., Adu, F., Baba, M., Gasasira, A., et al. (2011). Outbreak of Type 2 Vaccine-Derived Poliovirus in

Nigeria: Emergence and Widespread Circulation in an Underimmunized Population. J. Infect. Dis. *203*, 898–909.

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. *40*, 11189–11201.

Wong, W., Gauld, J., and Famulare, M. (2020). From Vaccine to Pathogen: Modeling Sabin 2 Vaccine Virus Reversion and Evolutionary Epidemiology. MedRxiv 2020.11.02.20224634.

Worby, C.J., Lipsitch, M., and Hanage, W.P. (2014). Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. PLOS Comput. Biol. *10*, e1003549.

Worby, C.J., Lipsitch, M., and Hanage, W.P. (2017). Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. Am. J. Epidemiol. *186*, 1209–1216.

Xue, K.S., and Bloom, J.D. (2019a). Reconciling disparate estimates of viral genetic diversity during human influenza infections. Nat. Genet. 1.

Xue, K.S., and Bloom, J.D. (2019b). Linking influenza virus evolution within and between human hosts. BioRxiv.

Xue, K.S., and Bloom, J.D. (2020). Linking influenza virus evolution within and between human hosts. Virus Evol. *6*.

Xue, K.S., Stevens-Ayers, T., Campbell, A.P., Englund, J.A., Pergam, S.A., Boeckh, M., and Bloom, J.D. (2017). Parallel evolution of influenza across multiple spatiotemporal scales. ELife *6*, e26875.

Xue, K.S., Moncla, L.H., Bedford, T., and Bloom, J.D. (2018). Within-Host Evolution of Human Influenza Virus. Trends Microbiol. *26*, 781–793.

Yamashita, M., Krystal, M., Fitch, W.M., and Palese, P. (1988). Influenza B virus evolution: Co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses. Virology *163*, 112–122.

Yamayoshi, S., and Kawaoka, Y. (2019). Current and future influenza vaccines. Nat. Med. 1.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, Z., Wong, W.S.W., and Nielsen, R. (2005). Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol. Biol. Evol. *22*, 1107–1118.

Yeh, M.T., Bujaki, E., Dolan, P.T., Smith, M., Wahid, R., Konz, J., Weiner, A.J., Bandyopadhyay, A.S., Van Damme, P., De Coster, I., et al. (2020). Engineering the Live-Attenuated Polio Vaccine to Prevent Reversion to Virulence. Cell Host Microbe.

Yoon, H., and Leitner, T. (2015). PrimerDesign-M: a multiple-alignment based multiple-primer design tool for walking across variable genomes. Bioinformatics *31*, 1472–1474.

Zhao, L., and Illingworth, C.J.R. (2019). Measurements of intrahost viral diversity require an unbiased diversity metric. Virus Evol. *5*.

Zhou, B., Lin, X., Wang, W., Halpin, R.A., Bera, J., Stockwell, T.B., Barr, I.G., and Wentworth, D.E. (2014). Universal Influenza B Virus Genomic Amplification Facilitates Sequencing, Diagnostics, and Reverse Genetics. J. Clin. Microbiol. *52*, 1330–1337.

Zwart, M.P., and Elena, S.F. (2015). Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. Annu. Rev. Virol. *2*, 161–179.

(2020a). Masking strategies for SARS-CoV-2 alignments.

(2020b). Issues with SARS-CoV-2 sequencing data.