Evolutionary dynamics of influenza virus across spatiotemporal scales

Katherine S. Xue

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Jesse Bloom, Chair

Joshua Akey

Trevor Bedford

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Evolutionary dynamics of influenza virus across spatiotemporal scales

Katherine S. Xue

Chair of the Supervisory Committee:
Associate Member Jesse D. Bloom
Fred Hutchinson Cancer Research Center

RNA viruses like influenza mutate rapidly to form genetically diverse populations. Recent high-throughput deep sequencing techniques make it possible to track influenza's evolutionary dynamics at high resolution, showing how viral populations diversify and adapt in just days or weeks. In my dissertation, I examine how influenza viruses evolve across different spatiotemporal scales.

First, I characterize a cooperative interaction between two distinct influenza variants that differ by a single nucleotide mutation. In cell culture, a mixture of the two viral variants grows to higher titers than either variant alone, and populations maintain an equal mixture of the two variants through several passages. Next, I show that this mixture of cooperative variants arises primarily in cell culture rather than in clinical samples. This work provides one of the first examples of a specific cooperative interaction between RNA viruses.

In the rest of my thesis, I focus on how influenza viruses evolve within infected hosts. First, I characterize influenza's evolutionary dynamics within chronically infected individuals. In multi-week infections, I observe extensive parallelism in the mutations

that arise within and between hosts. The same small set of antigenic variants arises recurrently within an individual, in multiple individuals in our study, and in the global influenza population. Next, I resolve a discrepancy between two recent estimates of how much genetic diversity is present within acute influenza infections and what proportion of this genetic diversity is transmitted. I identify a major technical issue in the raw sequencing data for one study that contributes to that study's estimate of high genetic diversity and a large transmission bottleneck. Altogether, this work expands our understanding of the evolutionary forces that shape viral populations across multiple spatiotemporal scales.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

ACKNOWLEDGEMENTS

As a graduate student, I have been lucky to be part of diverse academic communities at the University of Washington and Fred Hutchinson Cancer Research Center. It is both gratifying and daunting to try to acknowledge these many influences, but I will do my best.

I thank my advisor Jesse Bloom for his constant patience and support through a PhD that has had many twists and turns, all of them for the better. Jesse is responsible for the rare combination of intellectual freedom and scientific guidance that I have enjoyed in the past few years, which has gradually given me both the skills and confidence to work on scientific questions that have increasingly felt like my own. I admire many things about Jesse: his incredible dedication to the lab and generosity towards trainees; his talent for communicating scientific concepts; his rigor and thoroughness in pursuing scientific questions; his seemingly inexhaustible patience and good humor; his commitment to doing what is right (except when he steals from the M&M jar). I am grateful for everything I have learned from Jesse and the many opportunities he has given me. If I had the chance, I would do it all again.

I am also grateful to all of the past and present members of the Bloom lab for their support and scientific insights. Katie Hooper taught me everything I know about working with viruses, and I thank Orr Ashenberg, Hugh Haddox, Shirleen Soh, Heather Machkovech, Juhye Lee, and Sarah Hilton in particular for their kindness, good humor, and support. The lab's dedication to rigor, commitment to clear presentations, and appreciation for scientific tangents has made it a fun place to learn and grow.

I am also grateful for my co-advisor Joshua Akey, who has always dared me to push boundaries. My fellow Akey lab members Josh Schraiber, Rajiv McCoy, Ben Vernot, Rachel Gittelman, Serena Tucci, Selina Vattathil, Anne Clark, and Aaron Wolf have helped me learn population genetics and computational biology, and they have been great company and welcoming hosts at conferences and visits across four continents.

I have been lucky to be part of robust scientific communities in both Genome Sciences at UW and Basic Sciences at the Hutch. I thank my committee members

Trevor Bedford, Maitreya Dunham, Harmit Malik, and Benjamin Kerr, who are an evolutionary dream team. Maitreya Dunham, Kelley Harris, and their labs were kind enough to adopt me after Josh moved to Princeton, and I'm grateful to members of both labs for welcoming me and sharing their science whenever I wandered by. I thank Michael Emerman, Julie Overbaugh, and the other members of the retrovirus group for their thoughtful feedback. I also thank Louise Moncla and Allison Black in the Bedford lab for their help and support with virus sequencing, as well as Damien Wilburn, Will deWitt, and Allie Greaney for helping organize molecular evolution supergroup meetings. I am also grateful for the many other graduate students at UW and the Fred Hutch who have made the past few years fun and intellectually stimulating as both colleagues and friends.

My thesis would not have been possible without the help of our clinical collaborators—Terry Stevens-Ayers, Steve Pergam, and Michael Boeckh at the Fred Hutch. A chance run-in between Jesse, Steve, and Michael on December 17, 2015, first sparked the idea of sequencing longitudinal samples from influenza infections, and I am grateful for the work that Terry, Steve, and Michael put in over the next several years to help us transform that idea into reality. Their dedication, support, patience, and mentorship have expanded my conception of what scientific questions are possible to answer.

I have been fortunate to be funded throughout my PhD by the NSF Graduate Research Fellowship and the Hertz Foundation Myhrvold Family Fellowship, which have given me substantial intellectual freedom. I am immensely thankful for the help of Brian Giebel, Alex Moreno, and Helene Obradovich in managing the associated administrative complexities.

As I started my PhD, I was torn between my interests in science and history of science. Through the graduate certificate in Science, Technology, and Society Studies, I think I've gotten the best of both worlds. I'm grateful to my STSS advisor Leah Ceccarelli, who has supported me as I have navigated the STSS world. I am also thankful to STSS faculty Alison Wylie and David Ribes, who have been nothing but enthusiastic about having a scientist in their midst, and to Megan Callow and the Writing Across Difference research cluster, who have helped me to engage with important

questions about how science is taught. I also thank my fellow STSS students Sarah Nelson, Deborah Silvis, Charlie Hahn, Caleb Knapp, Michael Esveldt, Meshell Sturgis, and others for the stimulating, inspiring, and hopeful conversations that arise whenever our wandering paths cross.

A highlight of my time as a PhD student has been the lively community around the UW Genomics Salon, and the friendships I have made through the salon have deeply shaped my time here. I am grateful to my co-founders Jolie Carlisle, who first believed in the idea of the salon, and Hugh Haddox, who both worked tirelessly to turn an uncertain idea into reality. Co-organizers Orlando de Lange, Bryce Taylor, Sarah Nelson, Hannah Gelman, Eliah Overbey, Michael Goldberg, Scott Spencer, Chelsea Grimmer, and Jey Saung have each brought ideas, interests, and creative approaches to the salon that have surprised and delighted me. I am especially thankful for Bryce's fearless dedication to co-leading the salon book club and for the communities that have formed around our discussions of *The Eighth Day of Creation* and *The Origin of Species.* And I am deeply grateful to the nearly 300 different people who have contributed their time and ideas to our salon discussions in the past three years.

I am also thankful for past mentors whose guidance has continued to shape me: Kirsten Bomblies, who introduced me to population genetics and molecular evolution; Christopher Marx, who sparked my interest in microbial evolution; Jeremy Gunawardena and Tathagata Dasgupta, who taught me to think about biological systems; Theresa Holtzclaw, who got me interested in biology; Steven Shapin, who introduced me to sociology of science; John Rosenberg and Jean Martin, who took a chance on a science student; and Verlyn Klinkenborg, whose words shape every sentence I write.

My love of learning comes from my parents, Ziling Xue and Yihui Yang, who gave me every opportunity to learn and showed me the value of hard work. They also taught me that learning what others had discovered wasn't enough - I was maybe six years old when my parents told me that I should strive to someday discover things that no one else knew. My education has been a long journey towards that goal, and I am thankful to my parents for their support and dedication.

I am also grateful to my brother Albert Xue, whose kindness, support, and irreverent comments are a source of irreplaceable brightness. And I am grateful for the many friends in Seattle and elsewhere who have filled these past few years with color and joy.

Most of all, I am thankful for Seungsoo Kim, who makes everything possible.

Chapter 1. INTRODUCTION

A version of this chapter has previously been published as:

Xue, K.S., Moncla, L.H., Bedford, T., Bloom, J.D. Within-host evolution of human influenza virus. *Trends in Microbiol.* (2018). DOI: 10.1016/j.tim.2018.02.007

1.1 WHY STUDY HOW INFLUENZA VIRUSES EVOLVE WITHIN HUMAN HOSTS?

Influenza viruses evolve rapidly on a global scale (Bhatt et al., 2011; Fitch et al., 1991; Ghedin et al., 2005; Rambaut et al., 2008a), and this evolution begins with mutations that arise *de novo* within infected hosts (**Figure 1.1**). As influenza viruses replicate during an infection, they quickly mutate (Bloom, 2014; Nobusawa and Sato, 2006; Pauly et al., 2017; Sanjuán et al., 2010; Suarez-Lopez and Ortin, 1994) to form genetically diverse populations (Andino and Domingo, 2015; Eigen, 1971; Holland et al., 1982; Lauring and Andino, 2010). A small proportion of within-host variants transmit and found a new infection (Brankston et al., 2007; Frise et al., 2016; Varble et al., 2014), and of those, a small number of variants may eventually fix in the global population of influenza viruses. Influenza virus's evolution at the within-host scale is important because it provides the substrate for global evolution.

How do influenza viruses evolve within human hosts, and how does this within-host genetic variation give rise to influenza virus's rapid global evolution? Within hosts, influenza viruses infect heterogeneous cell populations that are arranged in complex spatial structures (van Riel et al., 2006; Shinya et al., 2006). Viruses encounter innate immune defenses like mucus barriers and interferon responses (Iwasaki and Pillai,

2014), as well as adaptive immune responses like antibodies that accumulate over the lifetime of the host (Fonville et al., 2014; Smith et al., 2004a). In some cases, influenza viruses also encounter antiviral drugs like adamantanes and oseltamivir (De Clercq, 2006; McKimm-Breschkin, 2000; van der Vries et al., 2013a). These factors can shape how influenza viruses evolve within humans as well as what viral variants arise and eventually transmit from one individual to another (Grenfell, 2004).

In this review, we summarize recent progress in understanding how and why influenza viruses evolve during the course of an infection, and how evolution within human hosts relates to the virus's global evolution. High-throughput sequencing now makes it possible to "deep sequence" the viral population within a host to measure genetic diversity, so we begin by surveying current deep sequencing methods and their limitations. We then present studies that use deep sequencing to assess viral genetic variation during acute human influenza A infections as well as during chronic influenza infections in immunocompromised human hosts. We consider how factors like antigenic selection, antiviral treatment, tissue specificity, spatial structure, and multiplicity of infection may shape how influenza viruses evolve within hosts. Finally, we discuss how this within-host diversity might relate to global evolution.

1.2 How is deep sequencing used to measure within-host viral diversity?

Traditionally, the viral population within an influenza infection is summarized as a single consensus sequence, representing the most frequent nucleotide at each genome position. For instance, public databases contain tens of thousands of influenza virus

sequences, nearly all of which are consensus sequences (Bao et al., 2008; Bogner et al., 2006; Squires et al., 2012). But in reality, each influenza infection generates a genetically diverse cloud of viral variants that are formed through *de novo* mutation, and variants can also be transmitted from host to host (Andino and Domingo, 2015; Eigen, 1971; Holland et al., 1982; Lauring and Andino, 2010; McCrone and Lauring, 2018). Most mutations in a viral population are expected to reach very low frequencies (**Appendix A**), and very few of these viral variants ever reach majority status in an infection. But the genetic *diversity* within an infection can reveal important evolutionary dynamics—and provides the material on which Darwinian selection can act.

Recent advances in high-throughput sequencing have made it possible to assess mutation frequencies and measure within-host genetic diversity (**Figure 1.2a**) (Beerenwinkel et al., 2012; Posada-Cespedes et al., 2017). Common deep-sequencing approaches can detect viral mutations above frequencies of approximately 1% in the total within-host viral population (Kugelman et al., 2017; McCrone and Lauring, 2016), though it remains difficult to determine linkage among these mutations (Beerenwinkel et al., 2012; Posada-Cespedes et al., 2017). But despite its power, deep sequencing is subject to important technical limitations that are essential to consider when designing experiments and analyzing data (Kugelman et al., 2017; McCrone and Lauring, 2016; Posada-Cespedes et al., 2017).

1.2.1 Experimental design.

A fundamental challenge of viral deep sequencing is the fact that in clinical samples, viral genetic material is often dwarfed by that of the host and co-occurring microbes. To compensate, most studies rely on PCR amplification to enrich for viral

genetic material (Debbink et al., 2017; McCrone and Lauring, 2016; Rogers et al., 2015; Xue et al., 2017). This amplification is relatively straightforward for the influenza-virus genome, which contains conserved regions at the ends of each gene segment (Hoffmann et al., 2001). Following reverse transcription, the entire genome can be amplified using a single set of PCR primers complementary to these conserved regions (McGinnis et al., 2016; Zhou et al., 2009, 2014) or a primer cocktail that is complementary to the conserved regions along with non-coding sequence specific to each gene (Hoffmann et al., 2001; Xue et al., 2017).

Various aspects of the sample and its preparation affect how accurately deep sequencing measures the actual viral variant frequencies within an infected individual (Illingworth et al., 2017; Kugelman et al., 2017; McCrone and Lauring, 2016; Posada-Cespedes et al., 2017; Zanini et al., 2017). Of these factors, the most important by far is viral load (**Figure 1.2b)** (Gallet et al., 2017; McCrone and Lauring, 2016). During whole-genome amplification, anywhere from 20 to 35 cycles of PCR may be required to produce sufficient material for sequencing. When the number of starting viral template molecules is low, below about 1000 copies per uL total RNA (McCrone and Lauring, 2016), this amplification can significantly distort variant frequencies (Gallet et al., 2017; Kanagawa, 2003; McCrone and Lauring, 2016). By comparison, errors that accumulate during reverse transcription, PCR, and Illumina sequencing have smaller effects for samples with typical low viral loads (Gallet et al., 2017; McCrone and Lauring, 2016).

It is therefore essential to maximize the amount of viral genetic material used in each RNA extraction, reverse-transcription, and PCR reaction to ensure that deep sequencing accurately measures variant frequencies in the viral population. It is also

important to prepare and sequence replicate libraries (Illingworth et al., 2017), preferably beginning from independent reverse-transcription reactions (McCrone and Lauring, 2016). Replicate libraries make it possible to identify samples with low viral load (McCrone and Lauring, 2016; Xue et al., 2017) or effective sequencing depth (Illingworth et al., 2017) that should be excluded from downstream analyses (**Figure 1.2b**). They also make it possible to empirically set variant-calling thresholds and exclude specific low-confidence viral variants whose frequencies vary extensively between replicates in an otherwise high-quality sample (McCrone and Lauring, 2016).

1.2.2 Limitations.

Deep sequencing can identify rare mutations in a viral population, but it has limited power to determine patterns of linkage between mutations, which can reveal patterns of epistasis (Illingworth, 2015) and clonal competition (Xue et al., 2017). Short reads can sometimes reveal linkage between closely spaced mutations (Illingworth, 2016; Illingworth et al., 2014; Sobel Leonard et al., 2017a; Xue et al., 2017), but the reads produced by Illumina sequencing are unable to span even the smallest influenza-virus genes. Several groups have successfully assembled viral haplotypes and assessed their frequencies by combining low-coverage PacBio sequencing, which produces long reads, with high-coverage Illumina sequencing (Rogers et al., 2015). But even these methods cannot directly determine linkage between mutations on different gene segments, even though intergenic epistasis (Kryazhimskiy et al., 2011; Mitnaul et al., 2000; Neverov et al., 2015; Wagner et al., 2002) and gene reassortment (Lowen, 2017) both affect influenza-virus evolution. In the absence of sequencing data that directly observe patterns of linkage between mutations, computational methods can

sometimes infer longer haplotypes by assembling multiple short-read haplotypes (Beerenwinkel et al., 2012; Illingworth, 2015; Posada-Cespedes et al., 2017; Sobel Leonard et al., 2017a) and tracking concordant changes in allele frequencies between mutations located on different genes (Illingworth, 2015; Sobel Leonard et al., 2017a). Even with current technical limitations, deep-sequencing approaches to measure viral variation can still shed light on important within-host evolutionary dynamics.

1.3 How do influenza viruses evolve within human hosts?

Several recent studies have used deep sequencing to characterize the spectrum of genetic diversity within natural human influenza A infections, and we summarize their findings here. Most studies focus on typical acute infections in immunocompetent hosts, but some studies also examine viral evolution during the lengthy infections experienced by immunocompromised patients.

1.3.1 Acute infections.

Viruses like HIV and hepatitis C virus establish long-term infections and evolve over years or decades to avoid the immune system and develop antiviral resistance (Lemey et al., 2006; Rambaut et al., 2004; Simmonds, 2004). In contrast, influenza infections typically last five to seven days, and viral shedding peaks two to four days after infections begin (Baccam et al., 2006; Carrat et al., 2008). These short infections provide little time for *de novo* mutations to arise, for selection to act on these mutations, and for selected mutations to reach frequencies at which they are detectable by deep sequencing (**Appendix A**).

Most studies of natural, acute influenza infections analyze one or two nasal swab or nasal wash samples from each patient by deep sequencing the hemagglutinin gene (Cushing et al., 2015; Dinis et al., 2016) or the entire viral genome (Debbink et al., 2017; McCrone et al., 2017; Poon et al., 2016). The exact number of viral variants identified is highly dependent on sample quality and sequencing methodology. But several studies have observed relatively limited genetic diversity within acute human influenza infections (Debbink et al., 2017; Dinis et al., 2016; McCrone et al., 2017), identifying fewer than ten variants per infection across the influenza-virus genome at a limit of detection of approximately 1-2% (Debbink et al., 2017; McCrone et al., 2017). Most of these mutations are rare, present in less than 10% of the viruses within a host (Debbink et al., 2017; Dinis et al., 2016; McCrone et al., 2017), and the number and frequency of within-host viral variants does not seem to correlate with how many days post-infection the samples were collected (Debbink et al., 2017). However, some acute infections harbor high genetic diversity due to apparent co-infection by multiple, related viral strains (McCrone et al., 2017; Poon et al., 2016). One study has found evidence of mixed infections in approximately half of the patients sequenced (Poon et al., 2016), and the contribution of co-infection to within-host genetic diversity requires further careful study. Overall, the limited genetic diversity found in many acute human influenza infections agrees with prior studies in dogs and horses that sequenced viral clones to measure within-host viral variation (Hoelzer et al., 2010; Hughes et al., 2012; Murcia et al., 2010, 2013).

It remains unclear what influences the patterns of observed variation, although we discuss potential biological factors below. Generally, within-host variants tend to be

dispersed across the viral genome (Debbink et al., 2017; McCrone et al., 2017), though one study observes some low-frequency variation in putative antigenic sites (Dinis et al., 2016). Another study estimated that the ratio of nonsynonymous to synonymous within-host variants is about 0.64 and suggested that purifying selection removes some deleterious variants in human infections (McCrone et al., 2017). Even if most acute human infections do not contain high-frequency mutations, the sheer number of influenza infections every year may allow the rapid global evolution of influenza virus to arise from limited within-host genetic diversity.

1.3.2 Chronic infections.

The vast majority of influenza infections are acute, but immunocompromised patients can experience severe infections lasting multiple weeks or months (Memoli et al., 2014; Nichols et al., 2004; Vigil et al., 2010). These chronic infections differ from acute infections in that host immune responses may be weakened or absent, infections are commonly treated with long courses of antiviral drugs, and influenza virus commonly co-occurs with other respiratory pathogens (Memoli et al., 2014; Nichols et al., 2004; Vigil et al., 2010).

Nevertheless, chronic infections provide unusual opportunities to observe how influenza viruses evolve within humans over longer spans of time, when selection has more opportunities to shape viral variation. Immunocompromised patients often receive close clinical monitoring, and several studies have tracked within-host evolution longitudinally by deep sequencing clinical samples taken from different time points in an infection (Ghedin et al., 2011; Rogers et al., 2015; Xue et al., 2017). In these chronic infections, influenza viruses can display extensive evolution. Putative antigenic variants

can arise and reach high within-host frequencies (Baz et al., 2006; McMinn et al., 1999; Rocha et al., 1991; Xue et al., 2017). Multiple drug-resistant variants can also arise during these lengthy infections (Baz et al., 2006; Ghedin et al., 2011; Rogers et al., 2015; Xue et al., 2017). It is common for multiple beneficial mutations to compete with one another within a patient (Rogers et al., 2015; Xue et al., 2017), displaying clonal interference dynamics commonly observed in experimental evolution (Hegreness et al., 2006; Kao and Sherlock, 2008; Lang et al., 2013).

The relatively weak immune responses mounted by immunocompromised hosts can have important evolutionary consequences, regardless of the exact underlying medical conditions. Small viral populations can survive and replicate in the presence of weak selection, making it easier for multiple adaptive mutations to emerge simultaneously (Feder et al., 2016). In chronic influenza infections, relatively weaker immune responses can lead to much longer viral infections, enabling putative antigenic variants to arise in ways that sometimes parallel global evolutionary trends (Xue et al., 2017). Overall, though, it remains unclear how much the evolutionary forces that act within chronic infections resemble selective pressures within more common, acute infections.

1.4 What affects how influenza viruses evolve within humans?

Here, we consider evidence for how antigenic selection, antiviral treatment, tissue specificity, spatial structure, and multiplicity of infection may shape how influenza viruses evolve within humans (**Figure 1.3**).

1.4.1 Antigenic selection.

Human influenza viruses undergo constant antigenic drift and occasional antigenic shift on a global evolutionary scale (Bedford et al., 2014; Hensley et al., 2009; Smith et al., 2004a), but it is unclear how much immune selection takes place within a typical human infection. Recent deep-sequencing studies have identified few antigenic variants within acute infections (Debbink et al., 2017; Dinis et al., 2016; McCrone et al., 2017), though it remains unclear whether antigenic variants are enriched or depleted relative to the frequency of within-host variants as a whole. In immunocompromised patients, putative antigenic variants can arise, display complex clonal dynamics, and even fix during an infection (Baz et al., 2006; McMinn et al., 1999; Rocha et al., 1991; Xue et al., 2017). Some of the putative antigenic variants that arise in immunocompromised patients also reach a high frequency in the global population of influenza viruses (Xue et al., 2017).

Another source of antigenic selection might be vaccination, which boosts immune responses against influenza viruses. Two recent studies deep sequenced viral populations from vaccine recipients and control groups (Debbink et al., 2017; Dinis et al., 2016). They found that vaccination status did not seem to affect consensus viral sequences, suggesting that infections in vaccinated individuals are not caused by specific resistant viral strains (Debbink et al., 2017; Dinis et al., 2016). Moreover, they found that vaccination had no detectable effect on the number or population frequency of within-host variants (Debbink et al., 2017; Dinis et al., 2016). One interpretation is that antigenic selection does not act detectably in most infections. An alternative explanation is that many unvaccinated individuals may already have strong immunity

from natural infections, and vaccination may not alter immunity enough to exert additional antigenic selection.

1.4.2 Antiviral resistance.

Antiviral agents are used to treat only a minority of acute influenza infections, but they can still exert important influences on viral evolution (De Clercq, 2006; McKimm-Breschkin, 2000; van der Vries et al., 2013a). For instance, many influenza A strains are resistant to adamantanes (Dong et al., 2015; van der Vries et al., 2013a), and resistance to oseltamivir swept to fixation in seasonal H1N1 influenza viruses before they were replaced by pandemic H1N1 (Bloom et al., 2010; Renaud et al., 2011; van der Vries et al., 2013a). For antivirals like oseltamivir, where drug resistance is not yet widespread in current strains, influenza viruses can gain resistance within individual infections by acquiring one or more *de novo* mutations (van der Vries et al., 2013a). As with antigenic selection, it is unclear how frequently drug resistance arises within typical, acute infections. In one case report, resistance arose even when oseltamivir was used for prophylaxis (Baz et al., 2009), but deep sequencing of viral populations from thirteen individuals in a human challenge study detected no drug-resistant variants following early or standard oseltamivir treatment (Sobel Leonard et al., 2017a). There is ample evidence, however, that resistance can arise rapidly during longer infections (Baz et al., 2006; Boivin et al., 2002; Memoli et al., 2014; Nichols et al., 2004; Rogers et al., 2015; Vigil et al., 2010; van der Vries et al., 2013b; Xue et al., 2017). In some cases, multiple drug-resistant variants may even compete within a single patient (Rogers et al., 2015; Xue et al., 2017). Since the mutations and molecular mechanisms underlying antiviral

resistance are well established, antiviral resistance can serve as a useful comparison for studying how other selective pressures may act within hosts.

1.4.3 Tissue specificity and spatial structure.

Influenza viruses infect heterogeneous, spatially structured populations of cells in the human airways. Differences between tissues, along with neutral processes of migration and genetic drift, may have important effects on viral evolution. One major difference between the upper and lower human airways is their distribution of sialic acid receptors, which influenza viruses use to enter host cells. Most human influenza infections primarily take place in the upper human respiratory tract, which contains a higher proportion of α2,6-linked sialic acids than the lower airways, which contain a higher proportion of α2,3-linked sialic acids (van Riel et al., 2006; Shinya et al., 2006). These histological differences may affect which viruses are transmitted. In ferrets, for example, viruses tend to transmit from the upper respiratory tract (Varble et al., 2014), and viral variants that preferentially bind to α2,6-linked receptors transmit more frequently than variants that bind α2,3-linked receptors (Lakdawala et al., 2015). This combination of spatial structure and tissue specificity provides one possible explanation for why avian-derived viruses, which tend to be adapted to the α2,3-linked sialic acid receptors in avian airways, can cause severe, lower lung infections in humans but rarely transmit from one human host to another (van Riel et al., 2006; Shinya et al., 2006; Xu et al., 2013). Within these two broad linkage categories, sialic acid chains also vary extensively in length and chemical linkages and are distributed differently in the airways of avian and mammalian host species, potentially affecting influenza virus binding (Chandrasekaran et al., 2008; Matrosovich et al., 2004; Thompson et al., 2006).

Even in the absence of tissue-specific selection, spatial structure can also limit genetic exchange between different parts of the human airways. For instance, one case report of a human infection documented the presence of distinct viral populations in the right and left lungs (Hamada et al., 2012). More generally, though, no deep sequencing studies have systematically compared viral populations sampled from different parts of the human airways, and the extent of tissue-specific selection remains an important open question.

1.4.4 Multiplicity of infection.

Spatial structure affects how densely viruses populate different parts of the human airways, and in turn, this within-host multiplicity of infection (MOI) determines how often two or more viruses co-infect the same host cell. When multiple viruses co-infect the same cell, viral gene segments have an opportunity to reassort, and they do so readily in cell culture and animal models (Lowen, 2017; Marshall et al., 2013; Tao et al., 2015). New combinations of gene segments are important for purging deleterious mutations in an otherwise clonal population and for forming new, potentially advantageous combinations between mutations (Lowen, 2017). It is usually difficult to estimate rates of within-host reassortment because current deep sequencing techniques are unable to establish linkage across multiple gene segments. But one group has developed a population-genetics framework to infer recombination from longitudinal, short-read sequencing data and estimated that the rate of effective within-host reassortment is low in human infections (Sobel Leonard et al., 2017a). Rates of effective reassortment may be low even when viral load is high because spatial

structure limits viral exchange so that most co-infection and reassortment occurs between genetically similar viruses.

Viral co-infection also provides opportunities for genetic complementation, which can decrease the efficacy of selection. If a wild-type virus and a virus carrying a deleterious mutation co-infect the same cell, the progeny virions can package both viral genomes, allowing the deleterious mutation to persist. The effects of complementation are especially clear in cell culture, where most influenza viruses are grown at high MOIs: defective viruses that carry large gene deletions quickly arise and spread through the population (Davis et al., 1980; Frensing et al., 2013). Large internal deletions have been documented in human influenza infections (Saira et al., 2013; Vasilijevic et al., 2017), and studies of influenza outbreaks in pigs and horses have documented the transmission of nonsense variants as well (Hughes et al., 2012; Murcia et al., 2012). However, the overall prevalence of defective viral particles and their association with infection length and severity remain poorly understood.

Altogether, studies in cell culture and animal models suggest various biochemical and morphological factors that may affect how influenza viruses evolve within human hosts, but few deep sequencing studies so far have had the power to detect their effects. Additional sequencing of viral populations collected from different human hosts and tissues will improve our understanding of how influenza viruses evolve within a complex host environment.

1.5 How does influenza virus's diversity within hosts relate to its global evolution?

The within-host evolution of influenza virus ultimately provides the substrate for the virus's rapid global evolution, but the forces that transform within-host genetic diversity into global variation are largely unknown. Selection and drift can operate within hosts, but they also shape viral variation at transmission and at the host-population level.

1.5.1 Transmission.

Only a small fraction of the influenza viruses within an infected individual go on to initiate subsequent infections (**Figure 1.4**). Transmission bottlenecks can limit the genetic diversity passed from one host to another and introduce stochasticity in variant frequencies along a transmission chain (McCrone and Lauring, 2018; McCrone et al., 2017; Poon et al., 2016). Transmission bottleneck sizes also affect how often genetically distinct strains of influenza virus infect the same individual (Hughes et al., 2012; Tao et al., 2015), and looser bottlenecks increase the chance for beneficial reassortment (Lowen, 2017; Tao et al., 2015).

Deep sequencing of contact and recipient viral populations can help estimate transmission bottleneck sizes in natural infections. Narrower transmission bottlenecks increase the variance with which viral variants are transmitted (Sobel Leonard et al., 2017b). Animal studies suggest that vaccination status (Murcia et al., 2013) and route of transmission (Frise et al., 2016; Varble et al., 2014) can affect transmission bottleneck size, which appears to be looser for direct contact than for aerosol transmission (Frise et al., 2016; Varble et al., 2014). Studies of influenza outbreaks in pigs and horses have

suggested that transmission bottlenecks can be loose, with frequent mixed infections (Hughes et al., 2012; Murcia et al., 2010, 2012).

In human influenza infections, few studies have had the power to estimate transmission bottleneck sizes, and the two recent studies to do so have disagreed considerably in their results. Poon et al. estimate a relatively loose bottleneck size of approximately 200 distinct genomes for both H3N2 and pandemic H1N1 influenza virus based on a household cohort study performed during the first wave of the 2009 H1N1 pandemic (Poon et al., 2016), and a recent re-analysis of the same data supports these estimates (Sobel Leonard et al., 2017b). More recently, McCrone et al. use similar analytical methods to infer a very narrow transmission bottleneck of 1 or 2 distinct genomes in a household cohort study that primarily sampled seasonal H3N2 influenza viruses from 2010 to 2015 (McCrone et al., 2017).

It is unclear what accounts for the differences between these two estimates, although differences in study populations may contribute. For instance, influenza virus transmission depends on temperature and humidity (Lowen et al., 2007). The Poon *et al.* cohort was recruited in sub-tropical Hong Kong, while the McCrone *et al.* cohort was recruited in temperate Michigan, in the northern United States. Moreover, the Poon *et al.* study recruited index patients with acute respiratory illnesses and then prospectively followed their family members, whereas the McCrone *et al.* study prospectively enrolled households and queried participants weekly about symptoms of illness. Furthermore, estimates of transmission bottleneck size may also be highly sensitive to sample quality, library preparation and sequencing methods, and variant-calling thresholds.

Most studies assume that transmission bottlenecks act neutrally, but certain influenza-virus variants may be more likely than others to transmit and found new infections. For instance, one ferret study found that transmitted viruses tended to preferentially bind α2,6-linked sialic acid receptors and most closely resembled viral populations in the soft palate (Lakdawala et al., 2015). Selection can also affect maladapted strains of human influenza virus. In one recent human challenge study, volunteers were inoculated with viral stocks that had acquired passage-adaptation mutations during growth in eggs and cell culture. Many of these passage-adaptation mutations in the viral inoculum were purged from the viral population during or shortly after inoculation (Sobel Leonard et al., 2016). Selection may also act at transmission to promote global antigenic evolution if novel antigenic variants transmit and found new infections more frequently when host populations are mostly resistant to circulating strains. The strength and evolutionary effects of transmission bottlenecks remain important areas of study for understanding how the genetic diversity of influenza virus within hosts relates to its global genetic variation.

1.5.2 Comparing evolutionary scales.

New mutations must arise and fix in individual hosts before they can spread through a large host population, linking within-host evolutionary dynamics to global evolution (Alizon et al., 2011). How do drift, positive selection, and purifying selection act within and between hosts? Studies of Ebola virus (Park et al., 2015), Lassa virus (Andersen et al., 2015), and dengue virus (Holmes, 2003) have compared the proportions of nonsynonymous to synonymous within- and between-host variants to argue that purifying selection acts at within- and between-host scales to eliminate

deleterious variants. However, the d_N/d_S ratio was originally developed to compare fixed variation between distantly diverged lineages, and within-host population dynamics can complicate its interpretation (Kryazhimskiy and Plotkin, 2008; Mugal et al., 2014). In cases where longitudinal deep-sequencing data is available, standard population-genetics models can be used to infer the influence of selection upon particular variants based on the changes in their allele frequencies over time (Illingworth, 2015; Illingworth et al., 2014; Sobel Leonard et al., 2017a). But for most studies of within-host evolution, which lack longitudinal data, it remains a major challenge to develop appropriate methods that make use of deep-sequencing data to distinguish what evolutionary forces act on viral populations within hosts.

1.6 Concluding Remarks

By studying how influenza viruses evolve within humans, we can observe what biological factors affect the virus within its natural host environment. We can also determine what evolutionary and epidemiological forces transform within-host genetic diversity into global viral variation. As deep sequencing makes it easier to survey genetic diversity within hosts, it will be important to develop methodologies to systematically analyze within-host evolutionary dynamics and their relationship to global evolution.

1.7 FIGURES AND TABLES

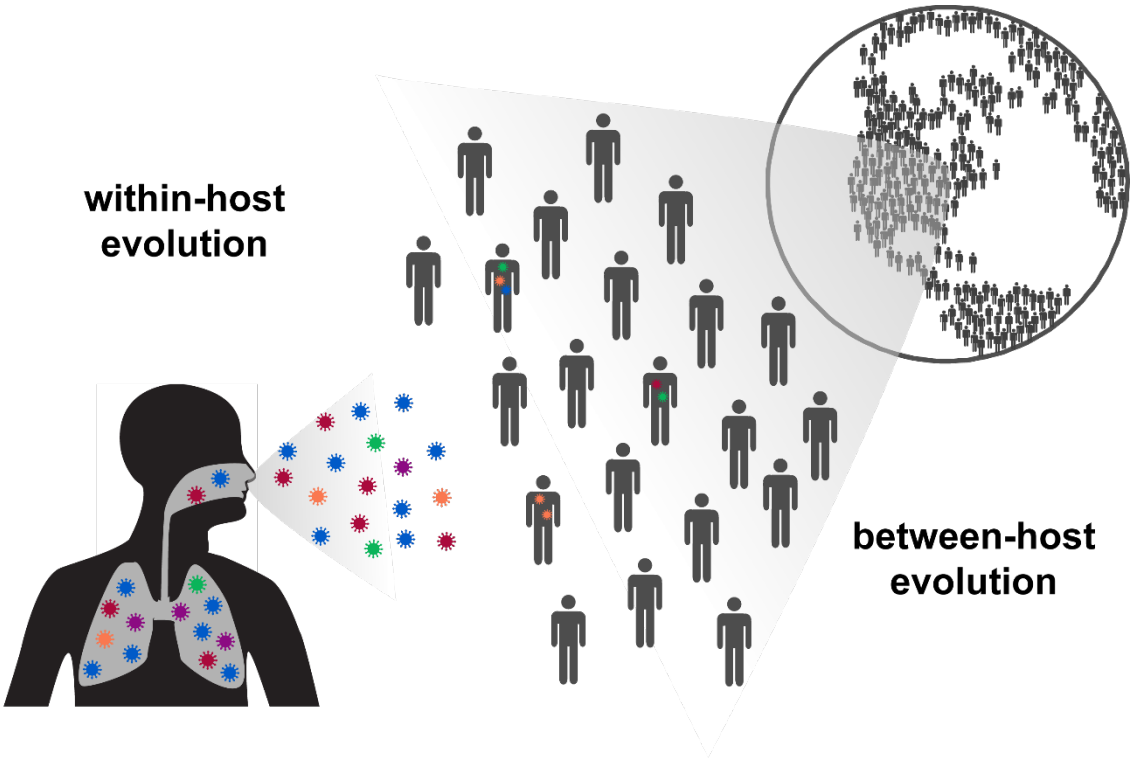

Figure 1.1. Within- and between-host evolutionary scales.

The rapid global evolution of influenza virus begins with *de novo* mutations that arise within individual infected hosts.

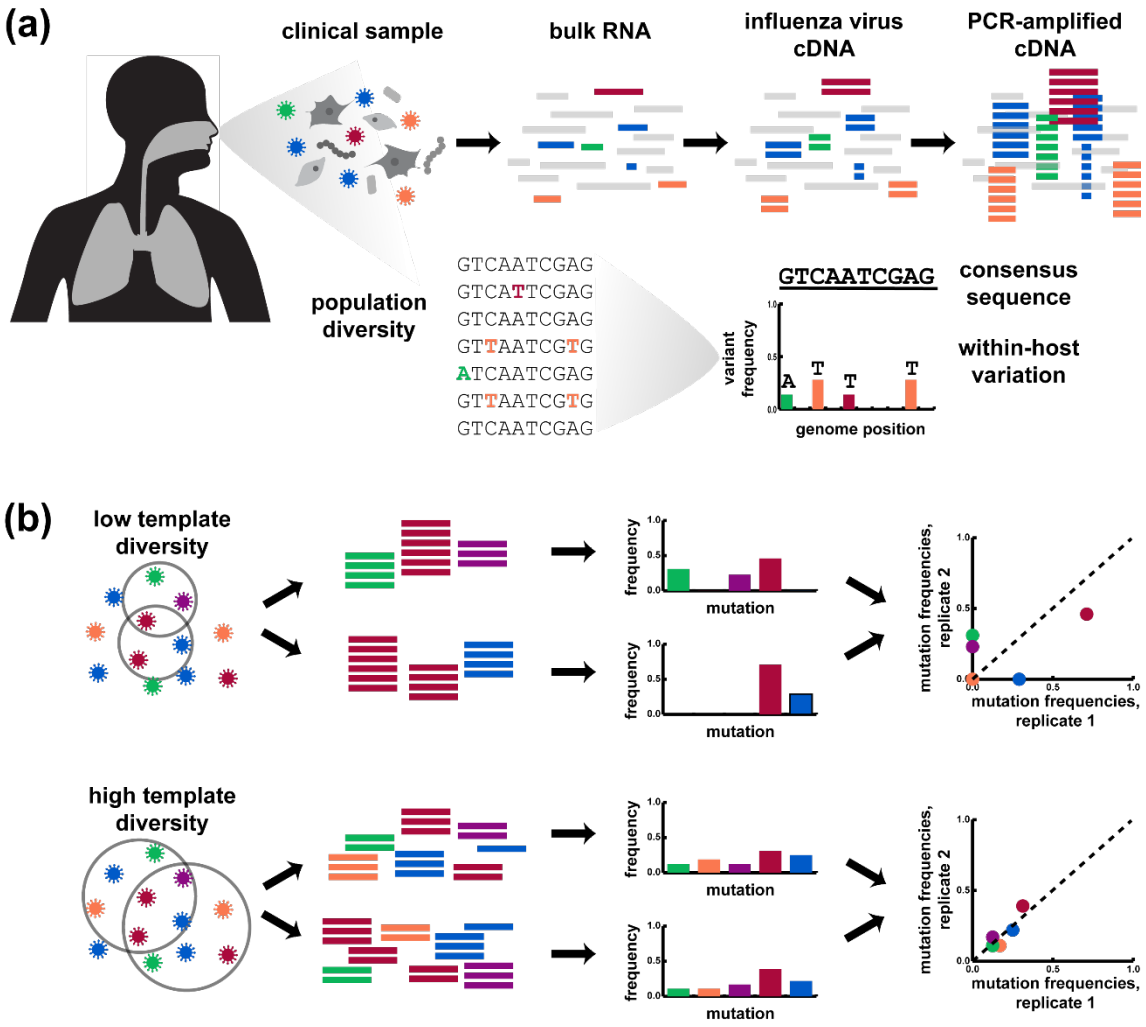

Figure 1.2. Deep-sequencing approaches to measuring within-host genetic diversity.

(A) Common deep-sequencing workflows can identify variants that make up approximately 1% of the within-host population. **(B)** Most studies amplify viral genetic material prior to deep sequencing. Low template diversity, typically due to low viral load, can distort the variant frequencies measured by deep sequencing. Replicate libraries are important for identifying and excluding samples with low viral load that should be excluded from downstream analyses.

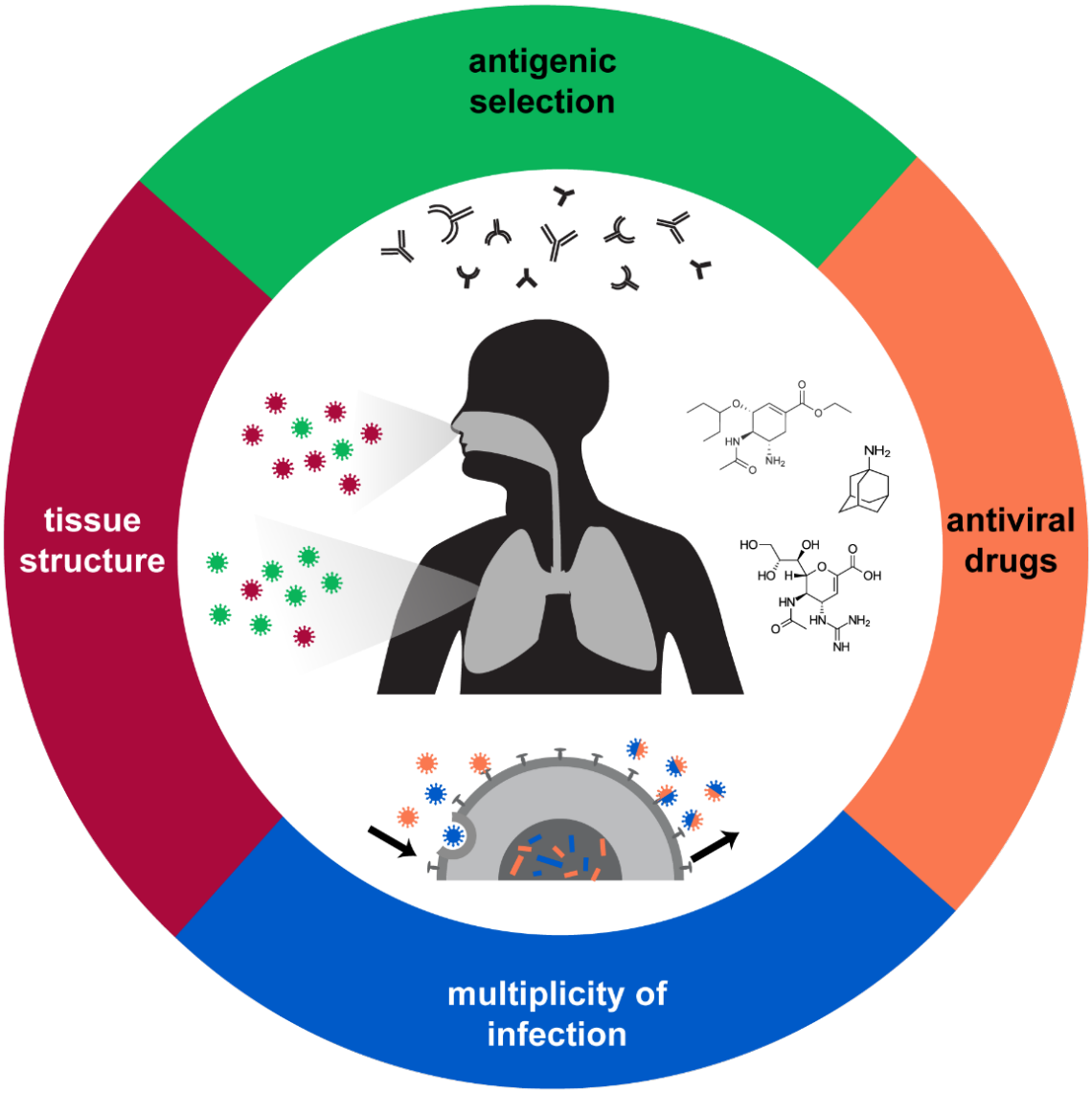

Figure 1.3. Factors affecting the evolution of influenza virus within human hosts.

Antigenic selection, antiviral drugs, tissue structure, and multiplicity of infection can affect how influenza viruses evolve within hosts.

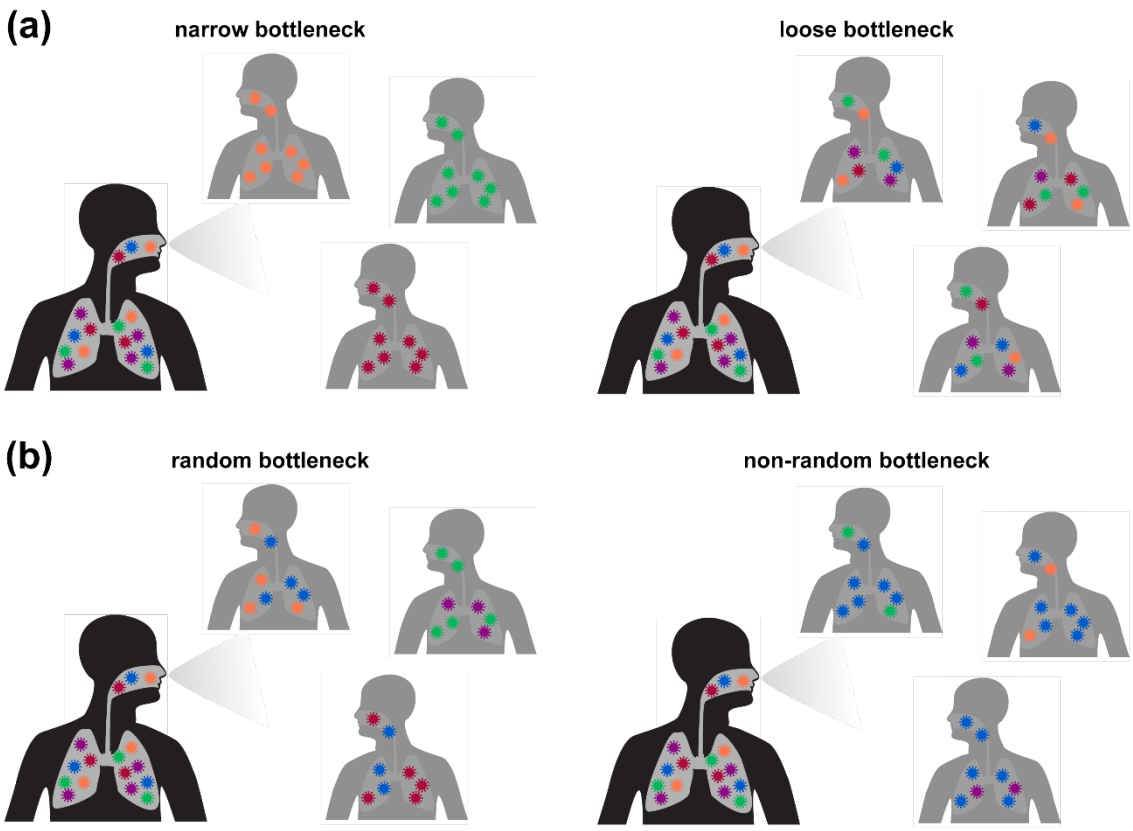

Figure 1.4. Transmission bottlenecks shape viral evolution.

The size **(A)** and randomness **(B)** of transmission bottlenecks affect how much of the viral genetic diversity generated within one host survives to initiate another infection.

Chapter 2. COOPERATION BETWEEN DISTINCT VIRAL VARIANTS PROMOTES GROWTH OF H3N2 INFLUENZA IN CELL CULTURE

A version of this chapter has previously been published as:

Xue, K.S., Hooper, K.A., Ollodart, A.R., Dingens, A., Bloom, J.D. Cooperation between distinct viral variants promotes growth of H3N2 influenza in cell culture. eLife 5: e13974 (2016). DOI: 10.7554/eLife.13974

The evolution of RNA viruses is characterized by high mutation rates and large population sizes, which together create genetically diverse populations known as quasispecies (Andino and Domingo, 2015; Eigen, 1971; Holland et al., 1982; Lauring and Andino, 2010). High levels of standing genetic diversity can provide a substrate for selection and rapid adaptation, an advantage for viruses that experience strong and varied selective pressures to escape immune recognition, develop drug resistance, and adapt to new hosts (Dutta et al., 2008; Najera et al., 1995; Pfeiffer and Kirkegaard, 2005).

Recently, several studies have suggested that cooperative interactions between variants in a quasispecies can also increase population-level fitness (Bordería et al., 2015; Ciota et al., 2012; Ke et al., 2013; Shirogane et al., 2012; Vignuzzi et al., 2006). Vignuzzi *et al.* (2006) found that genetically diverse poliovirus populations were required for wild-type neurotropism and pathogenesis, leading the authors to suggest that unknown cooperative interactions among minor variants promoted the overall fitness of

the population. More recent studies have suggested cooperation between distinct variants of measles (Shirogane et al., 2012), hepatitis B virus (Cao et al., 2014), and Coxsackie virus (Bordería et al., 2015). However, specific examples of robust cooperative interactions between defined variants in viral quasispecies remain rare (Holmes, 2010).

Here, we demonstrate that cooperation between two distinct variants of human H3N2 influenza promotes viral growth in cell culture. The two variants differ by a single amino-acid mutation in the neuraminidase (NA) protein, which normally mediates viral exit from the host cell. Both variants are reported numerous times in human H3N2 NA sequences deposited in the GISAID EpiFlu database, and both have been observed in mixed populations when clinical specimens are passaged in cell culture. We show that the two variants grow better together than apart, and that serial passage repeatably selects for mixed populations. We suggest that the cooperation arises because one variant is proficient at cell entry while the other is proficient at cell exit. Overall, our work represents a clear example of selection to generate and maintain two cooperating genotypes within a viral quasispecies.

2.1 RESULTS

2.1.1 Mutations at site 151 in H3N2 neuraminidase tend to occur in mixed populations.

Over the last decade, several groups have reported that mutations arise rapidly and repeatedly at residue NA 151 when human H3N2 influenza is passaged in cell culture (**Table 2.1**) (Chambers et al., 2014; Lee et al., 2013; Lin et al., 2010; McKimm-Breschkin et al., 2003; Mishin et al., 2014; Mohr et al., 2015; Tamura et al., 2013).

Residue 151 is in the NA active site and is highly conserved; until recently, it had an amino-acid identity of D in virtually all N2 NAs. Ordinarily, NA mediates viral exit from the host cell by cleaving sialic-acid receptors to release newly produced virions. The D151G mutation ablates the catalytic activity of NA and instead causes it to bind the receptors that it typically cleaves (Zhu et al., 2012). Mutations at this site seem to be more common in viruses that have been passaged in cell culture compared to the original clinical isolates (Chambers et al., 2014; Deyde et al., 2009; Lee et al., 2013; Lin et al., 2010; Mishin et al., 2014; Okomo-Adhiambo et al., 2010; Tamura et al., 2013). As a result, mutations at site 151 have been categorized as lab adaptations (Lee et al., 2013; Mishin et al., 2014; Okomo-Adhiambo et al., 2010; Tamura et al., 2013).

We examined whether mutations at site 151 exhibited patterns consistent with simple lab adaptation by determining the frequencies of amino acids at this position in human H3N2 NA sequences in the GISAID EpiFlu database for each year from 2000 to 2014 (**Figure 2.1**). Most isolates in this database are first passaged in eggs or cell culture, and then the consensus sequence of the viral population is determined by Sanger sequencing. Beginning in 2007, the frequency of mutations at NA site 151 rose dramatically, with mutant genotypes representing about a quarter of the sequences. G151 and N151 are each reported in about 1% of sequences, but ambiguous nucleotide calls at the codon make up the majority of non-wild-type sequences. Because these sequences usually represent consensus calls from Sanger sequencing, the ambiguous nucleotides likely indicate the presence of mixed D151+G151 and D151+N151 populations. The relative abundance of mixed populations in strains deposited in the GISAID EpiFlu database is consistent with the fact that, in tissue culture, mutations at

site 151 arise frequently but fix rarely (**Table 2.1**) (Lee et al., 2013; Lin et al., 2010; Mishin et al., 2014; Mohr et al., 2015; Okomo-Adhiambo et al., 2010; Tamura et al., 2013).

When viruses are passaged, they experience culture-specific selective pressures. To understand how variation at site 151 might depend on the passaging procedures, we classified sequences according to their passage histories as annotated in the GISAID EpiFlu database. Prior to 2007, site 151 is almost always reported to be D, regardless of passage history (**Figure 2.1B**). From 2007 onward, variants at site 151 are reported almost exclusively in isolates that have been passaged in cell culture (**Figure 2.1C**). Whereas nearly a third of isolates that have been passaged in cell culture are reported to have an amino acid other than D at site 151, only two of nearly 1800 unpassaged isolates and five of nearly 400 egg-passaged isolates are reported to have non-D amino acids at site 151. These observations accord with reports in the literature that mutations at site 151 are observed primarily in cell-culture-passaged isolates (Lee et al., 2013; Mohr et al., 2015; Tamura et al., 2013). However, it is important to remember that the methods used to determine most sequences in the GISAID EpiFlu database lack sensitivity to detect minority variants in a viral population. For instance, the experimental results that we describe below suggest that it is exceedingly unlikely that any of the more than 100 reported G151 sequences actually reflect the complete fixation of this mutation.

Overall, the results in **Figure 2.1** indicate that the G151 mutation tends to occur in mixed populations. We therefore sought to experimentally characterize the growth of pure D151 and G151 viral variants to determine whether mixed populations represent

incomplete fixation of a lab-adaptation mutation or whether they are the product of active selection.

2.1.2 Mixed populations of D151 and G151 viruses grow better than pure populations of either variant.

We compared the growth of the D151 and G151 variants both alone and in mixed populations by generating viruses of defined genotypes using reverse genetics. The A/Hanoi/Q118/2007 strain, henceforth referred to as Hanoi/2007, is a human H3N2 strain with a G151 genotype. Its NA protein sequence is identical to that of several other sequenced isolates, except that the other strains have D at site 151. We created reverse-genetics plasmids encoding the protein sequences of both the D151 and G151 variants of the Hanoi/2007 NA, as well as the HA from this strain. The internal genes were derived from the lab-adapted A/WSN/33 influenza strain with GFP packaged in the PB1 segment (Bloom et al., 2010). The two viral variants were therefore isogenic except for the variation at site 151.

We generated virus using reverse genetics by co-transfecting cells with plasmids encoding the D151 variant, the G151 variant, or an equal mix of the two, together with isogenic plasmids for the other viral genes, and we quantified the resulting titers (**Figure 2.2A**, **Figure 2.3**). Surprisingly, given its widespread designation as a lab adaptation, the G151 variant grows extremely poorly. However, a mixed population of D151 and G151 variants grows to substantially higher titers than the corresponding pure population of D151 viruses. The growth advantage of the mixed population suggests that cooperation between the two variants improves viral growth.

We next sought to determine whether there was also a cooperative effect when the D151 and G151 variants were mixed in direct infections, since generation of influenza virus by reverse genetics is a complex process that involves co-transfecting cells with plasmids encoding each of the eight viral genes. We generated pure populations of D151 and G151 viruses by reverse genetics, growing both populations in the presence of 50 nM oseltamivir, a small molecule that competes with sialic acid for binding to the NA active site. The addition of oseltamivir increases the titers of the G151 variant (**Figure 2.4**) and presumably prevents selection for *de novo* NA mutations by suppressing both the cleavage and binding activity of this protein. We then infected cells with pure D151 viruses, pure G151 viruses, or an equal mix of both variants at a total multiplicity of infection (MOI) of 0.2. One hour post-infection, we washed the cells to remove residual oseltamivir and then monitored viral replication. These experiments were performed in full biological triplicate, beginning with triplicate independent creations of each pure population by reverse genetics.

Once again, the mixed populations consistently grew more rapidly and reached higher maximal titers than either pure population (**Figure 2.2B**). The trends in the direct co-infections were similar to those observed when generating the viruses by reverse genetics. The pure G151 populations grew very poorly, again showing that this variant has very low fitness on its own. The pure D151 populations grew reasonably well on their own, but the mixed populations grew even better. These results show that cooperation between the D151 and G151 variants improves growth of the overall population.

Interestingly, viral titers increased sharply late in the passage in some G151 populations. One possibility is that *de novo* mutations to the D151 variant create a mixed population with higher fitness. To explore the possibility of *de novo* emergence of cooperation, we serially passaged pure and mixed populations as described below.

2.1.3 Serial passage selects for mixed populations of D151 and G151 viruses.

If the D151 and G151 variants cooperate, then we expect mixed populations to emerge by *de novo* mutation and to be stably maintained when they already exist. To test this prediction, we serially passaged pure and mixed viral populations and performed targeted deep sequencing of the NA gene at the end of each passage to assess changes in allele frequency at site 151. We again used reverse genetics to generate triplicate pure populations of D151 and G151 viral variants in the presence of 50nM oseltamivir, then infected cells with D151 viruses, G151 viruses, or an equal mix of the two at a total MOI of 0.2, washing the cells one hour post-infection to remove residual oseltamivir. We verified that the D151 and G151 populations used to inoculate the first passage were pure within our limit of detection of approximately 1%, which we determined by deep-sequencing pure plasmid. We performed a total of five serial passages for each replicate, in each case seeding the new passage with the supernatant from the previous one at a total MOI of 0.2.

The mixed D151+G151 populations maintained an approximately equal mix of the two variants through all five passages (**Figure 2.5**). In the pure populations, the opposite variant arose by *de novo* mutation, then rose in frequency as the population converged towards a roughly equal mix of the two variants. The D151 variant emerged rapidly during passage of the G151 populations, exceeding a frequency of 20% by the

end of the second passage in all three replicates. The G151 variant was slower to arise in the D151 populations but had reached a substantial frequency by the end of passage 4 in all three replicates. The changes in allele frequency during serial passage demonstrate that selection acts to balance the proportion of these two genotypes in the population.

In one of the D151 populations, N151 also emerged spontaneously, and by the end of passage 5, the population consisted of a mix of D151, N151, and G151. Like G151, N151 commonly occurs in mixed populations with D151 in sequences in the GISAID EpiFlu database (**Figure 2.1**) and is mentioned in reports of mutations at site 151 in cell culture (**Table 2.1**) (Chambers et al., 2014; Lee et al., 2013; Lin et al., 2012; McKimm-Breschkin et al., 2003; Mishin et al., 2014; Mohr et al., 2015; Okomo-Adhiambo et al., 2010; Tamura et al., 2013). We verified that N151 cooperates with D151 by creating the N151 variant of the Hanoi/2007 NA and generating pure and mixed populations by reverse genetics (**Figure 2.6**). N151 viruses behave similarly to G151 viruses: they grow very poorly on their own, but cooperate with D151 to outgrow either pure population.

These results show that serial passage selects for mixed populations of D151 and G151 variants, even when the starting population is isogenic. Furthermore, mixed populations are stably maintained; the G151 variant does not sweep to fixation, as would be expected for a simple lab adaptation. Cooperation between the D151 and G151 variants evidently selects for the generation and maintenance of a genetically diverse quasispecies.

2.1.4 The dynamics of cooperation depend on the multiplicity of infection.

Since each influenza virion typically packages only a single copy of the NA gene, co-infection of a cell by multiple viruses is likely to increase opportunities for interactions among viral variants. **Figure 2.2** shows that at an MOI of 0.2, the mixed populations of D151 and G151 variants have an advantage over pure populations. At a lower MOI, co-infection is less likely. We therefore sought to test whether cooperation also promotes growth at higher and lower MOIs. We infected cells with pure and mixed viral populations in biological triplicate at an MOI of 0.02 and an MOI of 0.5, and then monitored viral titers over the next 40 hours as in **Figure 2.2**.

At an MOI of 0.02, the mixed populations grew similarly to or slightly worse than the D151 populations for the first 24 hours post-infection (**Figure 2.7A**). Later in the infection, however, the mixed populations grew to substantially higher titers than the D151 populations. In contrast, at MOIs of 0.2 (**Figure 2.2B**) and 0.5 (**Figure 2.7B**), the mixed populations grew better than the pure populations throughout the entire infection.

We note that the effective MOI of an infection increases as the infection progresses as newly produced viruses accumulate in the supernatant (Wilke et al., 2004). For the infections inoculated at an MOI of 0.02, the sharp increase in titers for the mixed population late in the infection are likely a result of this higher effective MOI. We therefore conclude that the dynamics of cooperation depend on the multiplicity of infection, with the cooperative effect decreased at lower MOIs.

2.1.5 Changes in HA potentiated cooperation between the NA variants.

Mutations at NA site 151 become common in the EpiFlu database only starting in 2007 (**Figure 2.1A**), suggesting that other mutations to the influenza genome around

that date might have affected the potential for cooperation among NA variants at site 151. A candidate gene for these potentiating mutations is HA. Good viral growth requires a balance between the receptor binding of HA and the receptor cleaving of NA (Gulati et al., 2005; Neverov et al., 2015; Wagner et al., 2002). For reasons that remain unclear, the HAs of recent human H3N2 influenza have lost their affinity for many types of sialic acid (Gulati et al., 2013; Lin et al., 2012). We therefore hypothesized that recent mutations in HA might have potentiated cooperation by making it advantageous for viral populations to acquire the NA-mediated receptor-binding of the G151 variant (Zhu et al., 2012) to compensate for reduced HA binding.

To test this hypothesis, we examined the effects of the D151 and G151 NA variants in viruses that had the HA of an earlier H3N2 strain, A/Wisconsin/67/2005, henceforth referred to as Wisconsin/2005. We cloned the HA gene from the Wisconsin/2005 strain into a reverse-genetics plasmid and generated pure and mixed populations of D151 and G151 variants in the genetic background of either the Hanoi/2007 HA or the Wisconsin/2005 HA. Cooperation between the D151 and G151 variants was eliminated in the Wisconsin/2005 HA background (**Figure 2.8**). Therefore, some of the changes to HA that distinguish the Wisconsin/2005 and Hanoi/2007 homologs are important for potentiating cooperation between the NA variants.

If decreased HA receptor-binding potentiates cooperation between the receptor-cleaving D151 and receptor-binding G151 NA variants, then viral growth should depend entirely on this cooperation if NA is the only protein able to bind the receptor. To test this hypothesis, we used an HA that has been heavily engineered to eliminate its receptor-binding activity (Hooper and Bloom, 2013). We used reverse genetics to

generate pure and mixed populations of D151 and G151 NA variants paired with this binding-deficient HA, and we measured viral titers (**Figure 2.9**). In the absence of HA receptor binding, neither the D151 nor the G151 variant alone reached appreciable titers. However, the mixed population was still able to grow with the binding-deficient HA. These results show that cooperation becomes obligate in the absence of HA receptor binding, presumably because NA must serve as the sole source of both binding and cleaving.

2.2 Discussion

We have shown that cooperation between two distinct variants of human H3N2 influenza promotes viral growth in cell culture. These variants differ by a single amino-acid mutation in NA, and each variant is present in many human H3N2 isolates that have been analyzed by Sanger sequencing after passage in the lab. Prior work has assumed that the less common G151 variant is a lab-adaptation mutant that emerges as the more common D151 variant is passaged in cell culture. Our work shows, however, that evolution in cell culture selects for a balanced mix of both variants. The G151 variant can barely replicate alone, but it cooperates with the D151 variant to increase population fitness. After multiple serial passages, both pure and mixed populations converge to an equilibrium in which both variants are present at approximately equal frequencies. Our work therefore represents a clear example of cooperation between distinct variants in a viral quasispecies.

We propose that cooperation arises because one variant is proficient at cell entry, while the other is proficient at cell exit. Viruses with wild-type D151 NAs always

exit cells efficiently, since their NAs cleave sialic-acid receptors to facilitate viral release. However, the HAs of recent human H3N2 strains have reduced affinity for many sialic-acid receptors (Gulati et al., 2013; Lin et al., 2012), reducing the efficiency with which those viruses can attach to many cells via HA. G151 viruses are proficient at cell entry, since their NA binds strongly to sialic acid, but they cannot detach effectively from host cells due to a lack of catalytic activity (Zhu et al., 2012). But in combination, D151 and G151 enable both efficient cell exit and entry. Indeed, our experiments with a binding-deficient HA indicate that in a mixed D151 and G151 population, NA can act as the exclusive source of both receptor binding and receptor cleaving (**Figure 2.9**). Our results evoke prior work showing that fitness in a Coxsackie virus population is enhanced by the combination of multiple receptor-binding variants (Bordería et al., 2015).

How do the D151 and G151 variants collaborate to enable both viral entry and exit at the level of individual virions? Co-infection of the same cell with both D151 and G151 variants would produce progeny that have both NA variants on their surface, even though each new virion would package only a single copy of the NA gene. We suspect that much of the observed cooperation may result from co-infections that produce such mixed-NA virions, which would then carry proteins that make them proficient at both cell entry and cell exit. We found that MOI affects cooperative dynamics, supporting this interpretation (**Figure 2.7**). However, other mechanisms could also contribute. In a mixed population, D151 viruses may cleave G151 viruses from the cell surface without both protein variants being present on the same infectious particle, since sialidase activity can promote viral growth in *trans* (Liu and Air, 1993). More detailed molecular

characterization of virions in mixed populations will be necessary to establish the exact mechanism of cooperation.

It remains unclear whether the D151 and G151 variants cooperate in clinical infections. When we analyzed the passage histories of sequenced isolates, we found that mixed populations were reported almost exclusively in isolates that had been passaged in cell culture (**Figure 2.1C**). Several groups have reported that mutations emerge at NA site 151 when clinical isolates are expanded in cell culture – but with one exception (Lin et al., 2010), Sanger sequencing or pyrosequencing of matched clinical and passaged isolates has so far failed to detect variation in site 151 in unpassaged isolates (Chambers et al., 2014; Lee et al., 2013; Mishin et al., 2014) (see also **Table 2.1**). However, the sequencing methods used by these studies are relatively insensitive to low-frequency variation. Given how quickly and frequently D151 mutations sometimes arise—one group found that nearly a quarter of isolates showed variation at appreciable frequencies after a single passage in MDCK cells (Lee et al., 2013)—pre-existing variation at site 151 in the original clinical isolates could contribute to the observed evolution. More sensitive deep sequencing of unpassaged clinical isolates will be necessary to resolve these questions.

Our work demonstrates that cooperation between distinct viral variants can enhance the population's overall fitness. This cooperation is not a rare event; the cooperating variants that we describe emerge rapidly and repeatedly, both in our own experiments and apparently in hundreds of clinical isolates passaged by numerous labs. Our work emphasizes that genetic diversity in viral populations can be more than a transient state that facilitates adaptation: it can itself be a beneficial trait that is

generated and maintained by selection. As the deep sequencing of viruses becomes increasingly common in microbiology and epidemiology, it will be important to better understand the broader role that cooperation plays in the evolution and maintenance of population-level diversity.

2.3 Materials and Methods

2.3.1 Analysis of GISAID EpiFlu sequences

We downloaded the set of 15079 sequences in the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database (Bogner et al., 2006) corresponding to all full-length NA coding regions from human H3N2 influenza A isolates collected from January 1, 2000 to December 31, 2014. We pairwise aligned each sequence to the A/Hanoi/Q118/2007 (H3N2) coding sequence (Genbank accession CY104446) using the program needle from EMBOSS version 6.6.0 (Rice et al., 2000), which implements a Needleman-Wunsch alignment. We identified the genotype of each sequence at site 151 and parsed the sequence metadata to determine the year in which it was collected and the sequence's passage history. Mixed genotypes were assigned on the basis of IUPAC nucleotide ambiguity codes; for instance, the triplet GRT could refer to GAT or GGT, corresponding to amino acids aspartic acid (D) and glycine (G), respectively. We occasionally observed the triplet RRT, which could correspond to a mix of aspartic acid (D; GAT), glycine (G; GGT), asparagine (N; AAT), and serine (S; AGT). We chose to annotate triplet RRT as a mix of D, G, and N, given that this mixed population has been previously observed by multiple groups (Mishin et al., 2014; Mohr et al., 2015; Tamura

et al., 2013), whereas serine is two mutations away from the D consensus identity and is not present in the H3N2 GISAID sequences that we analyzed.

Passage histories are not recorded in a standardized fashion and are frequently missing altogether. In parsing the passage histories of isolates in the EpiFlu database, therefore, we sought only to sort sequences into broad categories of which we could be reasonably certain: egg-passaged, cell-culture-passaged, and unpassaged isolates. For instance, sequences with passage annotations containing “MDCK,” “SIAT,” “RHMK,” “MEK,” and various other cell-culture signifiers were combined into the broad category of cell-culture-passaged isolates. Our exact parsing procedures and the computer code used for analysis are available at https://elifesciences.org/articles/13974.

2.3.2 Viral strains

HA and NA sequences from the A/Brisbane/10/2007 (H3N2) strain were cloned into the bidirectional pHW2000 backbone to generate virus by reverse genetics (Hoffmann et al., 2000). We performed site-directed mutagenesis on the HA and NA to match the amino-acid (but not nucleotide) sequence from A/Hanoi/Q118/2007 (Genbank accessions AEX34134 and AEX34137 for the HA and NA, respectively), which has a G at NA site 151 (Bao et al., 2008). The HA and NA protein sequences are identical to those from A/California/UR06-0565/2007 (Genbank accessions ABW40191 and ABW40194 for HA and NA, respectively) aside from a single site in HA, as well as the genotype at NA site 151. We performed further rounds of site-directed mutagenesis to generate the D151 and N151 variants of the NA. The HA sequence from the A/Wisconsin/67/2005 (H3N2) influenza strain (Genbank accession CY163744) was similarly cloned into the bidirectional pHW2000 backbone for reverse genetics. The

binding-deficient HA is derived from the A/Hong Kong/2/1968 (H3N2) HA and contains extensive mutations and deletions that eliminate receptor-binding activity; this is the variant referred to as the “PassMut HA” in (Hooper and Bloom, 2013). Coding sequences for the HA and NA genes used in this study are available in Supplementary file 1.

The remaining six viral genes were expressed from bidirectional reverse-genetics plasmids derived from the A/WSN/33 strain (pHW181-PB2, pHW182-PB1, pHW183-PA, pHW185-NP, pHW187-M, and pHW188-NS) and were kind gifts from Robert Webster of St. Jude Children’s Research Hospital. For all experiments not otherwise indicated, we used a plasmid (PB1flank-eGFP) that carried GFP flanked by PB1 packaging signals in place of pHW182-PB1 plasmid, and propagated the viruses in 293T and MDCK-SIAT1 cells expressing the WSN PB1 under the control of a CMV promoter as described in (Bloom et al., 2010).

2.3.3 Viral reverse genetics

To generate GFP-expressing virus using reverse genetics, we transfected co-cultures of 293T-CMV-SIAT-PB1 and MDCK-SIAT1-CMV-PB1 cells with plasmids encoding the eight viral genes, with PB1flank-GFP rather than PB1 as described in (Bloom et al., 2010). We plated 2 x 10^5 293T-CMV-PB1 cells and 0.2 x 10^5 MDCK-CMV-PB1 cells per well in six-well dishes in D10 (Dulbecco modified Eagle medium supplemented with 10% heat-inactivated fetal bovine serum [FBS], 2mM L-glutamine, 100 U/mL penicillin, and 100 µg/mL streptomycin) and transfected each well with 2 µg plasmid DNA, corresponding to 250ng of each of the eight plasmids, using the BioT transfection reagent (Bioland Scientific, Paramount, California). At 12 to 18 hours post-

transfection, the cells were washed once with phosphate-buffered saline (PBS), and the media was changed to low-serum influenza growth media (IGM; Opti-MEM supplemented with 0.01% heat-inactivated FBS, 0.3% bovine serum albumin, 100 U/mL penicillin, 100 µg/mL streptomycin, and 100 µg/mL calcium chloride). TPCK (toylsulfonyl phenylalanyl chloromethyl ketone)-trypsin was added to IGM at 3 µg/mL immediately before use. For reverse genetics carried out in the presence of oseltamivir, we added the indicated concentration of oseltamivir carboxylate (kindly provided by Roche) to the IGM at this point as well. We collected viral supernatant at 72 hours post-transfection, clarified by centrifugation at 285xg for 4 minutes, aliquoted, and froze at -80 degrees C before thawing aliquots for titering. To generate viral populations that expressed the A/WSN/33 PB1 gene rather than PB1 segment packaging GFP, we substituted 293T and MDCK-SIAT1 cells for 293T-CMV-PB1 and MDCK-SIAT1-CMV-PB1 cells in the protocol above.

2.3.4 Viral titering

For viruses grown with the PB1flank-eGFP gene, titers were determined using flow cytometry. We plated 10^5 MDCK-SIAT1-CMV-PB1 cells per well in 12-well plates in IGM and infected them 4-6 hours later with 0.1, 1, 10, or 100µL of viral supernatant. At 16 hours post-infection, we collected the cells into PBS with 1% paraformaldehyde from wells in which approximately 1-10% of cells were GFP-positive and used flow cytometry to determine the exact proportion of GFP-positive cells. We used the Poisson equation to calculate the number of infectious particles in the original inoculum as:

*[titer, in infectious particle per µL] = -log(1 – [fraction of GFP-positive cells]) * [number of cells plated, in this case 10^5] / [inoculum volume, in µL]*

For viruses grown with the WSN PB1 gene, titers were determined by staining for intracellular NP. Similar to the GFP titering described above, MDCK-SIAT1 cells were infected with serial dilutions of viral supernatant. At 12 hours post-infection, the cells were collected, fixed and permeabilized with the BD Cytofix/Cytoperm kit (product number 554722, BD Biosciences, Franklin Lakes, New Jersey) following the manufacturer's protocol but omitting the GolgiPlug, stained with a 1:20 dilution of mouse anti-NP FITC-conjugated antibody (clone A1 from MAB8257F, EMD MilliPore, Darmstadt, Germany), washed twice, and analyzed by flow cytometry to count NP-positive cells. The viral titer was computed from the fraction of positive cells using the Poisson equation above. All titers are plotted with the lower bound of the y-axis set at the limit of detection of this assay, approximately 10^{-1} infectious particles/µL.

Note that all of these titering methods quantify the number of virions that enter cells and express a functional polymerase complex that produces large amounts of the mRNA encoding the protein product being detected (GFP or NP). Unlike in a TCID50 or plaque assay, not all detected virions are necessarily able to undergo multi-cycle infections.

2.3.5 Viral serial passage in cell culture

For each passage, we plated 10^5 MDCK-SIAT1-CMV-PB1 cells per well in six-well plates in IGM and infected them 4-6 hours later with D151 viruses, G151 viruses, or an equal mix of the two at a total MOI of 0.2. An hour after the viruses were added, we washed the cells with PBS and added fresh IGM supplemented with 3 µg/mL TPCK-

trypsin to dilute the effect of any oseltamivir remaining from reverse genetics for the first passage. We collected viral supernatant at 40 hours post-infection, clarified by centrifugation at 285xg for 4 minutes, aliquoted, and froze it at -80 degrees C before thawing aliquots for titering. We collected cells remaining at the end of each passage in 1mL Trizol reagent and froze them at -20 degrees C. From passage 4 onwards, mutation accumulation in the PB1flank-eGFP gene caused widespread loss of GFP in many viral populations; by the end of this passage, cytopathic effect was clearly visible even though GFP fluorescence was not. To inoculate passage 5, we infected cells with 5uL viral supernatant from passage 4, a volume corresponding approximately to an MOI of 0.2 based on the titers for earlier passages.

2.3.6 Targeted deep sequencing of the NA gene

We extracted RNA from cells remaining at the end of each passage using Trizol reagent and performed reverse-transcription using the primers CAGGAGTGAAAATGAATCCAAATCAAAAGATAATAACGATTG and TTGCGAAAGCTTATATAGGCATGAGATTGATG, which target the full-length NA gene. We then used primers CTTTCCCTACACGACGCTCTTCCGATCTxxxCAACACTAAACAACGTGCATTCAAATGAC and GGAGTTCAGACGTGTGCTCTTCCGATCTCCTAACTCATTCATCAATAGGGTCCGATAAGG to amplify a targeted region of the NA gene surrounding site 151 and add the first half of the Illumina sequencing adaptor and a three-mer in-read barcode, represented here as xxx. We performed 25 cycles of amplification at an annealing temperature of 55 degrees C and an extension time of 40 seconds. We purified the

PCR product using 1.5X Ampure beads and used this product as template for a second round of PCR using primers AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC and CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTGGAGTTCAGACGTGTGCTCTTCC, which add the second half of the Illumina sequencing adaptors and a six-mer barcode, represented here as xxxxxx.

To sequence the viral stocks used to inoculate the first passage (these inoculating stocks are referred to as "Passage 0" in **Figure 2.5**), we plated 10^5 MDCK-SIAT1-CMV-PB1 cells per well in six-well plates in IGM and infected them 4-6 hours later with viral stocks at an MOI of 0.02. An hour after the viruses were added, we washed the cells with PBS and added fresh IGM to dilute the effect of any oseltamivir remaining from the generation of virus by reverse genetics. We collected the cells in 1mL Trizol reagent at 16 hours post-infection and froze them at -20 degrees C. The purpose of this inoculation was to ensure that sequencing of viral stocks detected only infectious particles. We then prepared PCR amplicons from these samples for deep sequencing as described above.

Reads were first screened to verify sequencing quality and correct identity. Reads were discarded if any position had a Q-score below 25 or if the read had more than 4 mismatches relative to the plasmid reference sequence. We translated the reads in the NA reading frame and tallied the amino-acid identities at each position, recording "X" for positions with a discrepancy between the forward and reverse reads. The FASTQ files and the computer code used to analyze them are available at https://elifesciences.org/articles/13974.

2.3.7 Acknowledgements

We thank Roche for providing the oseltamivir carboxylate used in our experiments, Juhye Lee for cloning the A/Wisconsin/67/2005 hemagglutinin gene, Ian Gong for assistance with preliminary work, Choli Lee and Seungsoo Kim for assistance with sequencing, and Michael Emerman and Wenying Shou for helpful comments on the manuscript. This work was supported by the NIGMS of the NIH under grant R01GM102198. KSX was supported by an NSF Graduate Research Fellowship under grant number DGE-1256082 and a fellowship from the Fannie and John Hertz Foundation. KAH was supported by NRSA training grant T32GM007270.

2.4 FIGURES AND TABLES

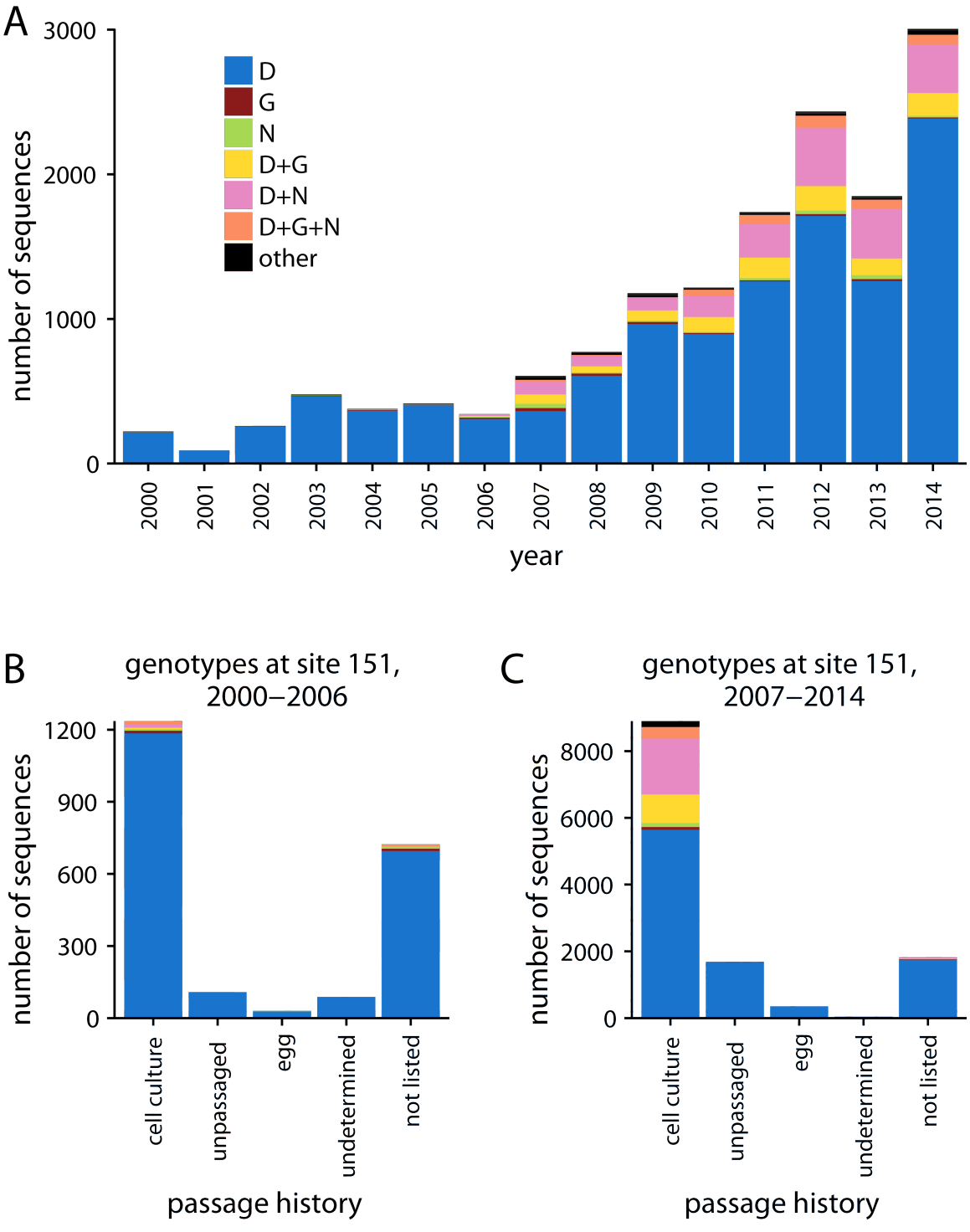

Figure 2.1. Ambiguous identities are common at NA site 151 after 2007.

(A) Shown are the number of human H3N2 influenza NA sequences in the GISAID EpiFlu database with the given identity at site 151 for each year from 2000 to 2014. Since 2007, ambiguous amino-acid identities have been present at residue 151 in about 20% of sequences. Sequences from **(B)** 2000 to 2006 and **(C)** 2007 to 2014 were classified into groups based on their passage history. Ambiguous amino-acid identities were present almost exclusively in isolates that had been passaged in cell culture. Sequences were classified as “undetermined” if the passage history was difficult to interpret and as “not listed” if the passage history was absent altogether. Mixed genotypes were inferred on the basis of IUPAC nucleotide ambiguity codes; for instance, the triplet GRT could refer to GAT or GGT, corresponding to amino acids D and G, respectively. Genotypes are indicated if they exceeded a frequency of 0.5% among all analyzed sequences; otherwise, they are categorized as “other.” The computer code used for analysis is available at https://elifesciences.org/articles/13974.

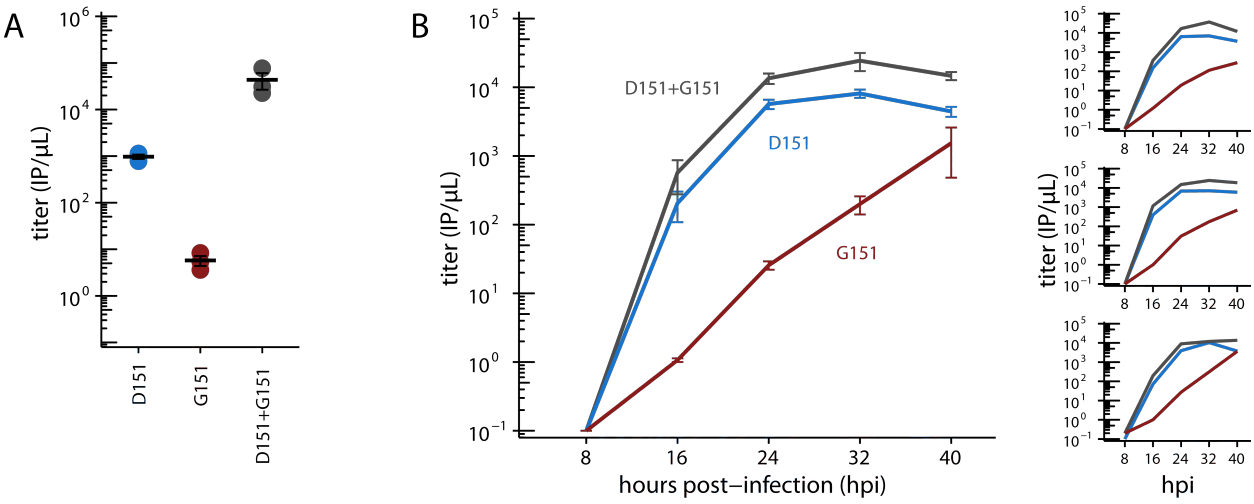

Figure 2.2. Mixed populations grow to higher titers than either pure population alone.

(A) Pure and mixed populations were generated by reverse genetics. Cells were transfected with a Hanoi/2007 NA plasmid encoding D151, G151, or an equal mix of the two, along with isogenic plasmids for the other genes. The total amount of NA plasmid was the same in all cases; that is, the pure populations were transfected with 250 ng of the indicated variant, and the mixed populations were transfected with 125 ng of each variant. The HA was also derived from Hanoi/2007, and the other genes were derived from the lab-adapted A/WSN/33 strain with GFP packaged in the PB1 segment. The titer was determined after 72 hours using the GFP reporter. Black lines indicate the mean and standard error of the titers for three biological replicates, with titers for each replicate plotted as points. Figure 2.3 shows a comparable effect when the virus does not package GFP. Figure 2.4 shows that growth of the G151 variant is improved by adding oseltamivir. **(B)** Cells were infected at an MOI of 0.2 with pure D151 virus, pure G151 virus, or an equal mix of the two. The total MOI of infecting virus was the same in all cases. The main plots show titers averaged across three biological replicates, with each replicate plotted individually in the small insets.

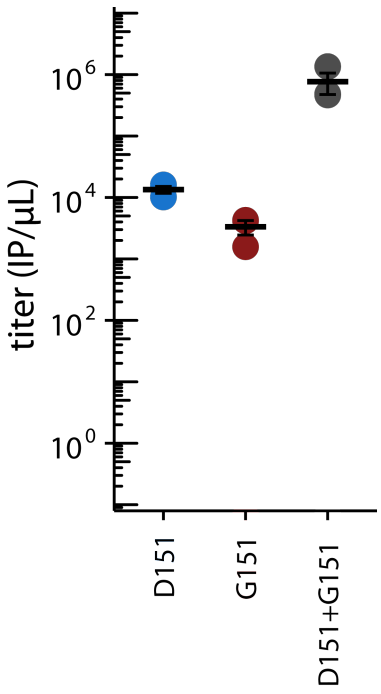

Figure 2.3. A mixed population outgrows either pure population when viruses are generated by reverse genetics with an unmodified PB1 gene.

The data here differ from Figure 2.2A in that the virus populations were generated by reverse genetics using the unmodified A/WSN/33 PB1 gene rather than the PB1 segment modified to package GFP. Titers were determined at 74 hours post-transfection by staining for nucleoprotein in infected cells. Black lines indicate the mean and standard error of the titers for three biological replicates, with titers for each replicate plotted as points.

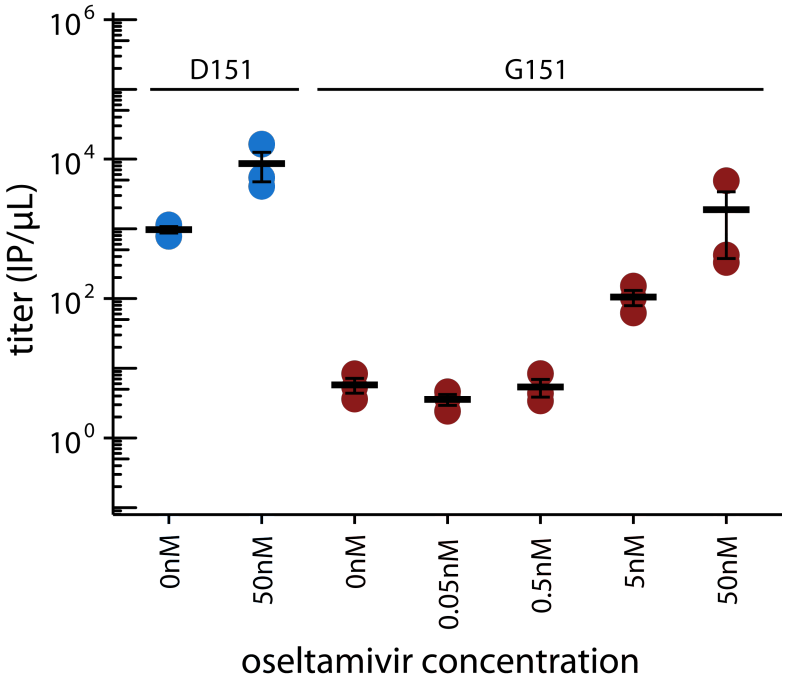

Figure 2.4. Growth of the G151 variant is improved by adding oseltamivir during the generation of viral populations by reverse genetics.

Presumably, this improvement occurs because oseltamivir blocks the binding of G151 NA to receptor, allowing newly formed virions to be released more efficiently. Oseltamivir also slightly increases the growth of the D151 variant. We speculate this is because oseltamivir blocks receptor cleavage by the D151 NA, leaving more receptors that can be bound by HA during secondary viral replication.

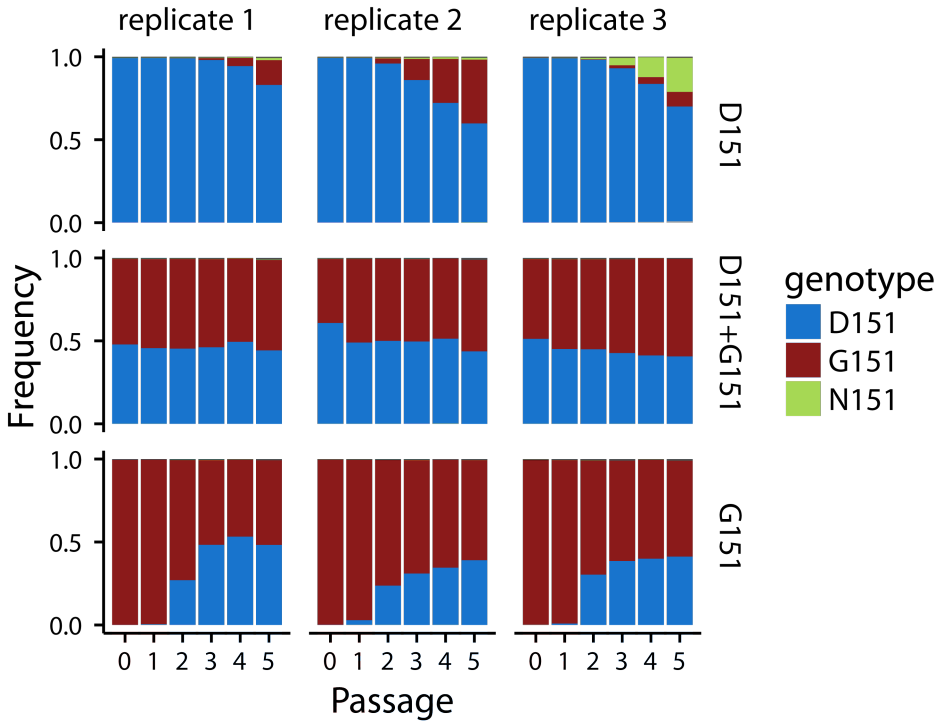

Figure 2.5. Serial passage selects for a stable mix of the two variants.

Shown are the allele frequencies at NA site 151 over five tissue-culture passages of initially pure D151 viruses, pure G151 viruses, or an equal mix of the two. Each passage was seeded at a total MOI of 0.2. Passage 0 refers to the ratio of variants in the viral inoculum for passage 1. Allele frequencies were determined by targeted Illumina deep-sequencing of the NA gene. Based on sequencing of pure plasmid, the error rate was less than 1%. The raw data and computer code are available at https://elifesciences.org/articles/13974.

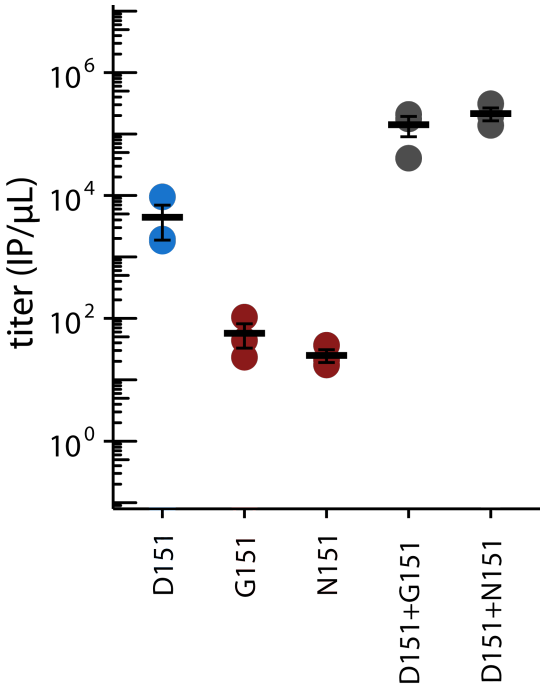

Figure 2.6. The N151 variant also cooperates with D151.

Shown are the titers after reverse genetics with the indicated variant of the Hanoi/2007 NA. The experiments here parallel those in Figure 2A. Black lines indicate the mean and standard error of the titers for three biological replicates, with titers for each replicate plotted as points.

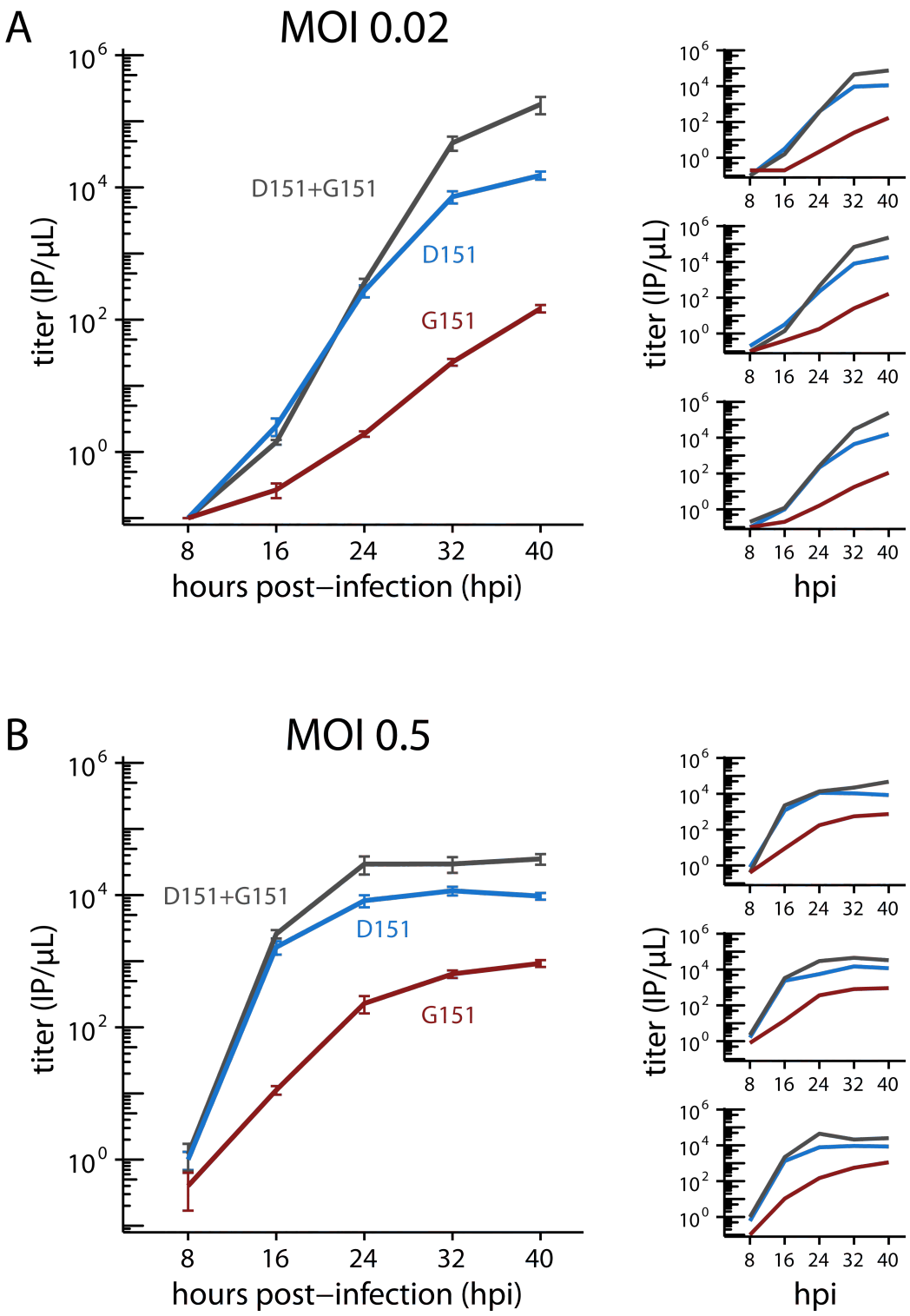

Figure 2.7. Cooperative dynamics depend on multiplicity of infection.

Cells were infected at an MOI of **(A)** 0.02 or **(B)** 0.5 with pure D151 virus, pure G151 virus, or an equal mix of the two. The total MOI of infecting virus was the same across the mixed and pure populations for infection at each MOI. The main plots show titers

averaged across three biological replicates, with each replicate plotted individually in the small insets. The experiments here parallel those in Figure 2.2B. Black lines indicate the mean and standard error of the titers for three biological replicates, with titers for each replicate plotted as points.

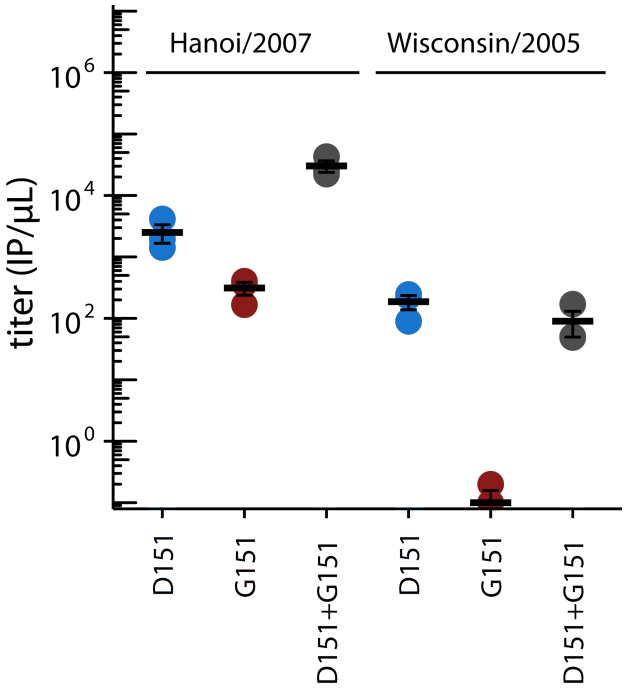

Figure 2.8. Changes in HA between 2005 and 2007 potentiated cooperation.

Cooperation occurs between the D151 and G151 NA variants in viruses with HA from the Hanoi/2007 strain, but not in viruses with HA from the Wisconsin/2005 strain. Shown are the titers after reverse genetics with the indicated HA and NA. The experiments here parallel those in Figure 2.2A. Black lines indicate the mean and standard error of the titers for three biological replicates, with titers for each replicate plotted as points.

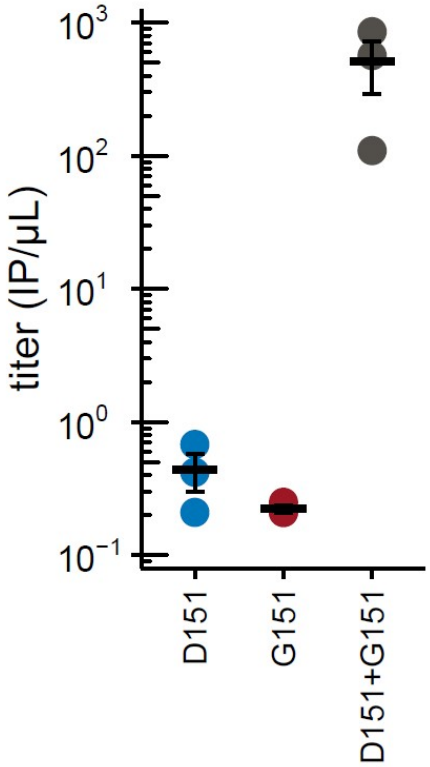

Figure 2.9. Cooperation is obligate when HA lacks receptor-binding activity.

Shown are the titers after reverse genetics with the indicated Hanoi/2007 NAs in combination with an engineered binding-deficient H3 HA with multiple mutations to the receptor-binding pocket (Hooper and Bloom, 2013). The experiments here parallel those in Figure 2.2A. Black lines indicate the mean and standard error of the titers for three biological replicates, with titers for each replicate plotted as points.

Table 2.1. Prior reports of variation at neuraminidase site 151 when H3N2 clinical specimens are passaged in cell culture.

Reference	Summary
(McKimm-Breschkin et al., 2003)	Sanger sequencing of 38 oseltamivir- and zanamivir-resistant MDCK-passaged clinical isolates found that 7 had G, N, E, or V at site 151.
(Lin et al., 2010)	Sanger sequencing of 18 isolates after passage in MDCK or MDCK-SIAT1 cells found 4 isolates as D+G, 3 as D+N, and 2 as D+A. Pyrosequencing detected low frequencies of G151 and N151 in some clinical samples.
(Tamura et al., 2013)	Pyrosequencing of 150 isolates after 1-4 passages in MDCK cells found that 85% developed mixed populations at site 151; 29% did so after a single passage. Mixed populations consisted of D+N, D+G, D+G+N, and D+G+A genotypes. T148I/K/P mutations were also observed in 23% of isolates.
(Lee et al., 2013)	77 clinical specimens were Sanger-sequenced before and after a single passage in MDCK cells. 18 acquired a mutation at NA site 151: 10 were D+N, 7 were D+G and one fixed D151N at the limit of detection. No mutations were detected in the unpassaged specimens.
(Chambers et al., 2014)	9 A/Victoria/361/11-like clinical specimens were passaged twice in MDCK cells and Sanger-sequenced before and after expansion. 4 isolates developed NA-dependent cell binding; 3 had D151G, the other D151N.
(Mishin et al., 2014)	Pyrosequencing of 150 MDCK-grown isolates found that 42 were D+G, 34 were D+N, and 57 were D+G+N. Pyrosequencing of 50 matched clinical specimens detected no variation at site 151.
(Mohr et al., 2015)	16 pairs of isolates cultured in parallel in MDCK cells and in eggs were sequenced using Ion Torrent. 5 MDCK isolates were D+N, 4 were D+G, and 2 were D+N+G. No egg-passaged isolates had mutations at site 151. T148I/K mutations were observed in 7 MDCK isolates.

Chapter 3. COOPERATING H3N2 INFLUENZA VIRUS VARIANTS ARE NOT DETECTABLE IN PRIMARY CLINICAL SAMPLES

A version of this chapter has previously been published as:

Xue, K.S., Greninger, A.L., Pérez-Osorio, A., Bloom, J.D. Cooperating H3N2 influenza virus variants are not detectable in primary clinical samples. mSphere 3: e00552-17 (2017). DOI: 10.1128/mSphereDirect.00552-17

RNA viruses like influenza mutate rapidly to form genetically diverse quasispecies. Several recent studies have suggested that interactions between different variants in a quasispecies can promote overall population fitness. In poliovirus, variants generated through spontaneous mutation are important for neurotropism, innate immune suppression, and overall pathogenesis in mouse models (Pfeiffer and Kirkegaard, 2005; Vignuzzi et al., 2006; Xiao et al., 2017). Other groups have identified cooperative interactions in measles virus (Shirogane et al., 2012), West Nile virus (Ciota et al., 2012), hepatitis B virus (Cao et al., 2014), and Coxsackie virus (Bordería et al., 2015). These cooperative interactions have primarily been observed in cell culture or animal models rather than clinical infections.

We previously described two distinct variants of H3N2 influenza virus that cooperate in cell culture (Xue et al., 2016). The two variants differ by a single mutation at amino acid 151 of neuraminidase (NA), the protein that releases new virions from host cells. The D151 viral variant, typically encoded as `GAT`, predominates among

clinical influenza samples, and it grows robustly in cell culture. The G151 viral variant, typically encoded as GGT, binds sialic-acid receptors rather than cleaving them (Lin et al., 2010; Zhu et al., 2012) and grows extremely poorly in isolation. However, a mixed population of D151 and G151 viral variants outgrows either single variant in cell culture.

An important question is whether cooperation between these two viral variants is purely a cell-culture phenomenon, or whether the D151 and G151 variants co-exist in natural infections. The D151G mutation is frequently observed when influenza virus is passaged through cell culture (Chambers et al., 2014; Lee et al., 2013; Lin et al., 2010; McKimm-Breschkin et al., 2003; Mishin et al., 2014; Mohr et al., 2015; Tamura et al., 2013), but it remains unclear whether the G151 variant exists within natural human infections or is primarily a cell-culture artifact. Prior groups that have performed matched clinical sequencing of unpassaged and passaged clinical samples have failed to detect the G151 variant before passaging (Lee et al., 2013; Mishin et al., 2014), but these studies have used methods like Sanger sequencing and pyrosequencing that are relatively insensitive to rare variation. More sensitive characterization of clinical samples that give rise to D151G upon lab passage can determine whether this mutation reaches high frequencies in cell culture because it is amplified from low- to modest-frequency standing diversity, or whether it arises spontaneously in the lab.

We sought to determine whether the D151G mutation is present in viral populations isolated from natural human infections. We identified nine clinical samples that, based on prior Sanger sequencing, consisted of a mixture of D151 and G151 viruses after passage in cell culture. We deep-sequenced the original unpassaged nasal swab samples to survey the variation present prior to laboratory growth. The D151G

mutation did not exceed the frequency of library preparation and sequencing errors in any of these samples. These results suggest that most variation observed at site 151 results from passage in cell culture rather than standing variation in human infections.

3.1 RESULTS

Most influenza-virus sequences in public databases are determined by Sanger sequencing of clinical isolates that have been passaged one or more times in cell culture (McWhite et al., 2016). A substantial number of recent human H3N2 influenza virus sequences in these databases contain an ambiguous nucleotide at NA site 151 because the lab-passaged samples often converge to a mix of the D151 and G151 variants (Xue et al., 2016). We sought to compare passaged samples that contained this ambiguous nucleotide at site 151 to unpassaged samples from the same viral infections. We first identified strains from western Washington state in the GISAID EpiFlu database (Bogner et al., 2006) for which Sanger sequencing had reported an ambiguous nucleotide at NA site 151 corresponding to a mix of the D151 and G151 variants (Xue et al., 2016). Based on the annotations available in the GISAID EpiFlu database, most of these strains had been passaged in cell culture prior to Sanger sequencing.

We obtained original, unpassaged nasal swab samples for the nine strains in **Table 3.1** that contained a mixture of D151 and G151 variants after passage in cell culture. These samples had been collected between 2013 and 2015 and had undergone one to three passages in cell culture prior to sequencing. We performed whole-genome sequencing of the influenza genome from the unpassaged clinical samples using

influenza-specific reverse transcription and PCR (Xue et al., 2017). For each sample, we prepared sequencing libraries in duplicate, beginning from separate reverse-transcription reactions (McCrone and Lauring, 2016). We sequenced each viral sample to an average sequencing depth of 100x-10,000x (**Figure 3.1**)**,** allowing us to observe viral variants at frequencies below the limit of detection of Sanger sequencing or pyrosequencing.

We identified all minor viral variants present at a frequency of at least 3% in the viral genome in both library replicates (**Table 3.2**). We did not observe the D151G variant in any of the nine clinical samples under these variant-calling criteria. To ensure that we were not missing extremely low-frequency variation, we calculated the frequency of D151G in each clinical sample based on the frequency of G-to-A mutations at the second nucleotide position of NA site 151. We compared this frequency to the frequency of G-to-A mutations at other sites across the genome (**Figure 3.2**). Minor-variant frequencies at NA site 151 fell well within the range of error expected through library preparation and sequencing errors. Therefore, we conclude that the D151G variant was not present at appreciable frequencies in the original clinical infections. Instead, the mutation must have arisen *de novo* or been enriched from an extremely low frequency during passage in cell culture.

3.2 DISCUSSION

The results of our deep-sequencing study support prior studies that failed to detect the D151G mutation in unpassaged clinical samples using Sanger sequencing or pyrosequencing methods (Lee et al., 2013; Mishin et al., 2014). In the GISAID EpiFlu

database, mixed populations of D151 and G151 viral variants are common in clinical samples that have been passaged in cell culture, but these mixed populations are rare among unpassaged and egg-passaged populations (Xue et al., 2016). It is impossible to rule out the possibility that the D151G mutation reaches appreciable frequencies in some natural human infections, but strong and repeated selection for cooperation in cell culture seems to account for its prevalence among sequences in public databases.

It is interesting to speculate about what biological factors might cause a variant that is rare in natural human infections to be strongly selected in cell culture. Influenza strains often acquire stereotypical mutations when they are grown in eggs (Brand and Palese, 1980; Skowronski et al., 2014), but these passage adaptations appear to be less common in cell culture, particularly for MDCK-SIAT1 cells (McWhite et al., 2016; Oh et al., 2008). Nevertheless, differences in the types and distributions of cell-surface receptors between MDCK-SIAT1 cells and human airways could account for some of the differences in genotypes we observe at NA site 151.

We also previously observed that cooperation is stronger at high multiplicities of infection (MOI) (Xue et al., 2016). Viral load can be high during natural infections (**Table 3.1**)**,** but recent studies of natural human infections have found that the effective reassortment rate is limited, suggesting that spatial heterogeneity within the host may limit viral circulation and co-infection (Sobel Leonard et al., 2017a). Moreover, human influenza infections, as well as those in animal models (Varble et al., 2014), experience a severe transmission bottleneck that greatly limits the genetic diversity initially present in an infection (McCrone et al., 2017; Poon et al., 2016; Sobel Leonard et al., 2017b). In contrast, viral populations can rapidly reach high MOIs in cell culture (Novella et al.,

2004). These different growth conditions may also promote the emergence of D151G within cell culture, but not natural infections.

Our study also underscores the importance of sequencing directly from unpassaged clinical samples. Mutations like D151G accumulate in cell culture within just a few passages and affect downstream analyses like inferences of positive selection (McWhite et al., 2016). Careful records of passage histories combined with deep-sequencing of unpassaged clinical samples can help distinguish natural variation from that generated in the lab.

3.3 MATERIALS AND METHODS

3.3.1 Viral samples

We downloaded the set of 66 sequences in the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database (Bogner et al., 2006) corresponding to all full-length NA coding regions from human H3N2 influenza A isolates collected from January 1, 2000 to August 26, 2015 and submitted from Seattle, Washington, or Shoreline, Washington. We pairwise aligned each sequence to the A/Hanoi/Q118/2007 (H3N2) coding sequence (Genbank accession CY104446) using the program needle from EMBOSS version 6.6.0 (Rice et al., 2000). For each sequence, we determined the genotype at site 151 and assigned the genotype X if there was an ambiguous nucleotide at that site. We identified sequences with ambiguous identities at site 151, suggesting the presence of mixed viral populations, and we extracted passage histories based on the metadata available in the GISAID Epiflu database. For the nine strains

described in **Table 3.1**, we were able to obtain aliquots of the original, unpassaged nasal swab samples in viral transport media.

3.3.2 Viral deep sequencing

We performed viral deep sequencing as previously described (Xue et al., 2017). In brief, we extracted viral RNA from unpassaged clinical samples using the QIAamp Viral RNA Mini Kit (Qiagen) according to manufacturer's instructions. We reverse-transcribed the viral RNA using the Superscript III First-Strand Reaction Mix (Thermo Fisher) and an equimolar mix of the influenza-specific primers 5'-TATTGGTCTCAGGGAGCAAAAGCAGG-3' and 5'-TATTGGTCTCAGGGAGCGAAAGCAGG-3', which both bind to the conserved U12 region at one end of each influenza gene. The two primers differ by a single nucleotide to account for a known polymorphism in the region. We incubated the reverse-transcription reactions at 25 degrees C for 10 minutes (to help the short primer anneal), 50 degrees C for 50 minutes, and 85 degrees C for 5 minutes. We amplified the influenza genome using a mixture of 24 primers that bind to the ends of each influenza gene (Hoffmann et al., 2001). For each gene, one primer binds to the conserved U13 region at one end of the gene, and two primers bind to the conserved U12 region at the other end of the gene, allowing for the known polymorphism in the U12 region. We performed 35 cycles of PCR using an annealing temperature of 55 degrees C and an extension time of 3 minutes. We purified the PCR product using 1X AMPure beads (Beckman Coulter) and prepared libraries for Illumina sequencing using Nextera XT (Illumina) tagmentation. We sequenced the libraries on a NextSeq 500 platform (Illumina) with 150 bp paired-end reads. We performed all library preparation and

sequencing in duplicate, starting from independent reverse-transcription reactions (McCrone and Lauring, 2016).

3.3.3 Analysis of deep-sequencing data

We first used bowtie2 (Langmead and Salzberg, 2012) to filter out reads that mapped to the human genome. Remaining reads are available in the SRA as BioProject PRJNA412675. We trimmed adapters from the raw reads using cutadapt version 1.8.3 (Martin, 2011). We first aligned the reads to the A/Victoria/361/2011 genome using bowtie2 and the `--very-sensitive` setting, then we used custom scripts to generate a new consensus genome sequence for each viral sample. We then re-aligned the reads to the corresponding consensus sequence and removed PCR duplicates using picard version 1.43. We used custom scripts to filter out base calls with a quality score below 20, tally the total number of high-quality bases at each genome position, and annotate each variant's codon position. We performed these initial analyses separately for each replicate library. We reported only variants that were located in protein-coding sequence.

3.3.4 A note on codon numbering and gene annotation

We numbered HA codons according to the H3 numbering system. This HA numbering scheme assigns 1 to codon 17 of the full HA gene, which is the beginning of the mature HA protein. The codons for all other genes are numbered sequentially beginning with one at the N-terminal methionine. The M1 and M2 genes have 27 bp of in-frame and 44 bp of out-of-frame overlap, and the NS1 and NEP genes have 30 bp of

in-frame and 251 bp of out-of-frame overlap. We annotated variants separately for each gene if they occurred in these regions of overlap.

3.3.5 Data and code availability

Sequencing reads are available on the SRA as BioProject PRJNA412675. The computer code that performs the analyses is available on Github at https://github.com/ksxue/D151G-clinical-public.

3.3.6 Acknowledgements

This study was funded by R01GM102198 from the NIGMS, R01AI127893 from the NIAID, a Howard Hughes Medical Institute Faculty Scholar Award, and a Simons Foundation Faculty Scholar Award to JDB. KSX was funded by an NSF Graduate Research Fellowship under grant number DGE-1256082 and a graduate fellowship from the Fannie and John Hertz Foundation. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

3.4 FIGURES AND TABLES

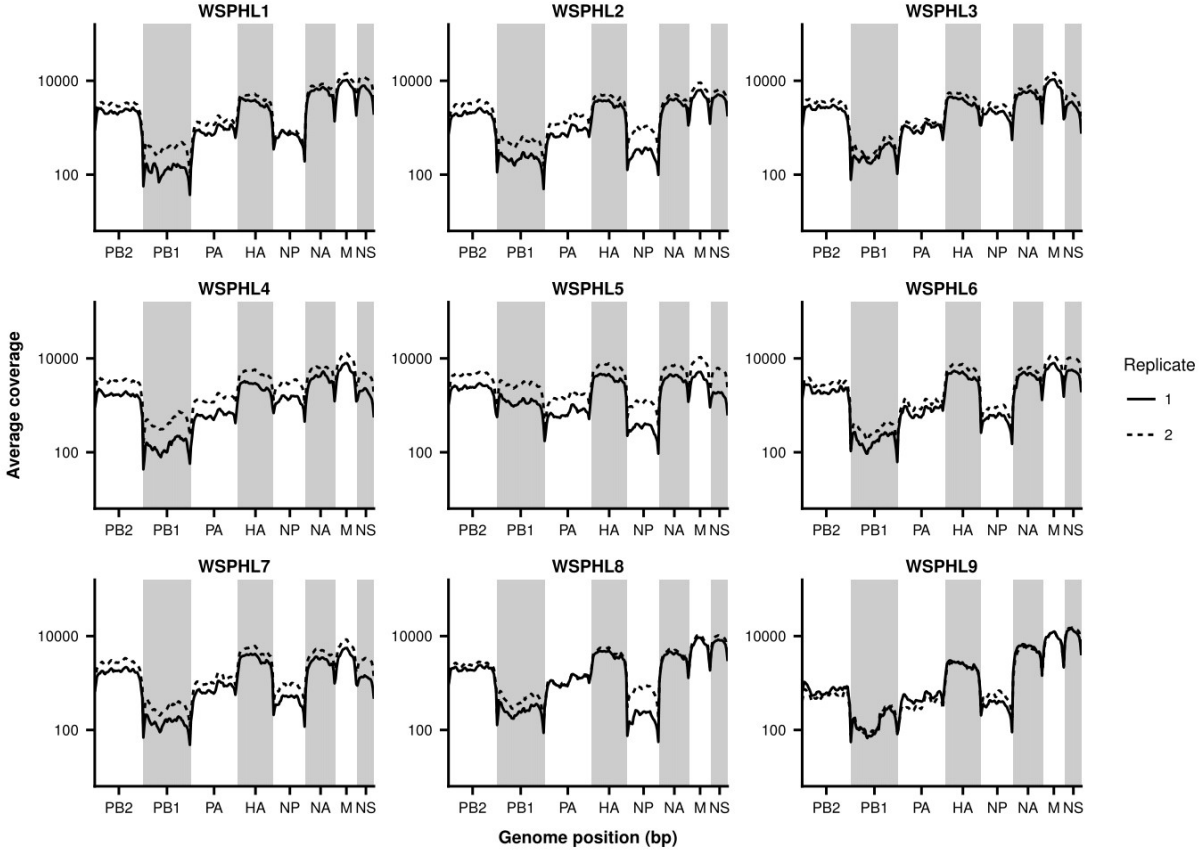

Figure 3.1. Sequencing coverage along the influenza genome.

Average sequencing coverage is plotted for 50 bp bins across the genome, with library replicates shown in solid and dashed lines.

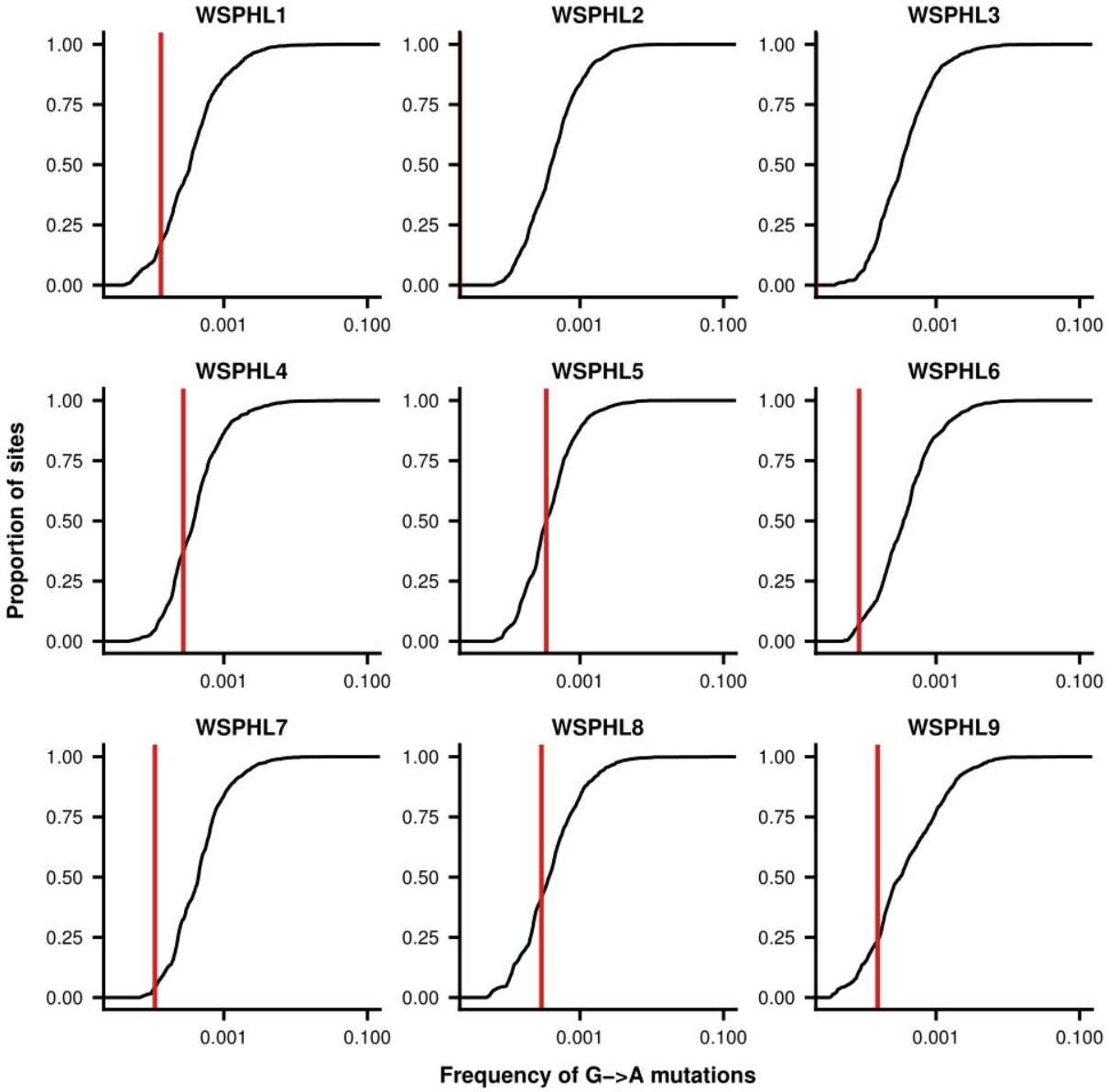

Figure 3.2. D151G does not exceed the frequency of library preparation and sequencing errors in unpassaged clinical samples.

Shown is the distribution of frequencies of G-to-A mutations across the genome for each clinical sample. Typically, the D151 viral variant is encoded by the nucleotides `GAT`, and the G151 variant is encoded as `GGT`, meaning that D151G arises as the result of a G-to-A mutation. The red vertical line shows the proportion of G-to-A mutations at codon position 2 of amino-acid site 151 of NA, which corresponds to the frequency of D151G. In cases where no G-to-A mutations were identified at this site, this red line is not shown. At each nucleotide site in the genome with consensus identity G, we calculated the total proportion of reads reporting an identity of A at that site and averaged this proportion between both replicate libraries. As expected, G-to-A

mutations make up less than 0.1% of total sequencing reads at most sites in the genome and are probably errors introduced through library preparation and sequencing.

Table 3.2. Strains deep-sequenced in this study.

Genotypes were determined through Sanger sequencing of passaged isolates, and are taken from those reported in the GISAID EpiFlu database. Annotations of passage history are not standardized, but C[N] generally refers to N passages of the virus in cell culture prior to sequencing, and S[N] generally refers to N passages of the virus in MDCK-SIAT1 cells (McWhite et al., 2016). For the genotype at site 151, an annotation of X indicates a mix of D151 and G151 in the original Sanger sequencing. The Ct value is the amount of viral material in the original clinical sample as determined by qPCR.

Sample	Strain	Passage history	Genotype, site 151	Ct Value
WSPHL1	A/Washington/10/2013	C1/C1	X	23.19
WSPHL2	A/Washington/13/2013	C1	X	17
WSPHL3	A/Washington/17/2013	C2	X	24.57
WSPHL4	A/Washington/18/2013	C3	X	21.52
WSPHL5	A/Washington/08/2014	C1	X	22.8
WSPHL6	A/Washington/07/2015	S3	X	23.78
WSPHL7	A/Washington/24/2015	S3	X	17.4
WSPHL8	A/Washington/32/2015	S3	X	18.69
WSPHL9	A/Washington/36/2015	S3	X	25.03

Table 3.3. Within-host variants identified through deep sequencing.

Sites were called as variable if a non-consensus base exceeded a frequency of 0.03, given a sequencing coverage of at least 100x, in both sequencing replicates.

Sample	Variant	Frequency
WSPHL1	NS1-G47S	0.042
WSPHL3	HA-D513Y	0.035
WSPHL4	NA-E83K	0.32
WSPHL4	PB2-E40G	0.035
WSPHL4	PB2-R175K	0.042
WSPHL4	HA-E325K	0.06
WSPHL6	PB1-M372I	0.038
WSPHL6	PB1-H562Y	0.059
WSPHL7	PB1-F254F	0.34
WSPHL7	PA-P238P	0.119
WSPHL7	HA-I202V	0.115
WSPHL7	NP-P419P	0.268
WSPHL8	NA-F42F	0.153
WSPHL8	NA-N86T	0.204
WSPHL8	PB2-M631V	0.061
WSPHL8	PB1-I392M	0.079
WSPHL8	HA-R208S	0.081
WSPHL8	HA-A425A	0.161
WSPHL9	PB1-N518N	0.248
WSPHL9	PB1-E731E	0.21

Chapter 4. PARALLEL EVOLUTION OF INFLUENZA ACROSS MULTIPLE SPATIOTEMPORAL SCALES

A version of this chapter has previously been published as:

Xue, K.S., Stevens-Ayers, T., Campbell, A.P., Englund, J.A., Pergam, S.A., Boeckh, M., Bloom, J.D. Parallel evolution of influenza across multiple spatiotemporal scales. *eLife* 6: e26875 (2017). DOI: 10.7554/eLife.26875

Viruses rapidly acquire *de novo* mutations as they replicate within infected hosts (Andino and Domingo, 2015), but only a small fraction of these variants transmit between hosts and eventually fix on a global scale. Within hosts, a mutation's impact on viral replication and immunogenicity affect whether it increases in frequency. At larger scales of space and time, transmission bottlenecks (Poon et al., 2016; Varble et al., 2014) and host heterogeneity also shape viral genetic diversity. The selective pressures at these various scales reflect complex molecular, immunological, and epidemiological constraints (Grenfell et al., 2004; Łuksza and Lässig, 2014; Neher et al., 2016; Pybus and Rambaut, 2009), which have formed the basis of recent efforts to forecast influenza evolution (Lässig et al., 2017; Łuksza and Lässig, 2014; Neher et al., 2014, 2016).

Influenza's rapid global evolution has been the subject of intense study (Ghedin et al., 2005; Rambaut et al., 2008b), but the origins of this variation within single infected hosts are still poorly understood. Recent deep-sequencing studies of human clinical samples suggest that influenza accumulates relatively limited genetic diversity within hosts during most acute infections (Debbink et al., 2017; Dinis et al., 2016; Poon

et al., 2016; Sobel Leonard et al., 2016), in line with earlier studies in dogs and horses (Hoelzer et al., 2010; Murcia et al., 2010). Some within-host mutations may confer novel antigenic properties (Dinis et al., 2016), but most lack clear functional interpretation. Altogether, it remains unclear how influenza's within-host diversity is transformed into global evolution.

Influenza infections usually last less than a week and provide limited opportunity for longitudinal study. But among some immunocompromised patients, infections can last weeks or months (Memoli et al., 2014; Nichols et al., 2004), making it possible to examine longer-term within-host evolutionary dynamics (McMinn et al., 1999; Rocha et al., 1991; Rogers et al., 2015). Here, we use deep-sequencing to characterize the evolutionary dynamics of influenza within immunocompromised hosts. We identify a small set of mutations that arise repeatedly within individual patients, across multiple patients in our study, and at the global scale, revealing surprising similarities in evolutionary dynamics across multiple spatiotemporal scales.

4.1 RESULTS

4.1.1 The same mutations often arise in multiple patients.

We deep-sequenced 37 viral samples collected longitudinally from four immunocompromised patients with long-term H3N2 influenza infections in the 2005-2006 and 2006-2007 seasons (**Figure 4.1)**. These patients developed influenza infections in the months after receiving hematopoietic cell transplantations when immune cell counts were still low, and nasal wash samples were collected approximately every week. All patients were treated with the neuraminidase inhibitor

oseltamivir for at least some duration of their infections (Campbell et al., 2015) (**Figure 4.1**, **Figure 4.2**).

We sequenced the full viral genome to high coverage directly from patient nasal wash samples by using influenza-specific reverse transcription and PCR (Hoffmann et al., 2001) to enrich for viral genetic material (**Figure 4.4**). To limit the impact of library preparation and sequencing errors on estimates of variant frequency (McCrone and Lauring, 2016), we prepared sequencing libraries in duplicate for each sample, beginning from separate reverse-transcription reactions. We excluded from downstream analyses eight low-quality samples for which sequencing coverage was low or variant frequencies differed greatly between replicates (**Figure 4.4**).

Across the influenza genome, *de novo* mutations arise most commonly in the surface proteins hemagglutinin (HA) and neuraminidase (NA) (**Figure 4.3A**), which undergo rapid global evolution (Bhatt et al., 2011). These mutations fluctuate in frequency but rarely fix, showing that complex evolutionary dynamics can emerge within single infected individuals (**Figure 4.3B, Figures 4.5-4.8**). We focused on within-host mutations that reached a frequency of at least 5% in two independent sequencing replicates from any patient sample. Many nonsynonymous mutations occur at sites that affect the antigenicity of HA (Koel et al., 2013) and the antiviral sensitivity of NA (Baz et al., 2006; van der Vries et al., 2013a) (**Figure 4.9**). In NA in particular, we observe the emergence and persistence of mutations T242I and R292K, which are known to be associated with oseltamivir resistance (Baz et al., 2006; van der Vries et al., 2013a), a phenomenon of strong clinical importance (Renaud et al., 2011) (**Figures 4.5-4.8**).

In several cases, the same mutations arise independently and reach high frequency in multiple patients (**Figure 4.3C**). We identified nine sites in the influenza genome where parallel mutations arose in two or more patients in our study: five in HA, three in NA, and one in the nonstructural (NS) segment (**Figure 4.3C, Figure 4.10**; HA: $p<0.001$; NA: $p<0.01$; permutation test). In subsequent analyses, we focused primarily on HA because of its prominent role in antigenic evolution (Koel et al., 2013).

4.1.2 Recurrent mutations drive clonal interference within individual patients.

Although the same HA mutations arise in multiple patients, we found that evolutionary outcomes sometimes diverge. For instance, A138S arises in patients W and Z, but it fixes only in patient Z. In three patients, N225D reaches a detectable frequency, but it fixes only in patient X (**Figure 4.3C**).

We suspected that the complex dynamics of these within-host mutations might arise from competition among mutant lineages. The influenza genome consists of eight linear segments that freely re-assort with one another but do not recombine (Boni et al., 2008), meaning that each segment evolves clonally. In the absence of homologous recombination, lineages carrying beneficial mutations rise and fall in frequency as they compete with one another, making it harder for any one variant to fix. This phenomenon, known as clonal interference, has been characterized extensively in experimental evolution (Hegreness et al., 2006; Kao and Sherlock, 2008; Lang et al., 2013; Neher, 2013) and affects influenza's global evolution (Strelkowa and Lässig, 2012).

We examined clonal dynamics within individual patients by analyzing patterns of linkage among within-host mutations. We identified read pairs that spanned multiple

variable sites to infer linkage, and we summarized these relationships as haplotypes: for instance, “0000” represents ancestral residues at four variable sites, and “1100” represents a double-mutant at the first two sites (**Figure 4.11A**).

In several instances, the same mutations arise in parallel on distinct genetic backgrounds within the same patient—echoing our observation that these same mutations arise in parallel in multiple patients in our study. In patient X, lineages carrying S193Y and N225D initially compete, but a double-mutant carrying both mutations eventually fixes (**Figure 4.11B**). The A138S and F193Y mutations also arise multiple times in parallel in patient W: once on the ancestral haplotype “0000” to form the single-mutant “1000” and “0100” lineages; once on these single-mutant lineages to form the double-mutant “1100”; and once on the double-mutant “0011” to form the triple-mutant “1011” and “0111” lineages (**Figure 4.11C**). These recurrent mutations also contribute to the large number of clonal lineages present. Several weeks into patient W’s infection, we observe at least five distinct HA lineages at a frequency of at least 5%, and the lineages differ from each other by one to three nonsynonymous mutations (**Figure 4.11, Figure 4.3C**). Eventually, all lineages that carry A138S, V223I, and N225D are outcompeted by a lineage that carries F193Y.

Our analysis shows that in large, clonally evolving influenza populations within hosts, a small set of beneficial mutations repeatedly arise and compete against one another in various combinations. Although many of these beneficial mutations are selected in parallel in multiple patients, the unpredictability of clonal competition determines which mutations eventually fix.

4.1.3 Within-host variants often arise at sites that are polymorphic in influenza globally.

We compared viral mutations that arose within our patients and at the global scale. Strikingly, many of the HA mutations that arise in parallel in multiple patients in our study also reach a high global frequency, which may reflect concordant antigenic selection at the within-host and global scales. We identified all variants that reached a frequency of at least 10% in any given year after 2000 in the GISAID database of global influenza sequences (Bogner et al., 2006) and compared them to variants that we identified in the patients in our study.

In HA, most sites that varied within hosts also varied in the global influenza population, compared to about a quarter of such sites in the other influenza genes (**Figure 4.15**). We tested whether this overlap between sites of variation within patients and globally was greater than expected by chance for HA, NA, and the rest of the viral genome combined. We calculated the expected overlap when the observed number of within-host and global variants were drawn at random from each gene (**Figures 4.16, 4.17**). Not all sites are expected to tolerate mutation, so we also performed simulations where we only considered sites for which there was variation in human H3N2 influenza globally between 2000 and 2015: for instance, in HA about 25% of codon sites show no variation within the GISAID database. We found significant parallelism in HA ($p<0.01$), but not in NA or in the rest of the genome ($p>0.05$) when we consider all sites of global variation. This parallelism in HA evolution remains statistically significant at a 0.05 threshold until we assume that less than 50% of HA codon sites tolerate variation.

The parallelism is especially striking at the sites of HA mutations found in multiple patients in our study. In particular, four of the five sites of recurrent within-host mutation

in HA are also sites of global influenza variation (**Figure 4.15, Table 4.2**). The V223I and N225D mutations arise in multiple patients, and then fix globally in the decade after the patient infections (**Figure 4.15D**). Mutations also reach high global frequencies at sites 138 and 193, although the F193 and S193 variants that spread globally differ from the Y193 variant that arises within our patients. However, the concordance is incomplete. Mutation L427F reaches a frequency of >75% in three patients but is rare or nonexistent in influenza globally (**Figure 4.15D**), suggesting that this mutation may have within-host benefits that are not reflected in global evolution. But overall across hemagglutinin, within-host variants tend to arise at sites that vary on the global scale.

4.2 Discussion

It is remarkable that influenza evolution shows such extensive parallelism at these disparate spatiotemporal scales despite heterogeneity in host immunity, viral genetic background, and the severity and duration of infection. In particular, the immunocompromised patients in our study had complex underlying conditions and diverse immune histories. Notably, the four HA sites that displayed parallel within-host and global evolution in our study (138, 193, 223, and 225) also gave rise to mutations in another study that used Sanger sequencing to analyze laboratory-passaged influenza isolated longitudinally from an immunocompromised child (Baz et al., 2006). Another previous study used hemagglutination inhibition assays to show that antigenic drift of influenza within an immunocompromised patient resembled global antigenic change (McMinn et al., 1999). These similarities further support our finding that influenza evolution shows parallelism across diverse patients. The parallel evolution that we

observe in influenza at the within-host and global scales contrasts with HIV, where similar mutations can arise within hosts that share an HLA type, but tend to revert upon transmission to recipients with different HLA types (Herbeck et al., 2006; Lemey et al., 2006; Leslie et al., 2004; Zanini et al., 2015). Part of the difference may be that immune epitopes in influenza are broadly similar among individuals with some exceptions (Li et al., 2013; Linderman et al., 2014), whereas the targets of anti-HIV immunity vary more widely due to patient-specific factors such as HLA type.

We suggest that parallelism in HA evolution may emerge from the confluence of several evolutionary conditions (Lässig et al., 2017). First, if selection acts concordantly across environments, it will favor a common set of beneficial mutations. Second, in a constrained evolutionary landscape, this set of beneficial mutations will be relatively small. Finally, given sufficiently large population sizes, high mutation rates, and time, these beneficial mutations will emerge and be selected to detectable frequencies. Our observation that similar mutations arise repeatedly within single patients, within multiple different patients, and at the global scale, suggests that at least some of these conditions may hold true.

The parallelism and extensive evolution that we observe in long-term influenza infections contrasts with the limited within-host variation found in prior studies, which sample from acute infections of immunocompetent hosts (Debbink et al., 2017; Dinis et al., 2016; Hoelzer et al., 2010; Murcia et al., 2010; Poon et al., 2016; Sobel Leonard et al., 2016). For instance, one recent study deep-sequenced HA from several hundred patients but only found a small number of antigenic variants, and mostly at low frequencies (Dinis et al., 2016). But our study suggests that influenza may experience

many of the same selective pressures within acute infections as it does globally, even if the short durations of these infections make it difficult for selected mutations to reach frequencies that are detectable with current methods. We suggest that within-host viral diversity may act as a noisy early measurement of global viral evolution, shaped by some of the same immunological and evolutionary constraints. As high-throughput sequencing continues to improve, detailed characterization of within-host variation will be increasingly valuable for understanding how molecular, immunological, and epidemiological forces interact to shape viral evolution.

4.3 MATERIALS AND METHODS

4.3.1 Patient material

Samples were prospectively collected during a surveillance study for respiratory viruses performed in allogeneic hematopoietic stem cell transplant (HCT) recipients undergoing transplantation between December 2005 and February 2010 at Fred Hutchinson Cancer Research Center (Campbell et al., 2015). Following written informed consent, weekly nasal wash samples (or nasopharyngeal swabs if nasal wash samples were precluded clinically) and oropharyngeal swab specimens were obtained at least once before and weekly after HCT up to 100 days. Afterwards, samples were collected as long as the patients continued to test positive for respiratory viruses, if they developed new symptoms, or at least every three months until one year post-transplantation. Nasal wash samples were collected using 5 mL of saline per nostril, and combined with oropharyngeal swabs for real-time PCR testing for a panel of 12 respiratory viruses, including influenza A and B. Samples were considered positive if the assay's cycle threshold was less than 40, for a limit of detection of approximately 2000 viral copies/mL. All samples sequenced in this study tested positive for influenza A. The timing of each sample during an infection was calculated as the number of days since the first influenza-positive nasal wash for that patient.

Descriptions of individual patients and their clinical courses are summarized below, with detailed information in **Figure 4.2**. All patients were severely immunocompromised: although their influenza infections occurred after transplant engraftment, their lymphocyte counts remained well below those found in

immunocompetent individuals, and they were concurrently treated with immunosuppressive medications. Influenza sometimes co-occurred with other respiratory viruses, and the patients were frequently taking multiple antiviral and antibiotic medications at any given point in the infection.

Patient W. A female in the 25-44 age group developed upper respiratory symptoms in early 2007, 30 days after receiving a non-myeloablative HCT for Hodgkin's disease and 18 days following engraftment. Patient nasal wash samples repeatedly tested positive for influenza A for the next 80 days until the patient died of pulmonary failure, with diffuse alveolar damage found on autopsy. The patient received a 12-day course of oseltamivir at 75 mg PO BID approximately 30 days into the infection and was treated continuously with oseltamivir for the last 26 days of her life, first at 75 mg PO BID and then increasing to 150 mg PO BID. The patient was co-infected with coronavirus for the duration of the influenza infection and also tested positive for human metapneumovirus for the last 26 days of her life.

Patient X. A male in the 65+ age group developed upper respiratory symptoms in early 2006, 45 days after receiving a non-myeloablative HCT for Hodgkin's disease. Patient nasal wash samples repeatedly tested positive for influenza A for the next 72 days, after which the patient chose to discontinue study participation. The patient was treated with two courses of oseltamivir: a 5-day course at 75 mg PO BID following the first positive nasal wash, and an 8-day course at 75 mg PO BID approximately four weeks into the infection. The patient also tested positive for cytomegalovirus (CMV) and *Aspergillus* early in the influenza infection.

Patient Y. A male in the 45-64 age group developed upper respiratory symptoms in spring 2006, 62 days after receiving a non-myeloablative HCT for acute myeloid leukemia (AML) and 52 days after engraftment. Patient nasal wash samples repeatedly tested positive for influenza A for the next 77 days, after which the patient began testing negative. The patient was treated with three courses of oseltamivir: an 8-day course following the first positive nasal wash, a 30-day course beginning approximately two weeks into the infection, and a second 30-day course starting approximately seven weeks into the infection, all at 75 mg PO BID. The patient also intermittently tested positive for CMV and coronavirus during the influenza infection.

Patient Z. A male in the 65+ age group developed upper respiratory symptoms in early 2007, 197 days after receiving a non-myeloablative HCT for AML and 175 days after engraftment. Nasal wash samples repeatedly tested positive for influenza A over the next 69 days, after which monitoring ceased due to severe illness, and the patient died 15 days after the last influenza-positive sample from relapsed AML. The patient was treated with two courses of oseltamivir: a 6-day course at 150 mg PO BID following the first flu-positive nasal wash, and a 66-day course starting approximately two weeks into the infection that began at 150 mg PO QD and increased to 150 mg PO BID. The patient also received 30g of IVIG 46 days into the flu infection. The patient intermittently tested positive for respiratory syncytial virus over the same period and also experienced Epstein-Barr viremia.

4.3.2 Viral deep sequencing

To deep-sequence viral populations, we extracted bulk RNA from nasal wash samples using the QIAamp Viral RNA Mini Kit (QIAGEN) according to manufacturer's

instructions. Where possible, we extracted RNA from 560uL of sample, the maximum volume recommended for use with the QIAamp kit, to capture as much viral diversity as possible.

To amplify the influenza genome, we modified the primers designed by Hoffmann et al. (Hoffmann et al., 2001) for full-length amplification of the influenza A genome (**Table 4.1**). We performed reverse transcription using Superscript III First-Strand Reaction Mix (Thermo Fisher) and an equimolar mix of the 5'-Hoffmann-U12-A4 and 5'-Hoffmann-U12-G4 primers, which bind to the conserved U12 region present on each influenza gene segment. To 6uL RNA eluent, we added 1uL annealing buffer and 1uL of 2uM primer mix, then incubated at 65 degrees C for 5 minutes. We added 10uL 2X First-Strand Reaction Mix and 2uL Superscript III/RNaseOUT Enzyme Mix on ice for a 20uL total reaction volume, then incubated at 25 degrees C for 10 minutes (this initial incubation is designed to help with the binding of short primers), 50 degrees C for 50 minutes, and 85 degrees C for 5 minutes.

We used the entire 20uL volume of the reverse-transcription reaction as template in a 100uL PCR reaction using KOD HotStart Reaction Mix (EMP Millipore) and a 24-primer cocktail as described in **Table 4.1** at a total concentration of 600nM. We performed 35 cycles of PCR amplification with an annealing temperature of 55 degrees C and an extension time of 3 minutes.

We purified the PCR product using 1X AMPure beads (Beckman Coulter) and prepared libraries for Illumina sequencing using Nextera XT (Illumina). We sequenced the libraries on a NextSeq 500 platform (Illumina) with 150 bp paired-end reads. We

performed library preparation and sequencing in duplicate, starting from independent reverse-transcription reactions.

4.3.3 Read mapping

We first used bowtie2 (Langmead and Salzberg, 2012) to filter out reads that mapped to the human genome. Remaining reads are available in the SRA as BioProject PRJNA364676. We used cutadapt 1.8.3 (Martin, 2011) to trim adapter sequences from the remaining reads, remove bases at the ends of reads with a Q-score below 25, and filter out reads whose remaining length was shorter than 20 bases. We locally aligned trimmed reads to the A/Brisbane/10/2007 (H3N2) genome (Genbank accessions CY035022 to CY035029) using bowtie2 (Langmead and Salzberg, 2012) and tallied the counts of each base at each genome position using custom scripts. We discarded reads with a mapping score below 20, as well as bases with a Q-score below 20.

4.3.4 Quality filtering

We calculated average sequencing coverage in 50-bp bins along the viral genome. Because we prepared sequencing libraries using Nextera tagmentation, we expect coverage to be low at the two ends of the eight viral gene segments, corresponding to 16 bins. We discarded samples with more than 16 bins with average coverage below 200x (**Figure 4.4A**). We also identified sites at which a non-consensus base reached a frequency of at least 1% in both sequencing replicates and compared variant frequencies between replicates. We discarded samples for which the average difference between variant frequencies in the two replicates exceeded 0.05 (**Figure 4.4B**). In total,

we excluded eight samples from downstream analyses. The samples shown in **Figure 4.1B** are high-quality samples only.

4.3.5 Variant calling and annotation

For each patient, we identified variable nucleotide sites in the viral genome. We defined these sites as positions with a sequencing coverage of at least 200x, at which multiple bases are present at a frequency of at least 5% in both replicate libraries. We used custom scripts to determine each variant's codon position and whether it created a synonymous or nonsynonymous substitution.

4.3.6 A note on codon numbering and gene annotation

We numbered HA codons according to the H3 numbering system. This HA numbering scheme assigns 1 to codon 17 of the full HA gene, which is the beginning of the mature HA protein. The codons for all other genes are numbered sequentially beginning with 1 at the N-terminal methionine. The M1 and M2 genes have 27 bp of in-frame and 44 bp of out-of-frame overlap, and the NS1 and NEP genes have 30 bp of in-frame and 251 bp of out-of-frame overlap. We annotated variants separately for each gene if they occurred in these regions of overlap.

4.3.7 Phylogenetic analysis

For each patient in our study, we determined the viral consensus sequence at the first sequenced time point. We also downloaded the set of 503 sequences in the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database (Bogner et al., 2006) corresponding to all full-length HA coding regions from human H3N2 influenza A isolates collected in the USA from January 1, 2004 to December 31, 2007 (GISAID

acknowledgement tables provided in **Supplementary File 1**). We analyzed only sequences with passage annotation “Unpassaged,” “Original”, or “P0,” indicating that the strains were sequenced directly from the clinical isolates, leaving 63 unique sequences for phylogenetic inference. We pairwise aligned each sequence to the A/Brisbane/10/2007 (H3N2) coding sequence (Genbank accession CY035022) using the program needle from EMBOSS 6.6.0 (Rice et al., 2000), which implements a Needleman-Wunsch alignment. We used RAxML 8.2.3 (Stamatakis, 2014) to infer a phylogeny from this alignment using a GTRCAT codon-substitution model and visualized the tree using the R package ggtree (Yu et al., 2017).

4.3.8 Haplotype inference

We identified paired-end reads that spanned *n* variable sites of interest within a single gene and determined which bases were present at each variable site. We summarized this information as an *n*-digit binary haplotype, in which each digit represented one variable site, 0 represented the ancestral base, and 1 represented the derived base. We discarded reads that did not span all sites of interest, or that contained genotypes other than the most common derived base. We estimated the rate of PCR recombination as described in **Figure 4.12**. In **Figure 4.13** and **Figure 4.14**, we show the number of paired-end reads used to infer the haplotypes in **Figure 4.11**.

4.3.9 Analysis of global variation

To identify sites of global variation in influenza, we downloaded all sequences in the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database (Bogner et al., 2006) corresponding to all full-length influenza coding regions from human H3N2

influenza A isolates collected from January 1, 2000 to December 31, 2015. Acknowledgement tables are provided as **Supplementary File 1**. We pairwise aligned each sequence to the A/Brisbane/10/2007 (H3N2) coding sequence (Genbank accession CY035022) using the program needle from EMBOSS 6.6.0 (Rice et al., 2000), which implements a Needleman-Wunsch alignment. We calculated the amino-acid distance of each sequence from the Brisbane/2007 reference and excluded outliers whose distance deviated significantly from the other sequences originating from that year, since these sequences may have been misannotated. We tallied the amino acids present at each codon position in each year, discarding sequences that contained indels, and we identified sites at which multiple amino acids were present at a frequency of at least 10% within a single year, or at which the consensus base changed from year to year.

4.3.10 Statistical tests of parallelism

We sought to test the probability that the parallel emergence of mutations across multiple patients in our study was due to chance. We began with a simple null model in which all sites are equally likely to mutate, and we drew sites from each gene at random without replacement, matching the number of mutations observed in each patient. We calculated the number of unique sites of mutation among all four patients in this simulated data set, and we compared this distribution to the number of unique sites observed in our sequencing data: fewer unique sites of mutation indicates more parallelism (**Figure 4.10A**). This null model is overly simplistic, since some sites in a protein experience more evolutionary constraint. To estimate this constraint, we limited the number of sites considered mutable to the sites that show at least two instances of

nonsynonymous mutation in the global H3N2 population between 2000 and 2015 (see Analysis of Global Variation) (**Figure 4.10B**). The p-values given in the main text are calculated under this more conservative null model. We also performed permutation tests for a range of possible proportions of mutable sites and calculated the fraction of simulations that matched or exceeded the amount of parallelism observed in our data (**Figure 4.10C**).

We used a similar approach to test whether the overlap of mutations observed within patients in our study and in the global flu population was likely to be due to chance. We drew two independent sets of sites from each gene at random without replacement, matching the total number of unique variable sites within all patients and the number of variable sites observed in the global population. We then calculated the overlap between these two sets of sites. We used the approach above to calculate the overlap under a simple null model in which all sites in the gene are equally like to mutate; a constrained null model in which the only mutable sites are ones that show nonsynonymous mutation between 2000 and 2015; and across a range of possible constraints (**Figure 4.17**).

4.3.11 Data and code availability

The FASTQ files are available on the SRA as BioProject PRJNA364676. The computer code that performs the analysis is available at https://github.com/ksxue/parallel-evolution and in **Supplementary File 2** (Xue 2017).

4.3.12 Acknowledgements

We thank Choli Lee and Seungsoo Kim for assistance with sequencing; Darneshia Smith, Louise Kimball, and Alpana Waghmare for interpretation of patient clinical data; and Seungsoo Kim, Alexander Greninger, and Trevor Bedford for comments and discussion about the manuscript. We also thank Thomas Friedrich, Louise Moncla, and Nick Florek for helpful discussions about methods for deep-sequencing and Mike Famulare for helpful discussions about evolution at the within- and between-host scales. MB reports research support and consulting fees from Aviragen Therapeutics, Gilead Sciences, and Ansun BioPharma, and research support from GlaxoSmithKline. JAE reports research support from Gilead Sciences, GlaxoSmithKline, Chimerix, and Pfizer, and fees for participation in a Data Safety Monitoring Board for GlaxoSmithKline. The other authors declare no competing financial interests.

4.4 FIGURES AND TABLES

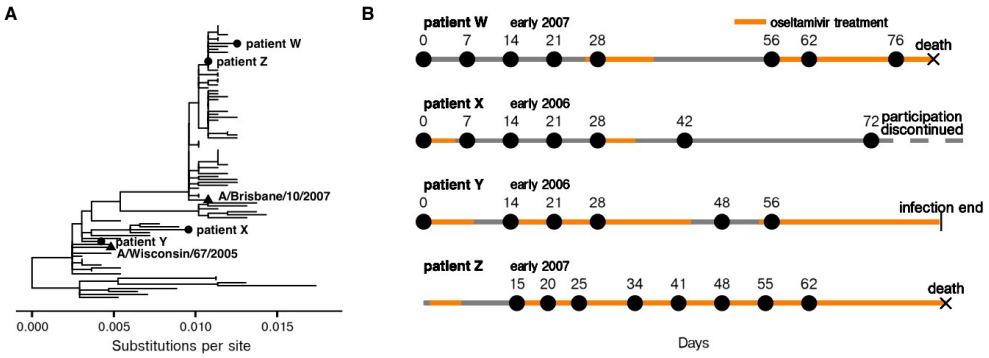

Figure 4.1. Long-term H3N2 influenza infections in four immunocompromised patients.

(A) Phylogenetic relationship between initial patient consensus sequences and 63 unique circulating influenza strains collected in the USA from 2004 to 2007, as inferred from the HA gene. **(B)** Overview of patient influenza infections and treatments. Periods of oseltamivir treatment are shown in orange. Dates of sequenced nasal wash samples are calculated relative to the first influenza-positive nasal wash. Low-quality samples are not shown here and were excluded from downstream analysis. **Materials and Methods** and **Figure 4.2** give full clinical histories.

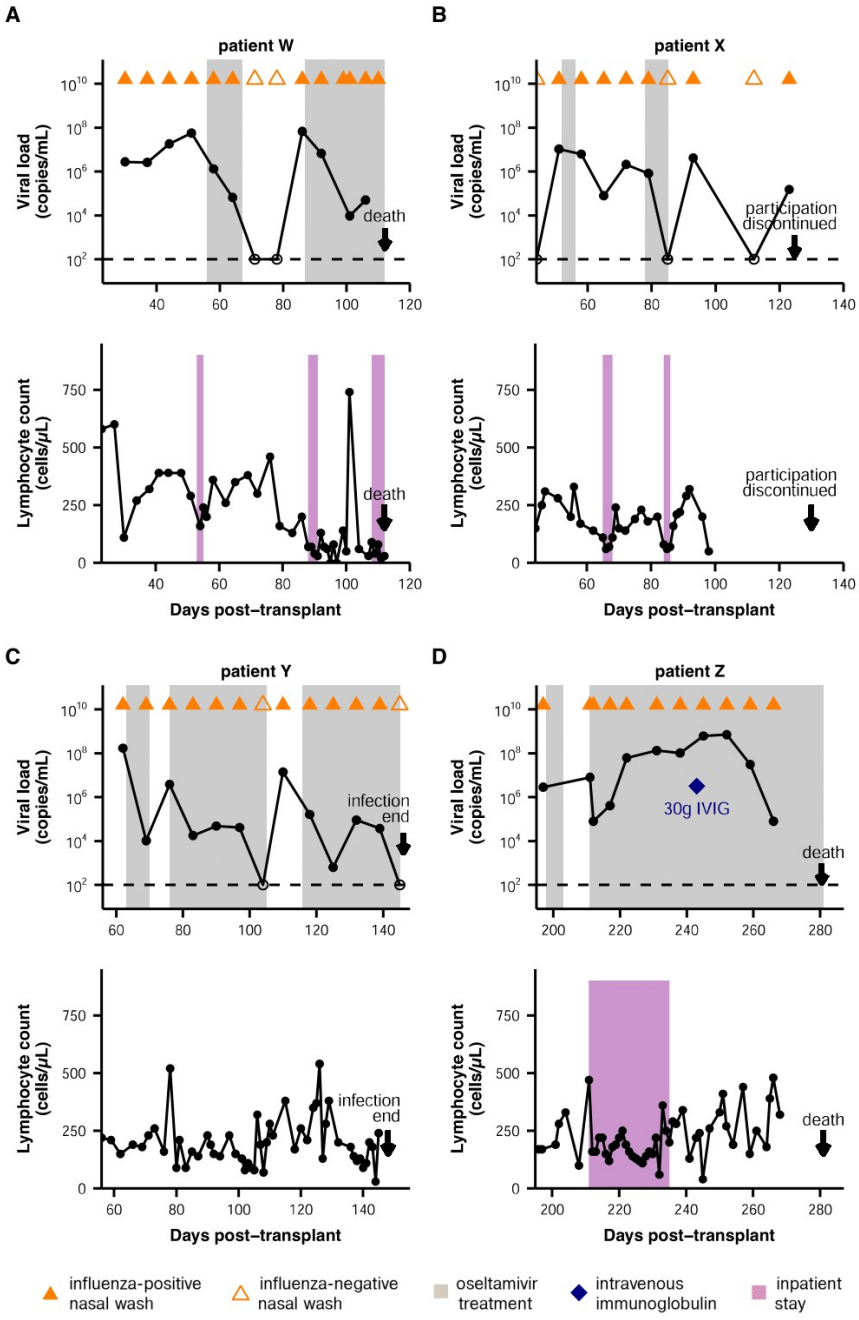

Figure 4.2. Summary of patient infections.

Influenza viral load over time, as quantified by qRT-PCR, and lymphocyte counts over time are shown for all patients. For the plots of viral load, nasal wash samples are marked with triangles, with influenza-positive nasal washes indicated by solid coloring. Oseltamivir treatment is indicated by gray shading. Intravenous immunoglobulin treatment is marked with a blue diamond. The limit of detection is marked with a dashed line. For the plots of lymphocyte counts, inpatient stays are indicated in purple.

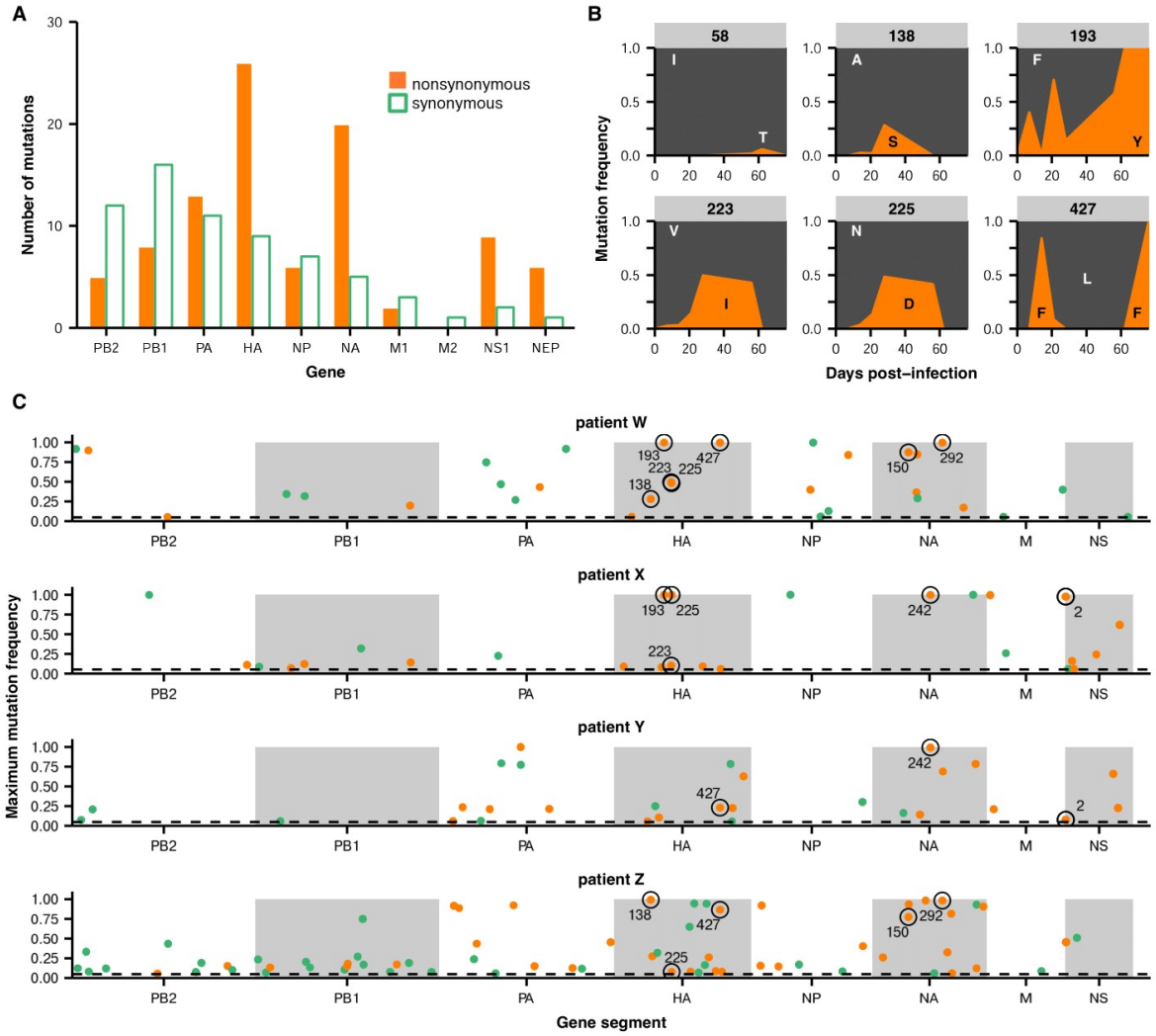

Figure 4.3. Within-host influenza variants.

(A) Number of nonsynonymous (orange) and synonymous (green) variants in each influenza gene. We identified within-host viral mutations that reached a frequency of at least 5% in two independent sequencing replicates from any patient sample. **(B)** Frequencies over time for all HA mutations in patient W. Each subplot represents a site in HA and is labeled by codon number. Ancestral identities are colored in gray and mutant ones in orange. **(C)** Maximum frequencies reached by all nonsynonymous (orange) and synonymous (green) mutations in each patient. Mutations circled in black emerged independently in multiple patients and are labeled by codon number. The dotted line indicates the minimum frequency threshold of 5%. **Materials and Methods** and **Figure 4.4** describe procedures used for variant calling and quality control. **Figures 4.5-4.8** give full frequency trajectories for all mutations in all patients. **Figure 4.9** shows mutations in HA and NA on their respective crystal structures. **Figure 4.10** describes permutation tests that assess the significance of the observed parallelism between patients.

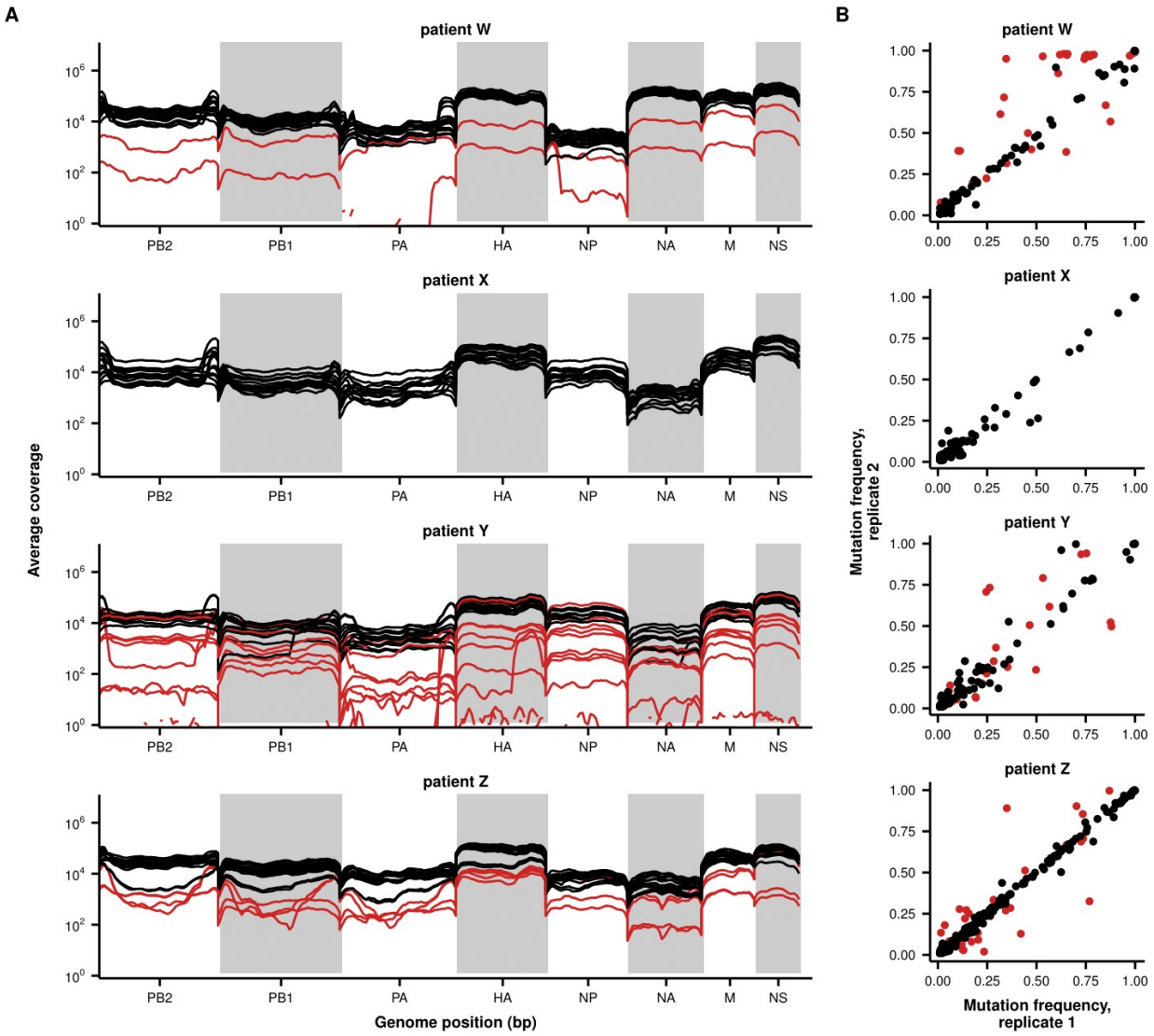

Figure 4.4. Sample quality controls.

(A) Sequencing coverage across the genome. Average sequencing coverage for each sample and library replicate is shown in 50-bp bins across the genome. Each line indicates a separate library replicate. Low-quality samples are colored in red and were not included in downstream analyses. **(B)** Correlation between mutation frequencies in replicate sequencing libraries. Mutations in low-quality samples are colored in red and were excluded from downstream analyses. These samples were not shown in **Figure 4.1B**. A mutation was called if a base other than the initial consensus base reached a frequency of at least 1% with a total coverage of at least 200 reads in both replicates. Note that this variant calling threshold is more lenient than the 5% threshold used to call variants in downstream analyses; this more lenient threshold identifies more variants to get a better measurement of replicability. Samples were excluded from downstream analyses if the average difference between mutation frequencies exceeded 0.05. Replicate libraries were prepared beginning with independent reverse-transcription reactions.

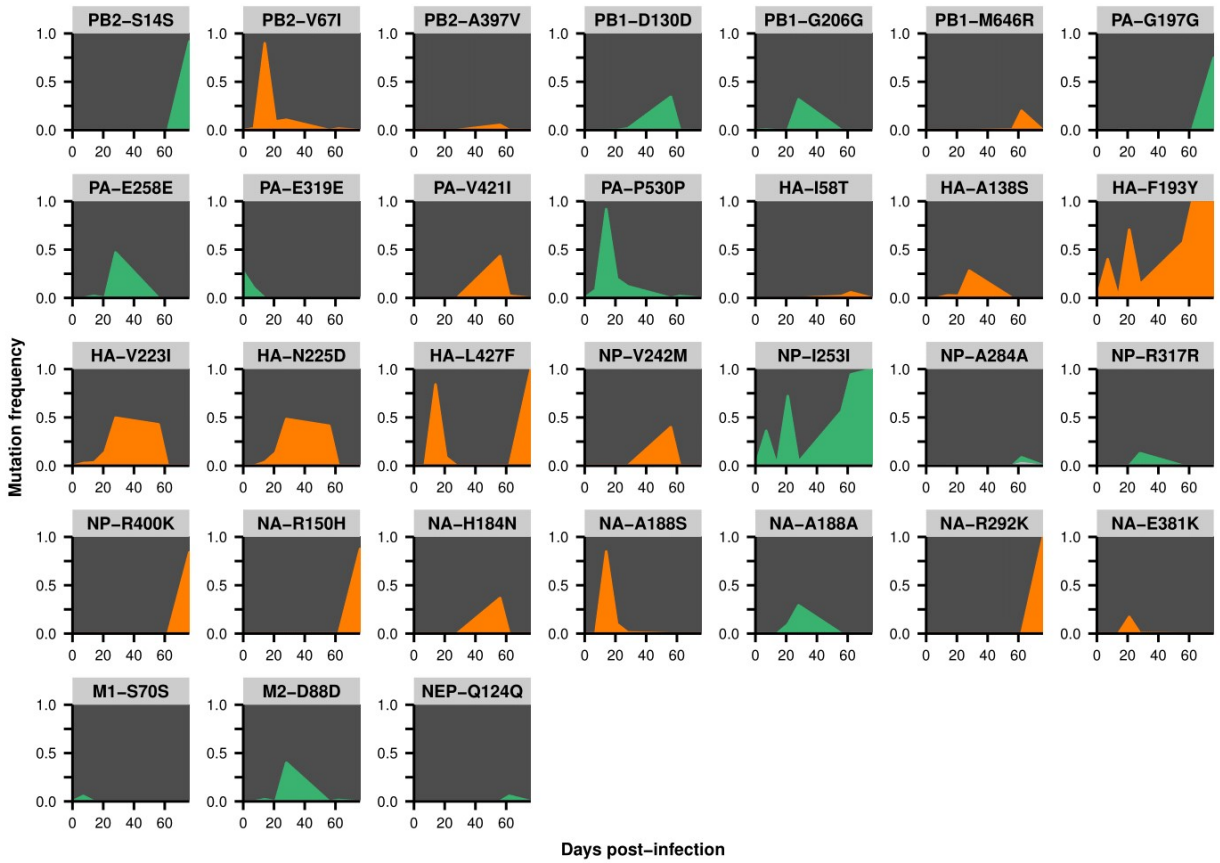

Figure 4.5. Within-host variants in patient W.

Mutation frequencies over the course of the infection are shown at all variable sites in patient W. Ancestral identities are colored in gray, derived nonsynonymous identities in orange, and derived synonymous identities in green. Sites were called as variant if a base other than the initial consensus reached a frequency of at least 5% in both replicate libraries of at least one sequenced time point. The frequencies shown here are the average of the two replicates. Mutations located in the overlap of M1/M2 or NS1/NEP genes are displayed twice with the appropriate annotation for each gene.

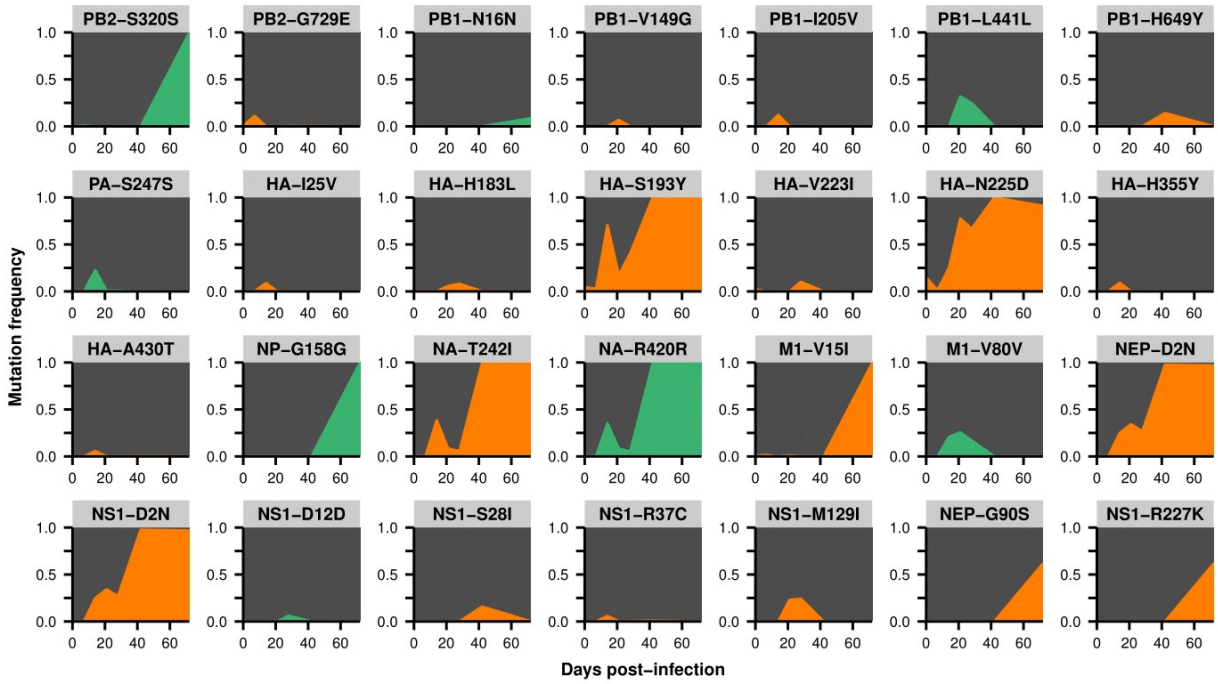

Figure 4.6. Within-host variants in patient X.

Mutation frequencies over the course of the infection are shown at all variable sites in patient X. Ancestral identities are colored in gray, derived nonsynonymous identities in orange, and derived synonymous identities in green. Sites were called as variant if a base other than the initial consensus reached a frequency of at least 5% in both replicate libraries of at least one sequenced time point. The frequencies shown here are the average of the two replicates. Mutations located in the overlap of M1/M2 or NS1/NEP genes are displayed twice with the appropriate annotation for each gene.

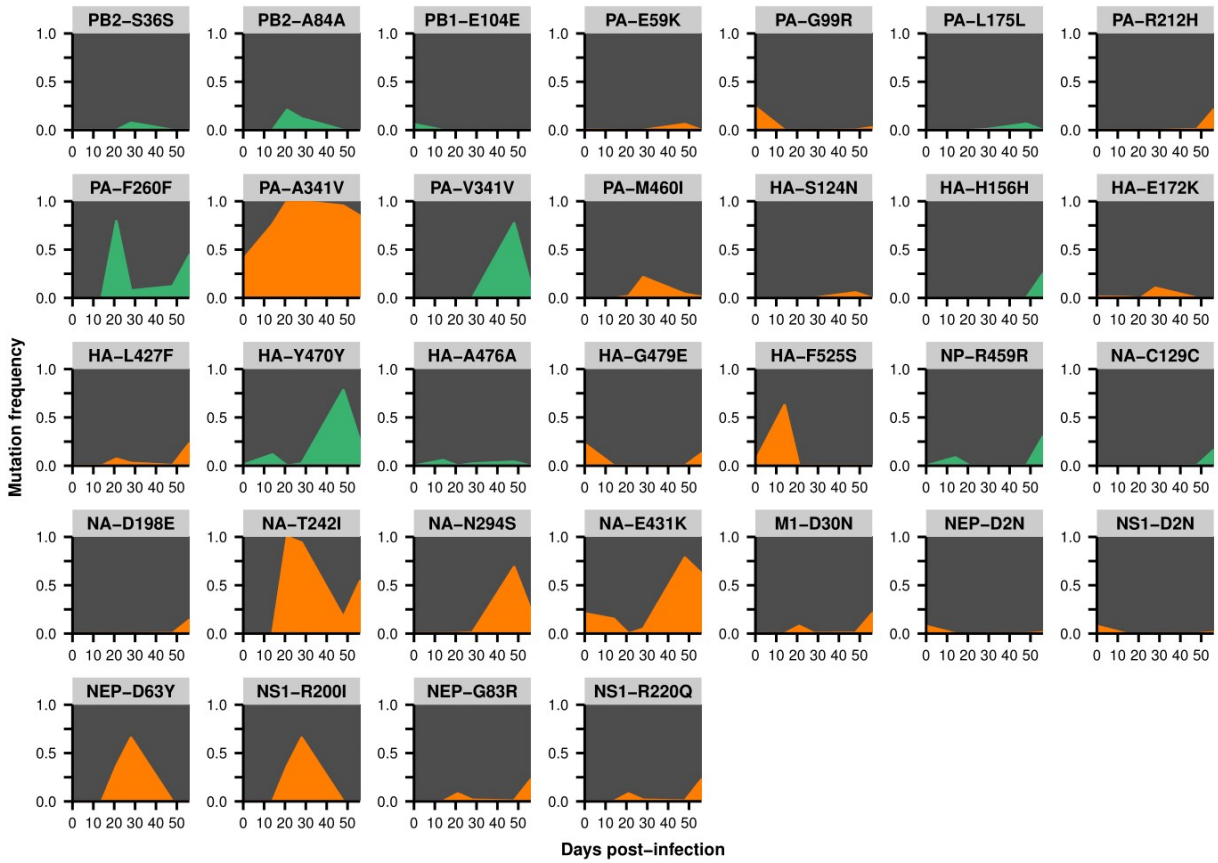

Figure 4.7. Within-host variants in patient Y.

Mutation frequencies over the course of the infection are shown at all variable sites in patient Y. Ancestral identities are colored in gray, derived nonsynonymous identities in orange, and derived synonymous identities in green. Sites were called as variant if a base other than the initial consensus reached a frequency of at least 5% in both replicate libraries of at least one sequenced time point. The frequencies shown here are the average of the two replicates. Mutations located in the overlap of M1/M2 or NS1/NEP genes are displayed twice with the appropriate annotation for each gene.

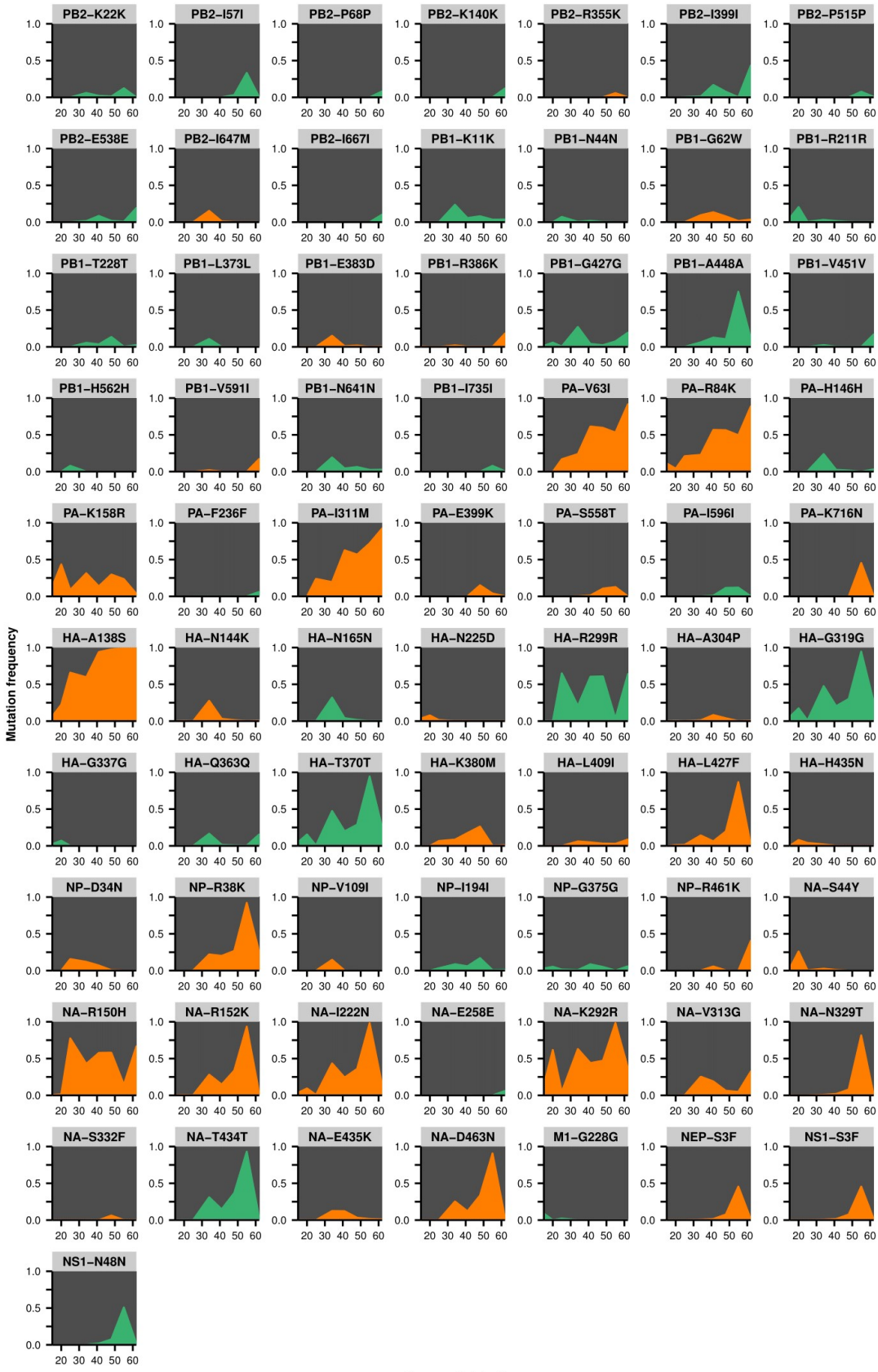
PB2–K22K
PB2–I57I
PB2–P68P
PB2–K140K
PB2–R355K
PB2–I399I
PB2–P515P
PB2–E538E
PB2–I647M
PB2–I667I
PB1–K11K
PB1–N44N
PB1–G62W
PB1–R211R
PB1–T228T
PB1–L373L
PB1–E383D
PB1–R386K
PB1–G427G
PB1–A448A
PB1–V451V
PB1–H562H
PB1–V591I
PB1–N641N
PB1–I735I
PA–V63I
PA–R84K
PA–H146H
PA–K158R
PA–F236F
PA–I311M
PA–E399K
PA–S558T
PA–I596I
PA–K716N
HA–A138S
HA–N144K
HA–N165N
HA–N225D
HA–R299R
HA–A304P
HA–G319G
HA–G337G
HA–Q363Q
HA–T370T
HA–K380M
HA–L409I
HA–L427F
HA–H435N
NP–D34N
NP–R38K
NP–V109I
NP–I194I
NP–G375G
NP–R461K
NA–S44Y
NA–R150H
NA–R152K
NA–I222N
NA–E258E
NA–K292R
NA–V313G
NA–N329T
NA–S332F
NA–T434T
NA–E435K
NA–D463N
M1–G228G
NEP–S3F
NS1–S3F
NS1–N48N
Mutation frequency
Days post–infection

Figure 4.8. Within-host variants in patient Z.

Mutation frequencies over the course of the infection are shown at all variable sites in patient Z. Ancestral identities are colored in gray, derived nonsynonymous identities in orange, and derived synonymous identities in green. Sites were called as variant if a base other than the initial consensus reached a frequency of at least 5% in both replicate libraries of at least one sequenced time point. The frequencies shown here are the average of the two replicates. Mutations located in the overlap of M1/M2 or NS1/NEP genes are displayed twice with the appropriate annotation for each gene.

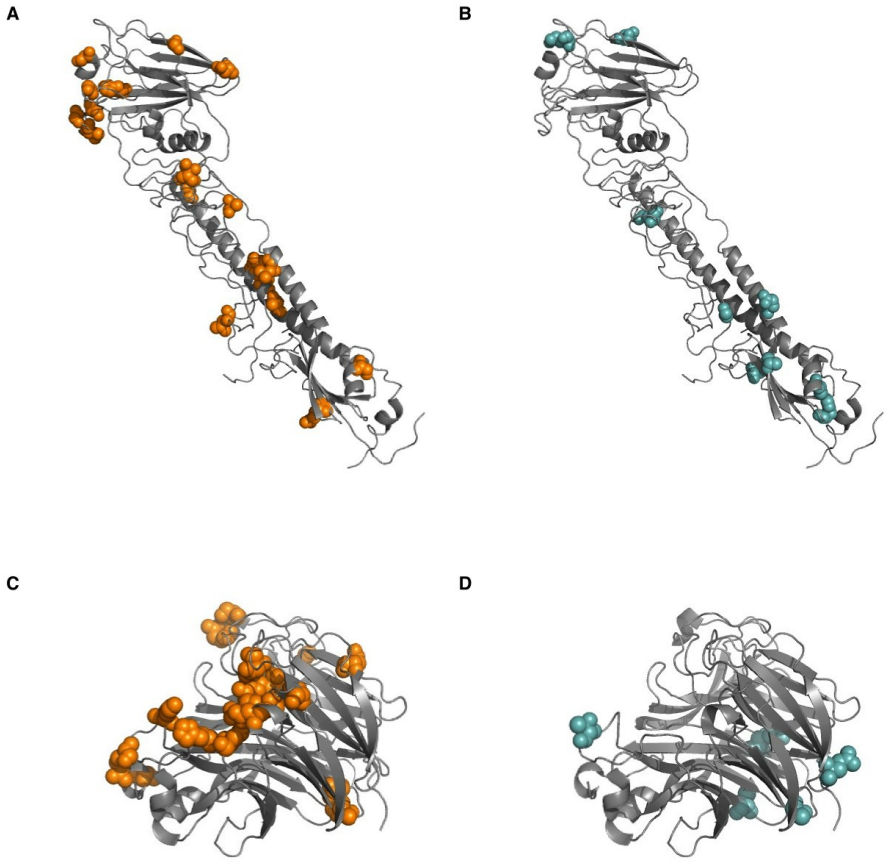

Figure 4.9. Sites of within-host mutation.

Sites of **(A)** nonsynonymous and **(B)** synonymous within-host mutations are shown on an HA crystal structure (PDB 4HMG (Weis et al., 1990)). Sites of **(C)** nonsynonymous and **(D)** synonymous within-host mutation are shown on an NA crystal structure (PDB 2BAT (Varghese et al., 1992)). All sites of synonymous and nonsynonymous mutation are shown here, in contrast to **Figure 4.15A**, which only shows sites in HA at which nonsynonymous mutations arise in multiple patients in our study.

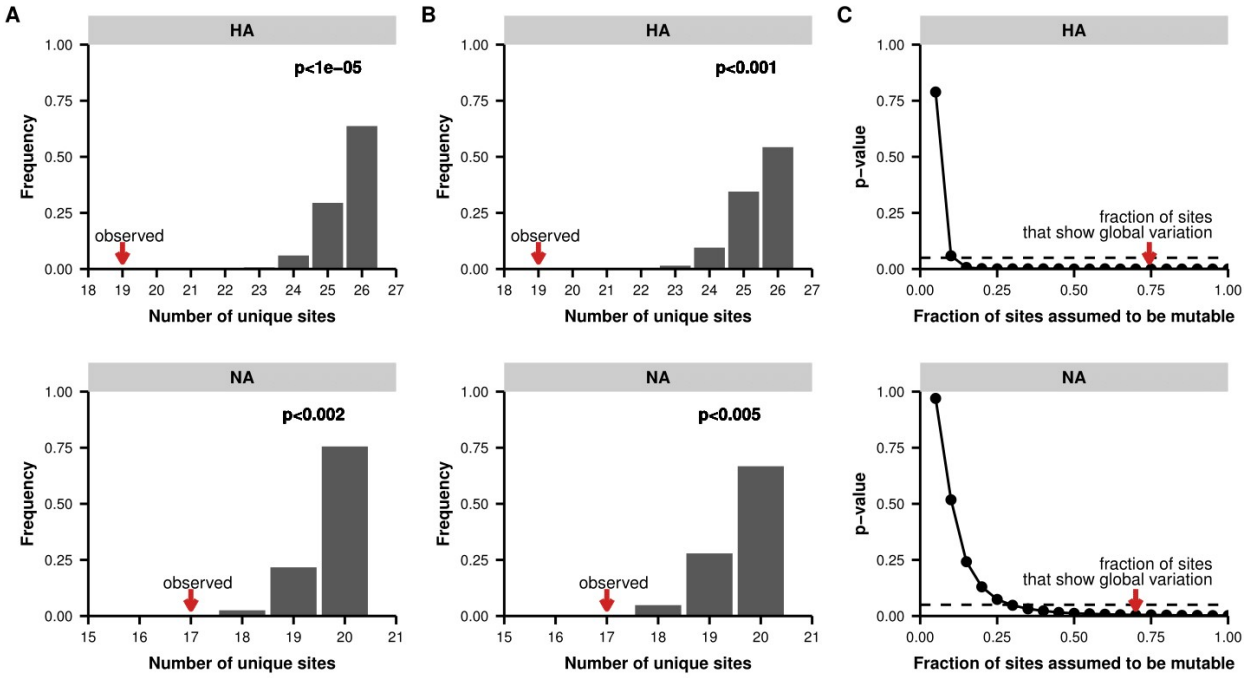

Figure 4.10. Permutation tests for parallel evolution between patients.

(A) Distribution of unique variable sites when sites are drawn at random along the full length of the indicated gene, matching the number of variants empirically observed in each patient. These simulations test a simple null model in which each site in the gene is equally likely to mutate. We calculated the number of unique sites of mutation in each simulation as a metric of parallelism: fewer unique sites means that more parallel mutation has occurred. The p-value indicates the proportion of 100,000 simulations in which the number of unique sites is less than or equal to what is empirically observed. **(B)** Distribution of overlapping sites for simulations as described in **(A)**, but with a constrained null model in which the fraction of sites considered mutable is the fraction that shows at least two instances of nonsynonymous mutation in the global H3N2 influenza population between 2000 and 2015 (see **Materials and Methods**) to account for constraints on protein evolution. For both HA and NA, the observed overlap of mutations between patients is statistically significant under this set of constraints. **(C)** p-values as described in **(A)**, calculated across a range of constraints on the fraction of mutable sites. The constrained null model indicated in **(B)** is indicated with a red arrow. For both HA and NA, the observed parallelism is statistically significant at a threshold of 0.05 unless it is assumed that fewer than 15% of the sites in HA or 35% of the sites in NA are mutable.

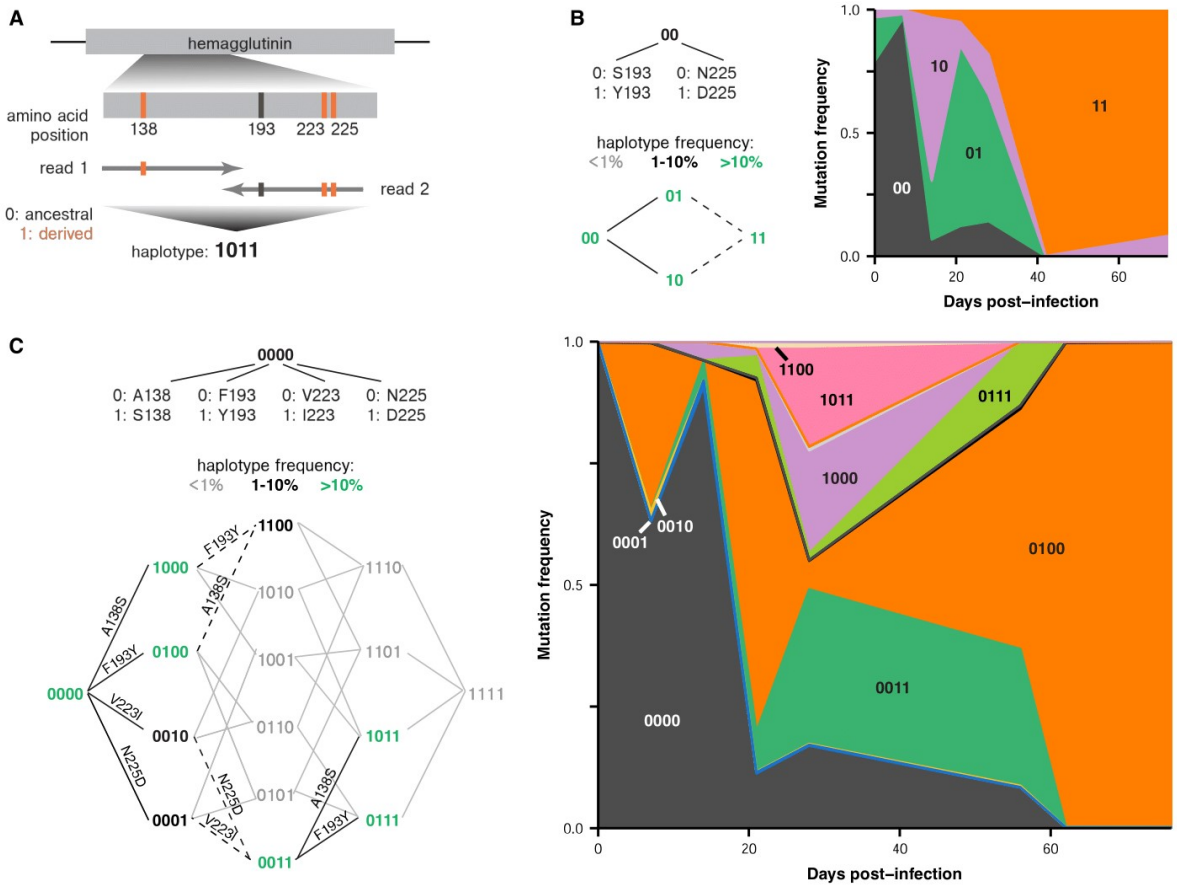

Figure 4.11. Parallel emergence of the same mutations within single infected hosts.

(A) Method for inferring partial haplotypes from short-read sequencing data. We identified paired-end reads that spanned multiple sites of interest along a gene and determined whether the read carried the ancestral or derived allele at each site. **(B)** Frequencies of haplotypes at HA sites 193 and 225 in patient X. Evolutionary paths from the ancestral to double-mutant state are shown, with haplotypes colored according to their maximum frequency during the infection. Solid black lines connect pairs of haplotypes that are both present at a frequency of above 1% and that unambiguously occurred through the indicated mutation. Dashed lines indicate that multiple mutations could have produced a particular haplotype. Gray lines indicate that a mutation did not arise at a detectable frequency on a particular haplotype background. **(C)** Frequencies of haplotypes at HA sites 138, 193, 223, and 225 in patient W. **Figure 4.12** estimates the rate of PCR recombination as described in **Materials and Methods**. **Figures 4.13** and **4.14** show the number of paired-end reads that spanned the mutations in the haplotypes in patients X and W.

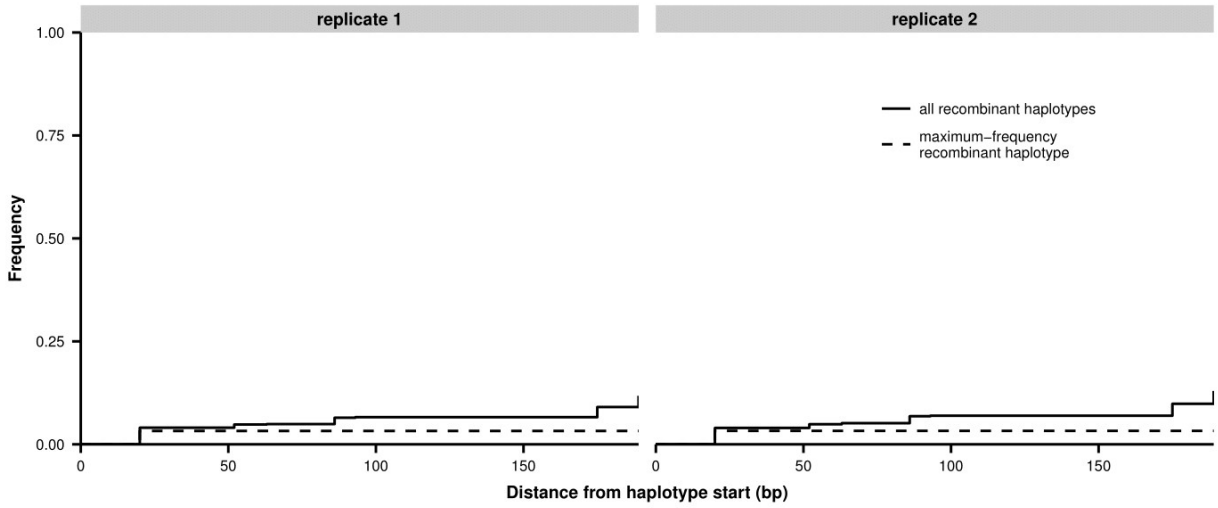

Figure 4.12. Estimate of PCR recombination rate.

The total frequency of all recombinant haplotypes and the frequency of the single most common recombinant haplotype are shown as a function of distance from the haplotype start. We mixed equal volumes of extracted RNA from the first influenza-positive nasal washes from patients W and X and prepared replicate libraries for sequencing as described in **Materials and Methods**. In the absence of PCR recombination, all haplotypes should consist entirely of bases from one or the other original sample: for instance, 00000000 or 11111111. Based on sequencing of the unmixed samples from patients W and X, we identified eight sites of fixed differences within a 200-bp region of the HA gene, and we inferred haplotypes in the mixture sample at these eight sites. Beginning at the first haplotype site, we tallied the proportion of haplotypes that had experienced recombination by each successive site in the haplotype. We did not seek to distinguish between PCR recombination and sequencing errors: the haplotypes 00100000 and 00111111 were both recorded as having experienced recombination by the third haplotype site. The maximum-frequency recombinant haplotype never exceeded 3.5% of the total haplotypes.

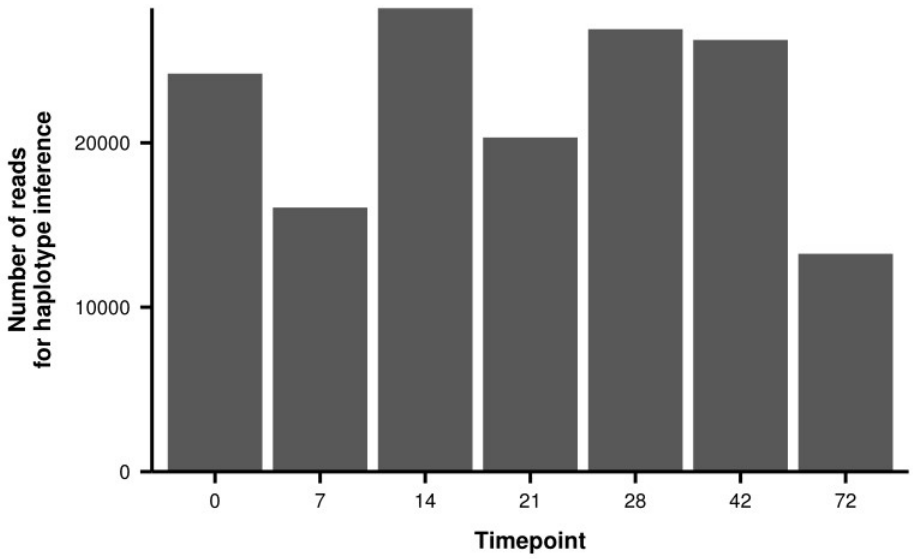

Figure 4.13. Number of paired-end reads used to infer haplotype dynamics in patient X.

Each bar represents the average number of paired-end reads that spanned both variable sites of interest in **Figure 4.11B** across the two sequencing replicates.

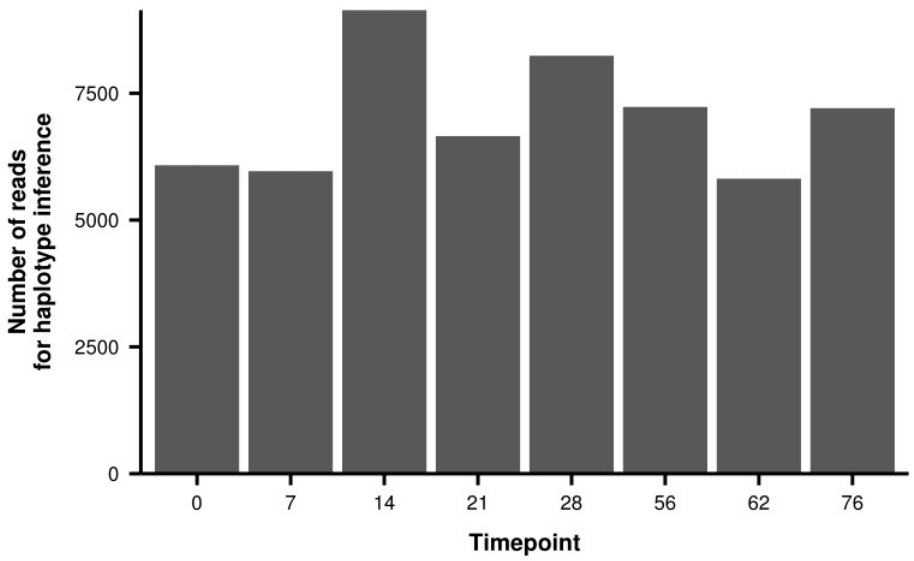

Figure 4.14. Number of paired-end reads used to infer haplotype dynamics in patient W.

Each bar represents the average number of paired-end reads that spanned the four variable sites of interest in **Figure 4.11C** across the two sequencing replicates.

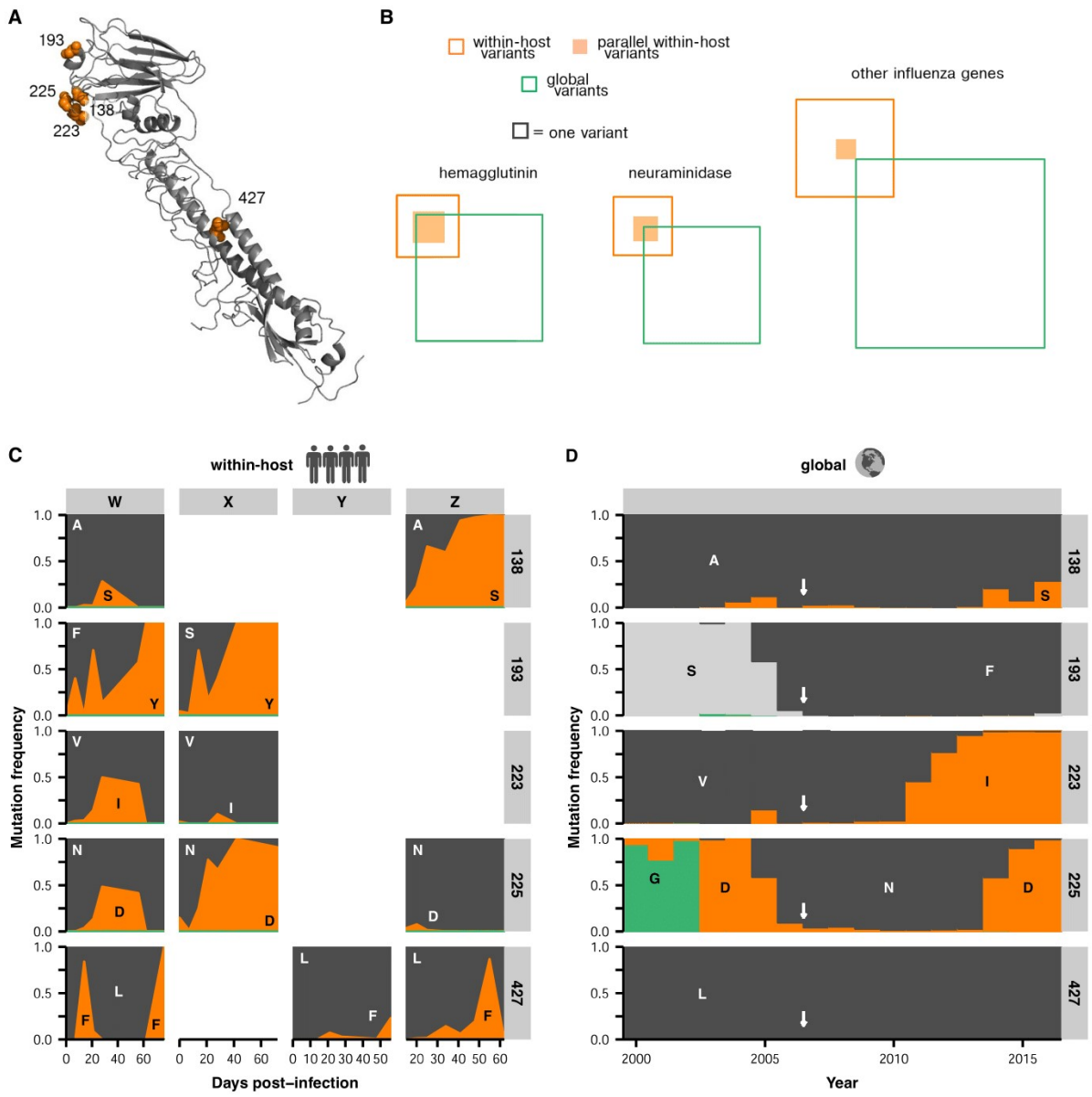

Figure 4.15. Parallel mutations at within-host and global scales.

(A) Sites of parallel within-host mutation plotted on an HA crystal structure (PDB 4HMG (Weis et al., 1990)). **(B)** Overlap of within-host (orange) and global (green) variable sites in HA, NA, and all other influenza genes. Sites at which mutations arise in more than one patient are indicated in solid orange. We defined global variable sites as those at which a variant reached a frequency of at least 10% in a given year after 2000 in the GISAID database of global influenza sequences (Bogner et al., 2006). Numbers of within-host and global mutations are given in **Table 4.2**. **(C)** Mutation frequencies over time within individual patients for parallel within-host mutations. Ancestral identities are colored in gray and mutant ones in orange. **(D)** Global variant frequencies between 2000 and 2015 in H3N2 influenza at sites of parallel within-host mutation. The approximate timing of the patient infections (2006-2007) is indicated by a white arrow.

Figure 4.16 displays variant frequencies for all sites of parallel mutation at the within-host and global scales. **Figure 4.17** describes permutation tests that assess the significance of the overlap in mutations at the within-host and global scales.

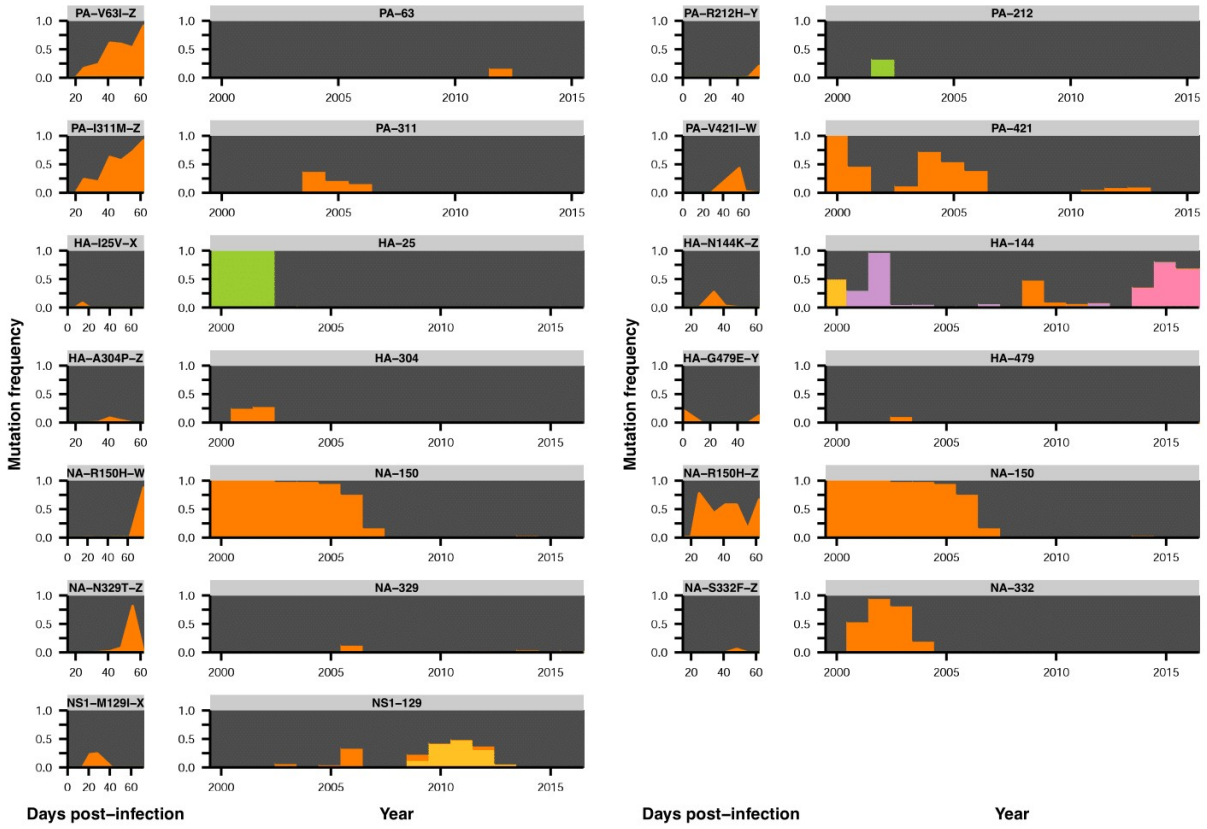

Figure 4.16. Parallel mutations at within-host and global scales.

Mutation frequencies over time are plotted within hosts and in the global population for sites that are variable across both scales. Parallel within-host mutations in HA are shown in **Figure 4.15C** and are omitted. This figure shows mutation frequencies for the remaining sites that were variable across scales. Within-host mutations are labeled by gene name, amino acid change, and patient ID. Ancestral identities are colored in gray and mutant ones in orange at sites of within-host mutation, and the same colors are applied to those amino acids in the global frequency plots where present. Global variant frequencies are shown twice for NA site 150 because mutations arise at that site in two independent patients.

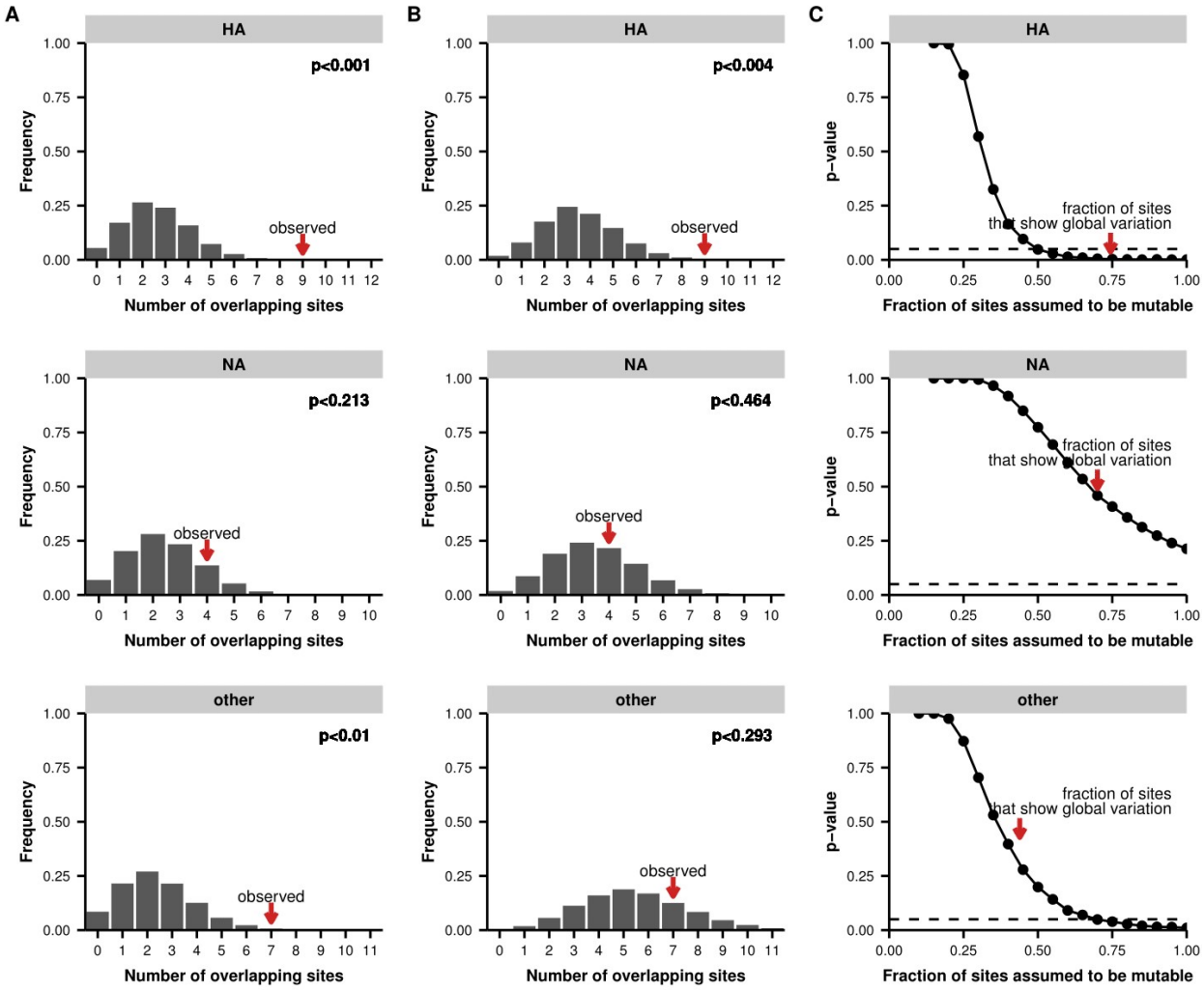

Figure 4.17. Permutation tests for parallel evolution across within-host and global scales.

(A) Distribution of overlapping sites when two sets of sites are drawn at random along the full length of the indicated gene or genes, matching the number of unique variable sites empirically observed in each patient and in the global influenza population (see **Materials and Methods**). These simulations test a simple null model in which each site is equally like to mutate. We calculated the overlap between the two sets of sites in each simulation as a metric of parallelism: greater overlap means that more parallelism has occurred. The p-value indicates the proportion of 100,000 simulations in which the number of overlapping sites is greater than or equal to what is empirically observed. **(B)** Distribution of overlapping sites for simulations as described in **(A)**, performed with a constrained null model in which the fraction of sites considered mutable is the fraction that shows at least two instances of nonsynonymous mutation in the global H3N2 influenza population between 2000 and 2015 (see **Materials and Methods**) to account for constraints on protein evolution. For HA, but not NA or the other influenza genes in aggregate, the observed overlap of mutations at the within- and global scales is

statistically significant under this set of constraints. **(C)** p-values as described in **(A)**, calculated across a range of constraints on the fraction of mutable sites. The constrained null model indicated in **(B)** is indicated with a red arrow. For HA, the observed parallelism is statistically significant at a threshold of 0.05 unless it is assumed that fewer than half the sites in the protein are mutable.

Table 4.4. Primers used for viral deep sequencing.

5’ primer tail sequence is indicated in `plain text`, homology to the U12/U13 regions in **`bold text`**, and gene-specific sequence in **`bold underlined text`**. Primers were modified from the universal primers described in Hoffmann 2001(Hoffmann et al., 2001) to account for the A/G polymorphism at the U4 site of the U12 region.

reaction	primer name	sequence	ratio
reverse transcription	5’-Hoffmann-U12-A4	TATTGGTCTCAGGG**AGCAAAAGCAGG**	1
	5’-Hoffmann-U12-G4	TATTGGTCTCAGGG**AGCGAAAGCAGG**	1
PCR	5’-Hoffmann-Ba-PB2-1-A4	TATTGGTCTCAGGG**AGCAAAAGCAGGTC**	1
	5’-Hoffmann-Ba-PB2-1-G4	TATTGGTCTCAGGG**AGCGAAAGCAGGTC**	1
	3’-Hoffmann-Ba-PB2-2341R	ATATGGTCTCGTATT**AGTAGAAACAAGGTCGTTT**	2
	5’-Hoffmann-Bm-PB1-1-A4	TATTGGTCTCAGGG**AGCAAAAGCAGGCA**	1
	5’-Hoffmann-Bm-PB1-1-G4	TATTGGTCTCAGGG**AGCGAAAGCAGGCA**	1
	3’-Hoffmann-Bm-PB1-2341R	ATATGGTCTCGTATT**AGTAGAAACAAGGCATTT**	2
	5’-Hoffmann-Bm-PA-1-A4	TATTGGTCTCAGGG**AGCAAAAGCAGGTAC**	1
	5’-Hoffmann-Bm-PA-1-G4	TATTGGTCTCAGGG**AGCGAAAGCAGGTAC**	1
	3’-Hoffmann-Bm-PA-2233R	ATATGGTCTCGTATT**AGTAGAAACAAGGTACTT**	2
	5’-Hoffmann-Bm-HA-1-A4	TATTGGTCTCAGGG**AGCAAAAGCAGGGG**	1
	5’-Hoffmann-Bm-HA-1-G4	TATTGGTCTCAGGG**AGCGAAAGCAGGGG**	1
	3’-Hoffmann-Bm-HA-890R	ATATGGTCTCGTATT**AGTAGAAACAAGGGTGTTTT**	2
	5’-Hoffmann-Bm-NP-1-A4	TATTGGTCTCAGGG**AGCAAAAGCAGGGTA**	1
	5’-Hoffmann-Bm-NP-1-G4	TATTGGTCTCAGGG**AGCGAAAGCAGGGTA**	1
	3’-Hoffmann-Bm-NP-1565R	ATATGGTCTCGTATT**AGTAGAAACAAGGGTATTTTT**	2
	5’-Hoffmann-Ba-NA-1-A4	TATTGGTCTCAGGG**AGCAAAAGCAGGAGT**	1
	5’-Hoffmann-Ba-NA-1-G4	TATTGGTCTCAGGG**AGCGAAAGCAGGAGT**	1
	3’-Hoffmann-Ba-NA-1413R	ATATGGTCTCGTATT**AGTAGAAACAAGGAGTTTTTT**	2
	5’-Hoffmann-Bm-M-1-A4	TATTGGTCTCAGGG**AGCAAAAGCAGGTAG**	1
	5’-Hoffmann-Bm-M-1-G4	TATTGGTCTCAGGG**AGCGAAAGCAGGTAG**	1
	3’-Hoffmann-Bm-M-1027R	ATATGGTCTCGTATT**AGTAGAAACAAGGTAGTTTTT**	2
	5’-Hoffmann-Bm-NS-1-A4	TATTGGTCTCAGGG**AGCAAAAGCAGGGTG**	1
	5’-Hoffmann-Bm-NS-1-G4	TATTGGTCTCAGGG**AGCGAAAGCAGGGTG**	1
	3’-Hoffmann-Bm-HA-890R	see above	2

Table 4.5. Overlap of mutations at the within-host and global scales.

Gene	Within-host variants	Parallel within-host variants	Global variants	Overlap, within-host and global variants	Overlap, parallel within-host and global variants
Hemagglutinin	19	5	79	8	4
Neuraminidase	17	3	67	3	1
Other genes	47	2	176	5	0

Chapter 5. RECONCILING DISPARATE ESTIMATES OF VIRAL GENETIC DIVERSITY DURING HUMAN INFLUENZA INFECTIONS

A version of this chapter has previously been published as:

Xue, K.S., Bloom, J.D. Reconciling disparate estimates of viral genetic diversity during human influenza infections. Nature Genetics (2019). DOI: 10.1038/s41588-019-0349-3

A key question in the study of influenza-virus evolution is how rapidly viral genetic variation arises within infected humans, and how much of this genetic diversity is maintained during transmission (McCrone and Lauring, 2018; Xue et al., 2018a). Recently, several studies have measured influenza's within-host genetic diversity in large cohorts of infected humans using high-throughput deep sequencing (**Table 5.1**) (Debbink et al., 2017; Dinis et al., 2016; McCrone et al., 2018; Poon et al., 2016). These studies have disagreed considerably in their estimates of influenza's within-host genetic diversity. In a *Nature Genetics* letter titled "Quantifying influenza virus diversity and transmission in humans" analyzing a household cohort in Hong Kong, Poon et al. (Poon et al., 2016) estimated that within-host genetic diversity is high and 200-250 viral genomes are transmitted between individuals. However, several recent studies conducted in Wisconsin (Dinis et al., 2016), Michigan (McCrone et al., 2018), and Washington (Xue et al., 2018b) that used similar methodologies have found much lower levels of viral genetic diversity. In particular, the Michigan study estimates a narrow

transmission bottleneck of just 1-2 viral genomes (McCrone et al., 2018). These large discrepancies between major published studies are concerning.

Various biological and technical explanations could account for these discrepancies. The studies differ in their geographic locations, cohort design, deep sequencing methods, and data analysis approaches. Any of these factors could affect estimates of viral genetic diversity (Illingworth et al., 2017; McCrone and Lauring, 2016; Sobel Leonard et al., 2017b). Here, we examine whether technical differences in the underlying deep-sequencing datasets or the methods used to analyze them explain the disparate estimates of within-host viral genetic diversity. We identify an anomaly in the Hong Kong data that provides a technical explanation for these discrepancies: read pairs from this study are often split between different biological samples, indicating that some reads are incorrectly assigned. These technical abnormalities explain the high levels of within-host variation and loose transmission bottlenecks reported by this study. Studies without these anomalies consistently report low levels of genetic diversity in acute human influenza infections.

5.1 Results

To systematically compare the results across studies, we used the same computational framework to re-analyze raw sequencing data for four large-scale studies of influenza's within-host genetic diversity, together encompassing more than 500 acute human infections (Debbink et al., 2017; Dinis et al., 2016; McCrone et al., 2018; Poon et al., 2016). For each study, we applied the same variant-calling thresholds as the Hong Kong study (Poon et al., 2016), identifying sites with a minimum coverage of 200 at

which a non-consensus base exceeds a frequency of 3% in the sequenced reads at that site (see **Materials and methods**). We averaged variant frequencies between sequencing replicates where available but otherwise used an analysis pipeline that was as similar as possible across studies to ensure comparable estimates of within-host genetic diversity.

Our analysis recapitulates the major results reported in the Hong Kong study. **Figure 5.1** shows within-host variation in the hemagglutinin gene in H3N2 patients in our re-analysis of the study's data, in the same format as the second figure of the original publication (Poon et al., 2016). In both the original study and our re-analysis, the same within-host variant is often present at similar frequencies in multiple, epidemiologically unrelated individuals. Moreover, the minority variant in one group of samples is typically the majority or consensus variant in the remaining samples (**Figure 5.1A**). Across the hemagglutinin gene, the original Hong Kong study and our re-analysis of that study's data identify the same patterns of within-host variation (**Figure 5.1B**).

Our analysis also identifies major differences between the Hong Kong dataset and the other studies. We find little within-host viral variation in the other three datasets, in line with these studies' stated conclusions (**Figure 5.2A**) (Debbink et al., 2017; Dinis et al., 2016; McCrone et al., 2018). Furthermore, only the Hong Kong dataset contains high-frequency within-host variants that are shared between epidemiologically unrelated individuals. In data from the Hong Kong study, the same within-host variants were shared among more than half of the patients at 42 sites in the H3N2 genome, and 9 sites in the pdmH1N1 genome (**Figure 5.3**). In contrast, we identified no such sites of extensively shared genetic variation among patients in the other three studies. These

results show that the large discrepancies between the Hong Kong study and other published work cannot be accounted for solely by methodological differences in variant calling pipelines.

The extensive shared genetic diversity in the Hong Kong study could result from genuine similarity in the mix of viruses that infect epidemiologically unrelated humans in Hong Kong. But they could also reflect cross-contamination or other abnormalities in the underlying sequencing data. In the course of our analysis, we identified abnormalities in the raw sequencing data from the Hong Kong study that can explain the apparently high levels of shared viral genetic diversity across different infected individuals. The deep sequencing for this study used paired-end Illumina reads. Both reads in a pair come from the same molecule of PCR-amplified viral genetic material, and so should always be assigned to the same infected human (**Figure 5.4A**). Illumina software assigns standard headers to each FASTQ-format sequencing read. These header lines contain information about each read, including the sequencing lane, a unique read-pair identifier, and whether a read is the first or second member of a pair (**Figure 5.4B**). When we analyzed FASTQ headers in the raw sequencing data for the Hong Kong study, we found that paired-end sequencing reads were frequently split between samples assigned to different individuals (**Figure 5.4C**). (**Figures 5.1** and **5.3** were generated by analyzing the sequencing data from the Hong Kong study as single-end data.) For instance, the read `@SOLEXA4_0078:1:1101:10000:101622#ATCACG/1` was associated with study subject 737-V1(0), whereas its pair `@SOLEXA4_0078:1:1101:10000:101622#ATCACG/2` was associated with study subject 741-V1(0), an epidemiologically unrelated individual.

It is biologically impossible for reads in a pair to be associated with distinct individuals, since both reads originate from the same DNA molecule. Across all samples, 70% of reads had corresponding pairs in a FASTQ file assigned to a different individual, and 25% of reads were not part of an identifiable pair (**Figure 5.4C**). Only 5% of the 500 million sequencing reads in this study were associated with the same sample as their corresponding pairs. This splitting of read pairs between samples indicates a problem in the sample index de-multiplexing or downstream computational analysis, and can be considered a form of technical cross-contamination.

Importantly, the problem appears to be with how read pairs were assigned to samples rather than with the FASTQ headers. We found that 93% of the read pairs reconstructed based on FASTQ header information mapped concordantly to the H3N2 or pandemic H1N1 influenza genome—that is, both reads in a pair mapped to the same gene segment in the expected relative orientation.

We analyzed patterns of read-pair splitting between all samples in the study (**Figure 5.4D**). We identified four disjoint sets of samples for which read pairs are split extensively within sets, but never between sets. Further analysis of FASTQ headers showed that all of the sequencing reads from each cluster were derived from the same flowcell lane. Poon et al. 2016 report that samples were amplified in duplicate and that replicates were sequenced on distinct flowcell lanes. Indeed, we find that each set of samples corresponds almost exactly to one set of replicate samples for one of the two influenza subtypes sequenced in this study (**Figure 5.4D**). This finding was robust to the computational analysis pipeline: the first author generated all of the figures in this paper, but the last author conducted an independent re-analysis of the data to reach similar

conclusions (see **Materials and methods**). Altogether, these analyses suggest that read pairs are split extensively between samples of a given influenza subtype in the Hong Kong study.

Without access to the full computational pipeline for the Hong Kong study, we cannot determine directly whether the first read, second read, or both members of split read pairs were assigned to samples incorrectly. However, when we analyzed only the first read of each pair, we found low within-host diversity, in line with other studies (**Figure 5.2B**, **Figure 5.4E**). In contrast, the second read of each pair was responsible for the high viral diversity reported in the Hong Kong study. These results suggest that the second member of each read pair may have been incorrectly assigned, and the first member may more accurately represent the low levels of within-host viral diversity.

5.2 Discussion

This splitting of read pairs between unrelated samples has important consequences for estimates of viral genetic diversity within human infections. Even if each individual were infected with a clonal population of influenza virus, read-pair splitting would create the appearance of high levels of shared genetic diversity between unrelated individuals. For instance, at a site in the influenza genome where some individuals exclusively have nucleotide A and others exclusively have nucleotide T, read-pair splitting would make it seem as though all individuals with majority identity A have minority variant T and vice versa, even in the absence of genuine within-host variation. The high-frequency shared viral diversity within human hosts in the Hong Kong study corresponds closely to what would be expected from read-pair splitting

(**Figure 5.1A**), suggesting that this abnormality may be responsible for the published results.

Read-pair splitting may also explain why the Hong Kong household cohort study estimates a loose transmission bottleneck for human influenza virus of 200-250 viral genomes (Poon et al., 2016; Sobel Leonard et al., 2017b), compared to a Michigan household cohort study that estimates a bottleneck size of 1-2 viral genomes (McCrone et al., 2018). Splitting of read pairs between samples would also create the appearance of shared within-host variation in donor and recipient individuals in a transmission chain, resulting in estimates of a looser transmission bottleneck.

Our finding of read-pair splitting in the Hong Kong dataset provides a technical explanation for major discrepancies in recent studies of the genetic diversity of human influenza viruses. In particular, these technical anomalies may account for the Hong Kong study's finding of shared, high-frequency viral diversity within human hosts (Poon et al., 2016) and its estimate of a loose transmission bottleneck between hosts (Poon et al., 2016; Sobel Leonard et al., 2017b). If we exclude the Hong Kong study, then all other studies report low levels of within-host genetic diversity for human influenza virus (Debbink et al., 2017; Dinis et al., 2016; McCrone et al., 2018).

5.3 Materials and methods

5.3.1 Data and code availability

We downloaded sequencing data generated by the Hong Kong study (Poon et al., 2016) from https://www.synapse.org/#!Synapse:syn8033988, following the methods of a study that re-analyzed data from the Hong Kong study to estimate transmission

bottleneck sizes using a new analytical method (Sobel Leonard et al., 2017b). We obtained sequencing data for the Wisconsin study (Dinis et al., 2016) by personal communication. We downloaded sequencing data for the other studies from SRA BioProject PRJNA344659 (Debbink et al., 2017) and PRJNA412631 (McCrone et al., 2018). See Life Sciences Reporting Summary for more details.

The computer code that performs the analysis is available at https://github.com/ksxue/compare-flu-within-hosts-public. All figures were generated by the first author using this code base, but the last author independently conducted an analysis of read pairing and came to similar conclusions (https://github.com/jbloomlab/reanalyze_Poon_et_al).

5.3.2 Variant calling

Here, we summarize our general computational pipeline for variant calling, with modifications for individual studies described below. We used cutadapt 1.8.3 (Martin, 2011) to trim Nextera adapter sequences, remove bases at the ends of reads with a Q-score below 24, and filter out reads whose remaining length was shorter than 20 bases. To determine the sample subtype, we used bowtie2 to map 1000 reads from each sample to reference genomes for each influenza subtype: A/Victoria/361/2011 (H3N2), A/California/04/2009 (pdmH1N1), and A/Boston/12/2007 (seasonal H1N1). We determined what proportion of the reads from each sample mapped to each reference genome and classified sample subtype based which reference genome resulted in the highest mapping rate.

For each sample, we first aligned reads to the subtype reference genome using bowtie2 and the --very-sensitive setting (Langmead and Salzberg, 2012), then we used

custom scripts to tally the counts of each base at each genome position and infer a consensus sequence for that sample. We then realigned the reads to the sample consensus sequence, discarding reads with a mapping score below 20 and bases with a Q-score below 20, and we removed read duplicates using Picard version 1.43. We used custom scripts to tally the counts of each base at each genome position among the remaining reads.

We defined sites of within-host variation as positions in the genome with sequencing coverage of at least 200 reads at which a minority base is present at a frequency of at least 3%, following the variant-calling criteria of the Hong Kong study (Poon et al., 2016). The Hong Kong study sequenced most samples in duplicate, and we required within-host variants to be present at a frequency of 3% in both sequencing replicates. To maintain consistent variant-calling criteria for all samples in the Hong Kong study, we did not include samples with only a single sequencing replicate in our downstream analyses. We performed no other filtering for sample or variant quality because common filtering metrics like sequencing replicates (Xue et al., 2017) and plasmid sequencing controls (McCrone and Lauring, 2016; McCrone et al., 2018) were not universally available.

5.3.3 Study-specific modifications

We tried to analyze all sequencing data using methods that were as similar as possible across studies. However, certain study designs or data formats required us to modify our basic analysis framework as described below. For more information, the code that performs the analysis is available at https://github.com/ksxue/compare-flu-within-hosts-public.

Hong Kong study. We obtained data from the Hong Kong study (Poon et al., 2016) from https://www.synapse.org/#!Synapse:syn8033988 in the form of a single FASTQ file per biological sample containing first and second members of read pairs. As described in the main text and in the section below, we found that read pairs were frequently split between different sample files, so we could not conduct a meaningful analysis of paired-end sequencing reads. To call variants and generate the data in Figures 1 and 2, we analyzed the sequencing data as single-end reads. We used single-end read-mapping settings for bowtie2, and we did not perform read deduplication, which typically makes use of paired-end read information. Most samples from this study were sequenced in duplicate. We only analyzed samples for which we were able to identify both sequencing replicates, and we required within-host variants to meet the variant-calling criteria in both replicates.

Wisconsin study. We obtained data from the Wisconsin study (Dinis et al., 2016) by personal communication in the form of a single FASTQ file per biological sample containing first and second members of read pairs. We reconstructed read pairs for each sample using read-pair information in the FASTQ headers, and we found no read pairs that were split between different sample files. We then analyzed the reconstructed read pairs as described above.

5.3.4 Analysis of read pairing

To analyze read pairing in data from the Hong Kong study, we parsed FASTQ read headers to identify first and second members of read pairs, as well as their associated sequencing lane. We interpreted the colon-delimited fields of a FASTQ header like @SOLEXA4_0078:1:1101:10000:101622#ATCACG/1 to contain information

about the sequencing instrument (@SOLEXA4_0078), lane number (1), tile (1101), cluster coordinate (10000:101622), sequencing index (ATCACG), and whether a read was the first or second member of a read pair (1), in accordance with Illumina specifications (for instance, see http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm). We did not make use of sequencing indices in this analysis. We used this FASTQ information to analyze all 500 million reads in the Hong Kong dataset to determine what proportion of reads had pairs in the same sample file and to determine which reads were associated with each sequencing lane. We used the --very-fast-local setting of bowtie2 to map reconstructed read pairs against a pan-influenza reference genome produced by concatenating the eight-segment A/Victoria/361/2011 (H3N2) reference genome and the eight-segment A/California/04/2009 (pdmH1N1) reference genome into a single reference genome containing sixteen segments.

5.3.5 Analysis of first and second reads

To separate the first and second reads assigned to each sample file in the Hong Kong dataset, we interpreted the colon-delimited fields of the FASTQ headers as described above. Because read pairs were frequently split between sample files, the first and second reads that we identified in each sample did not necessarily constitute complete pairs. We then used the read-mapping and variant-calling pipeline described above for the Hong Kong study to identify sites of within-host variation.

5.3.6 Acknowledgments

We thank P. Green for helpful comments on the manuscript. K.S.X is supported by the Hertz Foundation Myhrvold Family Fellowship. The work of J.D.B was supported by grant R01AI127893 from the NIAID of the NIH. J.D.B. is an Investigator of the Howard Hughes Medical Institute.

5.4 FIGURES AND TABLES

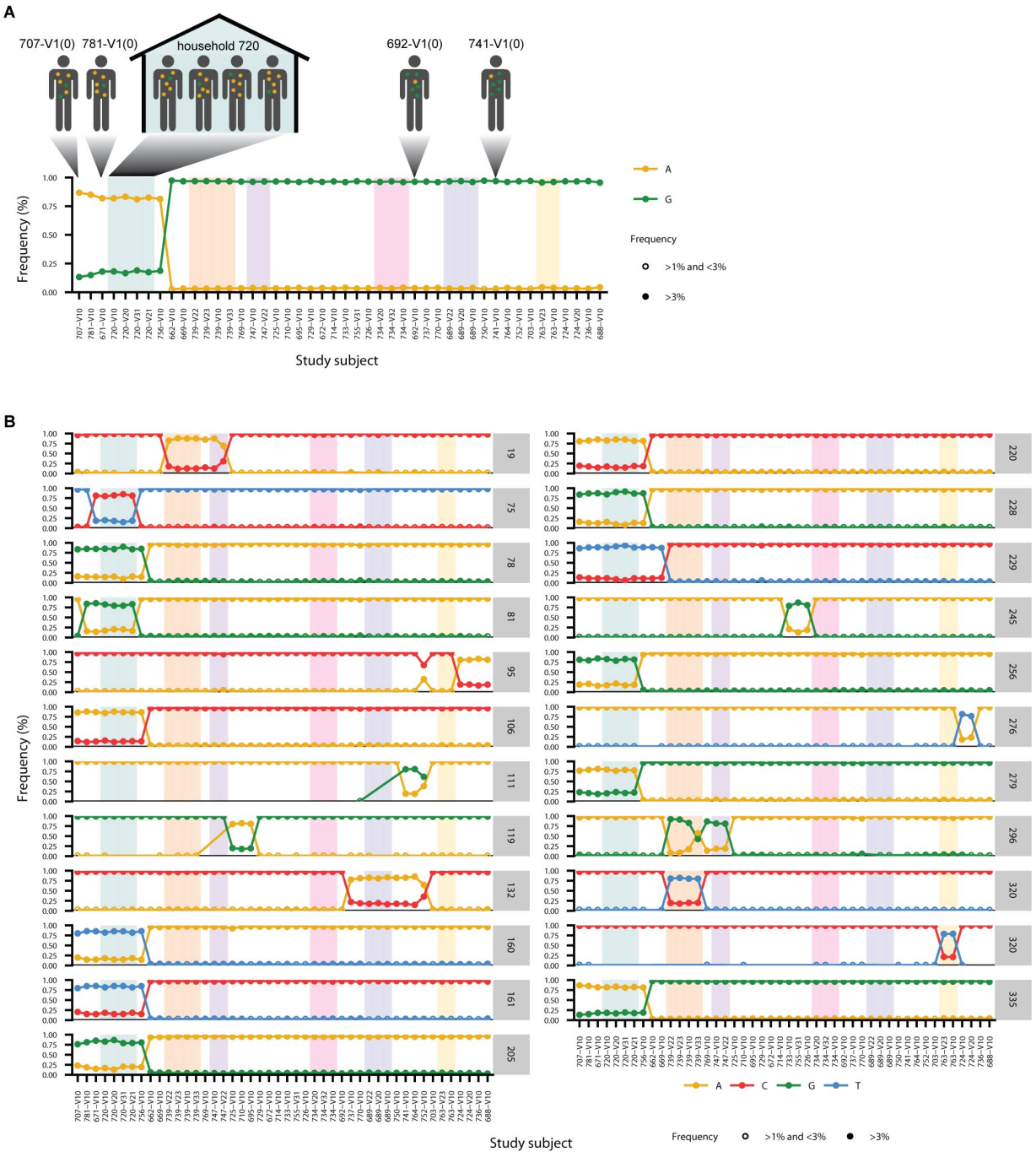

Figure 5.1. High within-host genetic diversity of human influenza virus in our re-analysis of sequencing data from the Hong Kong study.

This figure mimics the format of the second figure of Poon et al (2016) and shows that our re-analysis recapitulates the main reported results of high-frequency shared genetic

diversity between epidemiologically unrelated individuals. **(A)** Viral genetic diversity at hemagglutinin codon 335 in H3N2 human influenza infections. At this codon, both variants encode the same amino acid. This plot shows within-host variants that were present at a frequency of at least 1% in both sequencing replicates at sites with minimum sequencing coverage of 200 reads. Shaded regions indicate individuals from the same household. **(B)** Viral genetic diversity in the HA1 domain of hemagglutinin in H3N2 human influenza infections in our re-analysis. Each panel represents a separate site in the genome and is labeled by the codon it represents. Sites shown harbored within-host variation at a frequency of at least 3% in both sequencing replicates for at least two samples in the study.

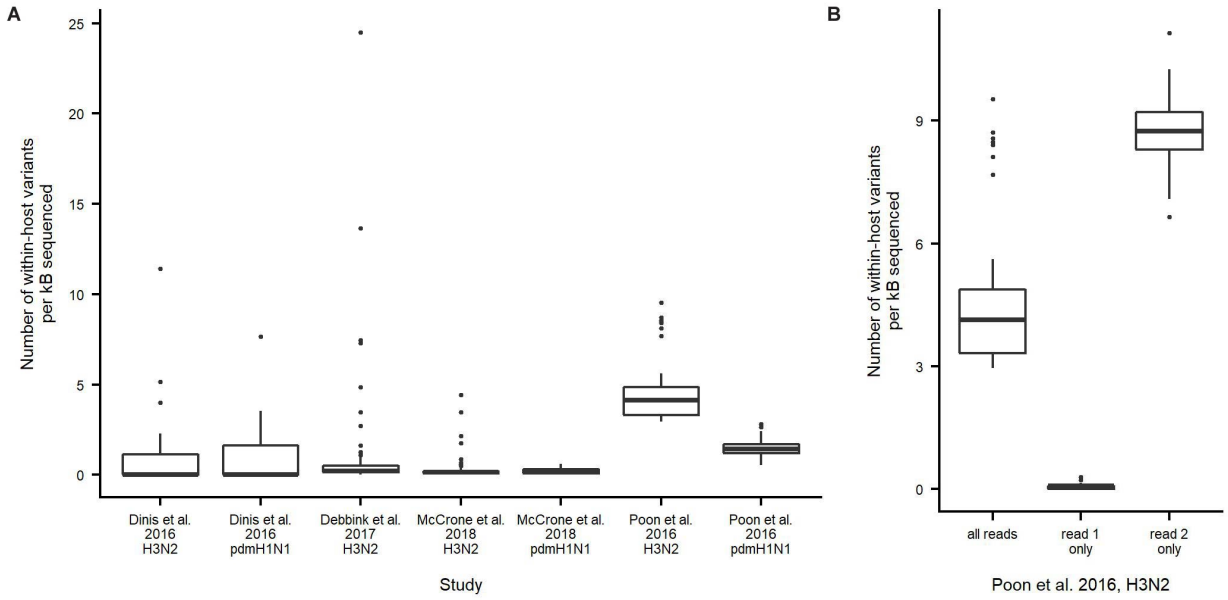

Figure 5.2. Comparison of within-host genetic diversity across studies and within different sequencing reads in a study.

(A) Number of within-host variants identified in each sample in each study, normalized to the length of the genome sequenced in each study. For each sample, we identified within-host variants that were present at a frequency of at least 3% at sites with minimum sequencing coverage of 200 reads. The center line of each box plot displays the median value; the box limits display upper and lower quartiles; and the whiskers extend up to 1.5 times the interquartile range. The number of samples in each study is listed in **Table 5.1**. **(B)** Number of within-host variants identified in the 46 H3N2 samples when analyzing both members of each sequenced read pair, just read 1, or just read 2. Variants were called and data plotted as in **(A)**.

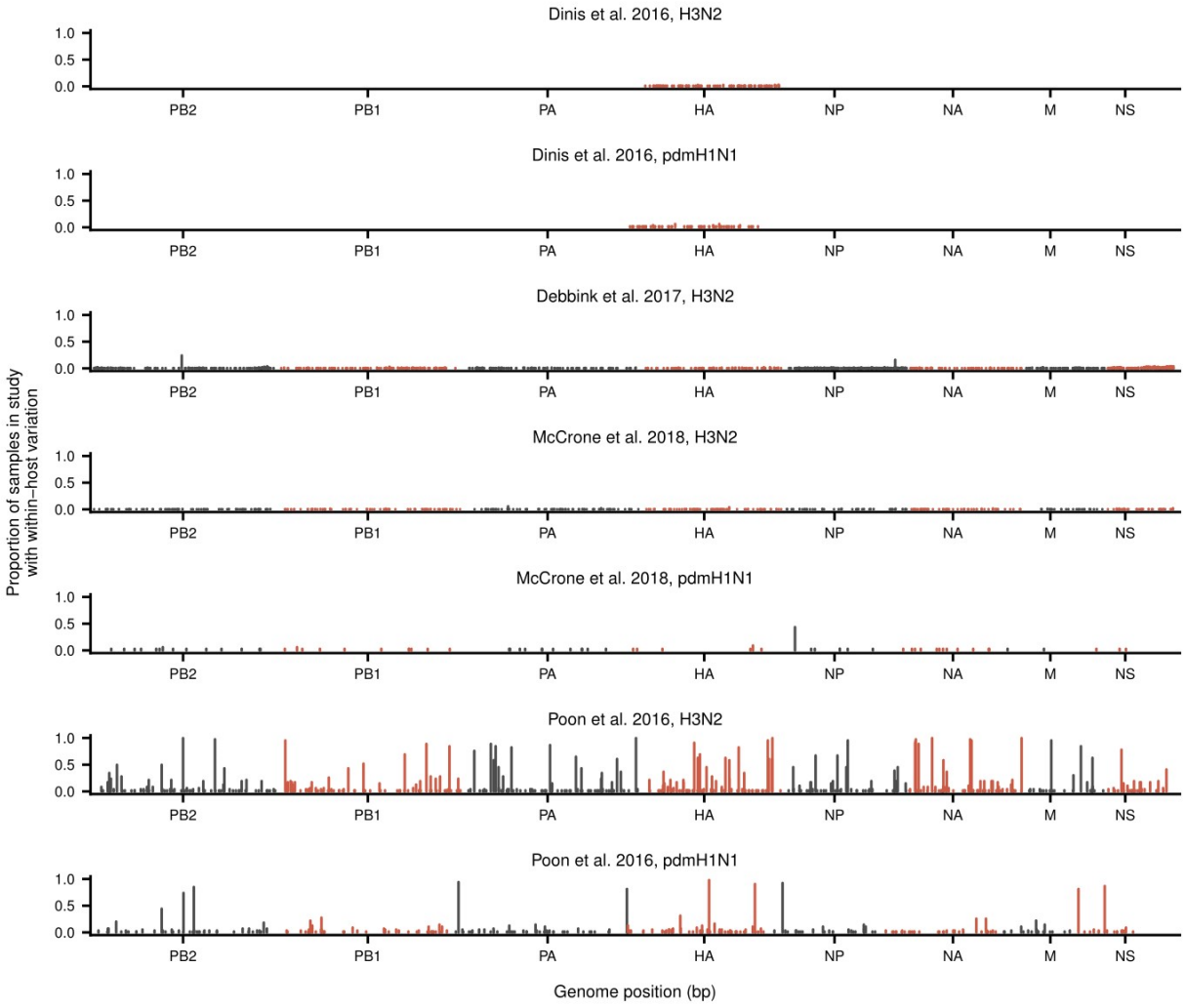

Figure 5.3. Comparison of shared within-host viral genetic diversity in four large-scale deep-sequencing studies of human influenza virus.

Proportion of samples in each study in which we identified within-host variation at each genome site. For each sample, we identified within-host variants that were present at a frequency of at least 3% at sites with minimum sequencing coverage of 200 reads. Our re-analysis is consistent with the previously reported results of each study: we find little shared genetic diversity in the data from the Dinis et al. (2016), Debbink et al. (2017), and McCrone et al. (2018) studies, but we observe high shared genetic diversity in the data from the Poon et al study.

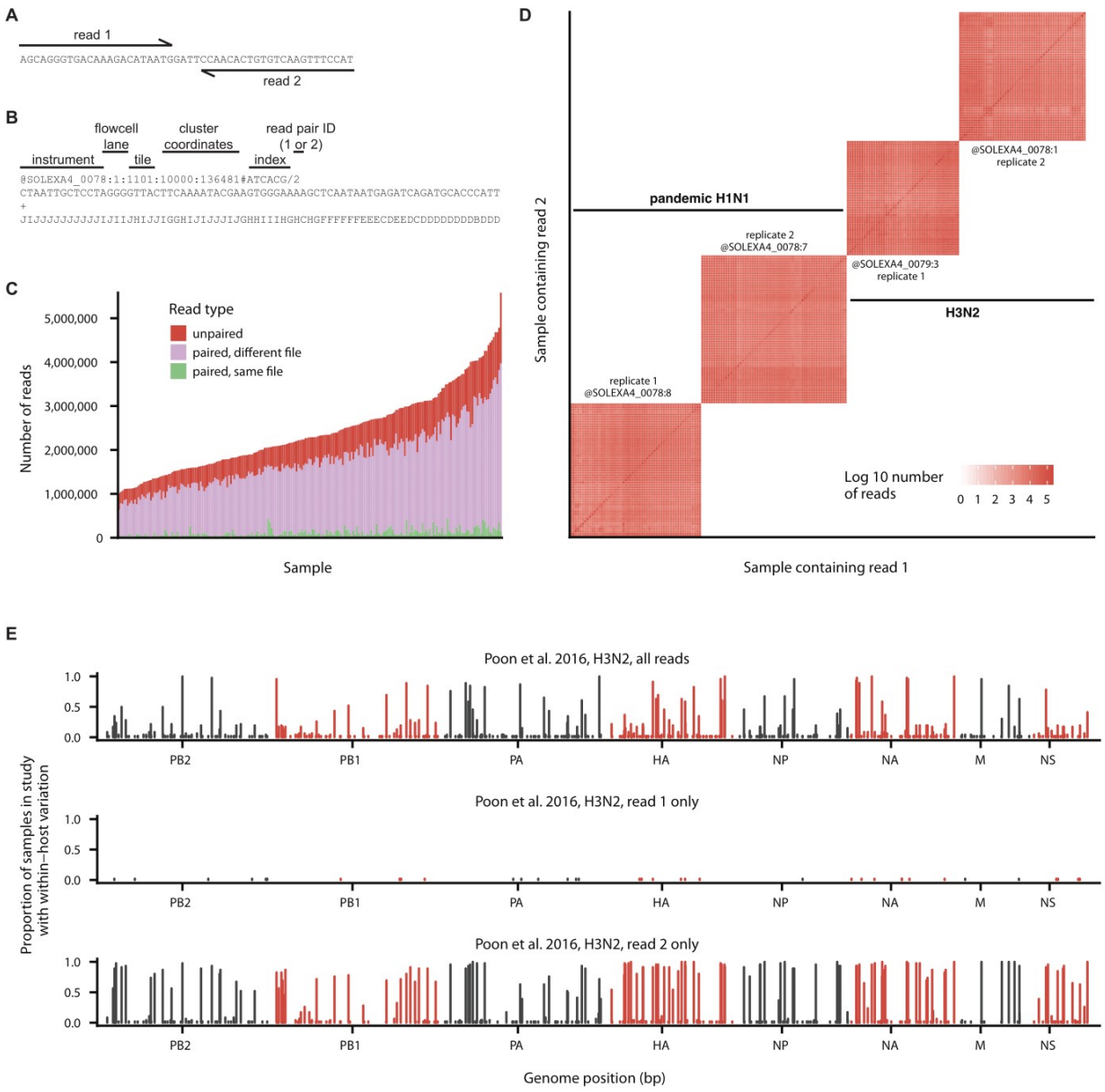

Figure 5.4. Paired-end sequencing reads are frequently split between samples that were run on the same sequencing lane.

(A) Paired-end sequencing reads are derived from the same physical DNA molecule. **(B)** The FASTQ header for each sequencing read provides information about the sequencing instrument, flowcell lane, tile, cluster coordinates, and sequencing index for each read, as well as whether the read is the first or second member of a read pair. **(C)** Sequencing reads from the Hong Kong dataset are frequently split between distinct biological samples. **(D)** Hierarchical clustering of the number of read pairs split between each pair of samples in the Hong Kong study. Sequencing reads from the Hong Kong dataset are split between four distinct clusters of samples. All sequencing reads in each cluster are derived from the same flowcell lane and correspond to one set of replicate samples for one of the two influenza subtypes sequenced in the study. **(E)** Proportion of

samples for which we identified within-host variation at each genome site when analyzing both reads for a pair, just read 1, or just read 2. For each sample, we identified within-host variants that were present at a frequency of at least 3% at sites with minimum sequencing coverage of 200 reads.

Table 5.6. Large-scale deep-sequencing studies of human influenza virus.

Study	Location	H3N2 samples	pdmH1N1 samples	Methods	Findings
Dinis et al. 2016	Wisconsin	68	46	Targeted deep sequencing of hemagglutinin gene.	Low genetic diversity. Possible low-frequency antigenic variants.
Poon et al. 2016*	Hong Kong	46	54	Whole-genome deep sequencing using multi-segment RT-PCR followed by sequence-independent, single-primer amplification.	High genetic diversity, frequent mixed infections, and a loose transmission bottlenecks in a household cohort study.
Debbink et al. 2017	Michigan	121	0	Whole-genome deep sequencing using multi-segment RT-PCR.	Low genetic diversity. No differences in genetic diversity between vaccinated and unvaccinated individuals.
McCrone et al. 2018**	Michigan	217	32	Whole-genome deep sequencing using multi-segment RT-PCR.	Low genetic diversity and narrow transmission bottlenecks in a household cohort study.

* The Hong Kong study performed sequencing in duplicate for most samples in the study. We only analyzed samples for which we were able to identify both sequencing replicates. We count samples collected from the same individual at different time points as separate samples in this summary.

** We count samples collected from the same individual at different time points as separate samples in this summary.

Chapter 6. CONCLUSION

In my dissertation, I investigated how evolutionary forces act on influenza populations across different spatiotemporal scales. I showed that deep sequencing can reveal evolutionary dynamics like viral cooperation and within-host evolution that are difficult to observe using traditional sequencing methods. Here, I summarize the main findings of my dissertation and suggest directions for future work.

6.1 How do cooperative interactions shape influenza evolution?

In the first part of my dissertation, I described a cooperative interaction between distinct variants of H3N2 influenza in cell culture. In chapter 2, I showed that a mixture of the neuraminidase variants D151 and G151 outgrew either variant alone, and viral populations stably maintained an approximately equal frequency of the two viral variants through several passages in cell culture (Xue et al., 2016). A mixture of the D151 and G151 variants had frequently been documented by clinical sequencing laboratories when influenza samples were passaged in cell culture, but it was initially unclear whether the two variants co-existed natural clinical infections as well. In chapter 3, I demonstrated that the G151 variant was undetectable in unpassaged clinical samples, even when these samples went on to develop a mixture of the D151 and G151 variants after being passaged in cell culture (Xue et al., 2018c). These findings suggested that cooperation was confined to cell-culture settings rather than natural clinical infections.

This work expands our understanding of viral evolutionary dynamics by providing one of the first examples of cooperation between distinct viral variants. RNA viruses

mutate rapidly to form genetically diverse populations, and prior work on poliovirus (Vignuzzi et al., 2006) and Coxsackie virus (Bordería et al., 2015) has suggested that cooperative interactions between spontaneously generated variants might contribute to overall population fitness. Additionally, recent work with measles (Shirogane et al., 2012) and hepatitis B (Cao et al., 2014) has identified specific viral variants that cooperate with one another. However, specific instances of viral cooperation remain rare and poorly understood. The example of viral cooperation that we have identified is unusual because it has arisen spontaneously and robustly in many laboratory settings (Chambers et al., 2014; Lee et al., 2013; Lin et al., 2010; McKimm-Breschkin et al., 2003; Mishin et al., 2014; Mohr et al., 2015; Tamura et al., 2013), to the point that it can interfere with common assays used to characterize influenza's antigenicity and resistance to antiviral drugs (Lee et al., 2013; Lin et al., 2010; Tamura et al., 2013).

Cooperative interactions arise rapidly between influenza variants in cell culture, but they appear to be absent in natural clinical infections (Lee et al., 2013; Mishin et al., 2014; Xue et al., 2018c). The discrepancy between cell-culture and clinical settings raises interesting questions about the mechanism of cooperation and the conditions that favor its emergence. We find, for instance, that cooperative interactions are enhanced at high multiplicities of infection (Xue et al., 2016), whereas natural human infections appear to have low effective rates of reassortment that suggest low multiplicities of infection (Sobel Leonard et al., 2017a). Moreover, the growth defects of the G151 variant may be tolerated better in cell culture than in natural human infections, where the neuraminidase protein helps cleave mucins and perform other functions important for viral replication (Cohen et al., 2013; Yang et al., 2014). Recent work also suggests

that influenza infections are founded by 1 to 2 viral genomes (McCrone et al., 2018), and the short duration of acute infections may make it unlikely for a mixture of variants at neuraminidase site 151 to reach detectable frequencies. Further work is needed to understand what combination of factors limits this instance of influenza to cell-culture settings.

We also found that cooperation between the D151 and G151 variants was limited to particular genetic backgrounds (Xue et al., 2016). Cooperative interactions were present in the background of the A/Hanoi/Q118/2007 (H3N2) but not the A/Wisconsin/67/2005 (H3N2) hemagglutinin gene, which may explain why mixtures of the D151 and G151 neuraminidase variants are more commonly observed in global influenza sequence databases after 2007. Epistatic interactions between the hemagglutinin and neuraminidase genes have been extensively documented (Gulati et al., 2005; Neverov et al., 2015; Wagner et al., 2002), but the mechanistic basis of this background-dependence remains unclear. Similar cooperative interactions may also occur in influenza viruses of different subtypes. Our study focused on H3N2 influenza, but one recent report suggests that cooperation may also occur between D151 and G151 neuraminidase variants in a pandemic H1N1 genetic background (Gong et al., 2019). Additional work that clarifies how genetic background influences the emergence of cooperative interactions will help elucidate the extent to which viral cooperation plays a role in influenza evolution.

6.2 How does antigenic selection shape influenza's evolution within hosts?

In the second part of my dissertation, I examined the evolution of influenza viruses within infected individuals. In chapter 4, I tracked influenza's evolutionary dynamics within four chronically infected patients (Xue et al., 2017). In hemagglutinin, a small set of antigenic variants arose recurrently within these patients and in multiple patients in our study, and many of these variants also reached high frequencies in the global influenza population. This work demonstrated that influenza can evolve rapidly within infected individuals in ways that mirror its global evolution.

Antigenic selection has long been recognized as a major force driving influenza's global evolution (Fitch et al., 1991; Petrova and Russell, 2017; Smith et al., 2004b), and our work shows that this antigenic evolution can occur rapidly at the level of individual hosts. However, many questions remain about how antigenic selection drives viral evolution across different evolutionary scales.

Although we observe antigenic variants reaching high frequencies in the chronically infected individuals in our study, chronic influenza infections are uncommon, and most antigenic variants are expected to arise in more typical, acute infections. However, it remains unclear to what extent antigenic variants are enriched within acute influenza infections. Several large-scale studies of influenza's within-host genetic diversity have documented the presence of antigenic variants at low frequencies in acute infections (Dinis et al., 2016; McCrone et al., 2018), but it is unclear whether these antigenic variants are enriched relative to non-antigenic variants (McCrone et al., 2018). The short duration (Carrat et al., 2008) and narrow transmission bottleneck

(McCrone et al., 2018) of acute influenza infections may make it difficult for antigenic variants to reach frequencies detectable by deep sequencing. Moreover, the adaptive immune response tends to activate late in acute infections (Beauchemin and Handel, 2011), which may mean that it exerts a limited influence on evolutionary dynamics within hosts. In contrast, the multi-week influenza infections that we studied appear to provide enough time for antigenic selection to act detectably on viral populations within hosts. In this respect, the evolutionary dynamics that we observe in chronic influenza infections resemble those in HIV and HCV infections, which are marked by the constant emergence and fixation of novel antigenic variants (Lemey et al., 2006; Rambaut et al., 2004; Simmonds, 2004).

The relative paucity of antigenic variants within acute infections may suggest that the emergence and initial transmission of antigenic variants is shaped heavily by genetic drift. However, further work is needed to characterize these early stages in the emergence of antigenic variants.

6.3 What factors influence the transmission bottleneck size of influenza?

Two large-scale household studies have produced widely differing estimates of the transmission bottleneck size of influenza. (Poon et al., 2016) estimated a transmission bottleneck size of 200 to 250 viral genomes, compared to an estimate of 1 to 2 viral genomes by (McCrone et al., 2018). In chapter 5, I resolve the discrepancy between these two estimates. I show that read pairs in the (Poon et al., 2016) raw sequencing data are commonly split between distinct samples, a form of technical

cross-contamination that contributes to the study's estimate of a loose transmission bottleneck (Xue and Bloom, 2019). Our identification of this technical issue helps resolve a major source of confusion in our understanding of influenza evolution.

Transmission bottlenecks play a major role in our understanding of how viral genetic diversity within hosts is transformed into variation between hosts (McCrone and Lauring, 2018). Loose transmission bottlenecks provide opportunities for *de novo* viral variants to increase in frequency through the course of several infections. But narrow transmission bottlenecks suggest that stochastic processes like genetic drift dominate the earliest stages of viral evolution. Our study supports the (McCrone et al., 2018) finding of a narrow transmission bottleneck. However, more independent studies are needed to understand how bottleneck size differs based on factors like mode of transmission.

It also remains unclear whether and how selection for antigenic variants might occur as viruses compete to found new infections. Two studies that characterized the transmission of avian-adapted influenza viruses in mammalian hosts (Moncla et al., 2016) and cell-culture-adapted influenza viruses in human hosts (Sobel Leonard et al., 2016) found evidence that selection acted strongly at transmission to favor better-adapted variants. It remains uncertain to what degree selection might act at transmission in natural human infections to remove deleterious variants or to favor variants with novel antigenic properties.

6.4 CONCLUDING REMARKS

High-throughput sequencing technologies act as a microscope, allowing us to observe previously invisible evolutionary dynamics that occur at small scales of space and time. In my dissertation, I use deep sequencing to characterize evolutionary dynamics like viral cooperation and to document early stages of viral evolution within infected individuals. Deep sequencing is still a relatively new technique, and best practices for generating and interpreting deep-sequencing data remain an active area of research and discussion (Grubaugh et al., 2019; McCrone and Lauring, 2016). Moreover, existing population-genetic methods for analyzing evolutionary dynamics are rarely suitable for interpreting deep-sequencing data that has relatively high error rates and poor haplotype resolution compared to traditional consensus-sequencing techniques. Continued development of techniques for producing and analyzing high-quality deep-sequencing data will shed light on key questions in viral evolution.

BIBLIOGRAPHY

Alizon, S., Luciani, F., and Regoes, R.R. (2011). Epidemiological and clinical consequences of within-host evolution. Trends Microbiol. *19*, 24–32.

Andersen, K.G., Shapiro, B.J., Matranga, C.B., Sealfon, R., Lin, A.E., Moses, L.M., Folarin, O.A., Goba, A., Odia, I., Ehiane, P.E., et al. (2015). Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell *162*, 738–750.

Andino, R., and Domingo, E. (2015). Viral quasispecies. Virology *479–480*, 46–51.

Baccam, P., Beauchemin, C., Macken, C.A., Hayden, F.G., and Perelson, A.S. (2006). Kinetics of influenza A virus infection in humans. J. Virol. *80*, 7590–7599.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. (2008). The influenza virus resource at the National Center for Biotechnology Information. J. Virol. *82*, 596–601.

Baz, M., Abed, Y., McDonald, J., and Boivin, G. (2006). Characterization of multidrug-resistant influenza A/H3N2 viruses shed during 1 year by an immunocompromised child. Clin. Infect. Dis. *43*, 1555–1561.

Baz, M., Abed, Y., Papenburg, J., Bouhy, X., Hamelin, M.-E., and Boivin, G. (2009). Emergence of oseltamivir-resistant pandemic H1N1 virus during prophylaxis. N. Engl. J. Med. *361*, 2296–2297.

Beauchemin, C.A., and Handel, A. (2011). A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead. BMC Public Health *11*, S7.

Bedford, T., Suchard, M.A., Lemey, P., Dudas, G., Gregory, V., Hay, A.J., McCauley,

J.W., Russell, C.A., Smith, D.J., and Rambaut, A. (2014). Integrating influenza antigenic dynamics with molecular evolution. Elife *2014*, e01914.

Beerenwinkel, N., Günthard, H.F., Roth, V., and Metzner, K.J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. *3*, 329.

Bhatt, S., Holmes, E.C., and Pybus, O.G. (2011). The genomic rate of molecular adaptation of the human influenza A virus. Mol. Biol. Evol. *28*, 2443–2451.

Bloom, J.D. (2014). An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol. Biol. Evol. *31*, 1956–1978.

Bloom, J.D., Gong, L.I., and Baltimore, D. (2010). Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science *328*, 1272–1275.

Bogner, P., Capua, I., Lipman, D.J., and Cox, N.J. (2006). A global initiative on sharing avian flu data. Nature *442*, 981–981.

Boivin, G., Goyette, N., and Bernatchez, H. (2002). Prolonged excretion of amantadine-resistant influenza a virus quasi species after cessation of antiviral therapy in an immunocompromised patient. Clin. Infect. Dis. *34*, E23-5.

Boni, M.F., Zhou, Y., Taubenberger, J.K., and Holmes, E.C. (2008). Homologous recombination is very rare or absent in human influenza A virus. J. Virol. *82*, 4807–4811.

Bordería, A. V, Isakov, O., Moratorio, G., Henningsson, R., Agüera-González, S., Organtini, L., Gnädig, N.F., Blanc, H., Alcover, A., Hafenstein, S., et al. (2015). Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. PLoS Pathog. *11*, e1004838.

Bozic, I., Gerold, J.M., and Nowak, M.A. (2016). Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. PLOS Comput. Biol. *12*, e1004731.

Brand, C., and Palese, P. (1980). Sequential passage of influenza virus in embryonated eggs or tissue culture: Emergence of mutants. Virology *107*, 424–433.

Brankston, G., Gitterman, L., Hirji, Z., Lemieux, C., and Gardam, M. (2007). Transmission of influenza A in human beings. Lancet Infect. Dis. *7*, 257–265.

Campbell, A.P., Guthrie, K.A., Englund, J.A., Farney, R.M., Minerich, E.L., Kuypers, J., Corey, L., and Boeckh, M. (2015). Clinical Outcomes Associated With Respiratory Virus Detection Before Allogeneic Hematopoietic Stem Cell Transplant. Clin. Infect. Dis. *61*, 192–202.

Cao, L., Wu, C., Shi, H., Gong, Z., Zhang, E., Wang, H., Zhao, K., Liu, S., Li, S., Gao, X., et al. (2014). Coexistence of hepatitis B virus quasispecies enhances viral replication and the ability to induce host antibody and cellular immune responses. J. Virol. *88*, 8656–8666.

Carrat, F., Vergu, E., Ferguson, N.M., Lemaitre, M., Cauchemez, S., Leach, S., and Valleron, A.-J. (2008). Time Lines of Infection and Disease in Human Influenza: A Review of Volunteer Challenge Studies. Am. J. Epidemiol. *167*, 775–785.

Chambers, B.S., Li, Y., Hodinka, R.L., and Hensley, S.E. (2014). Recent H3N2 influenza virus clinical isolates rapidly acquire hemagglutinin or neuraminidase mutations when propagated for antigenic analyses. J. Virol. *88*, 10986–10989.

Chandrasekaran, A., Srinivasan, A., Raman, R., Viswanathan, K., Raguram, S., Tumpey, T.M., Sasisekharan, V., and Sasisekharan, R. (2008). Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin. Nat. Biotechnol. *26*,

107–113.

Ciota, A.T., Ehrbar, D.J., Van Slyke, G.A., Willsey, G.G., and Kramer, L.D. (2012). Cooperative interactions in the West Nile virus mutant swarm. BMC Evol. Biol. *12*, 58.

De Clercq, E. (2006). Antiviral agents active against influenza A viruses. Nat. Rev. Drug Discov. *5*, 1015–1025.

Cohen, M., Zhang, X.Q., Senaati, H.P., Chen, H.W., Varki, N.M., Schooley, R.T., and Gagneux, P. (2013). Influenza A penetrates host mucus by cleaving sialic acids with neuraminidase. Virol. J.

Cushing, A., Kamali, A., Winters, M., Hopmans, E.S., Bell, J.M., Grimes, S.M., Xia, L.C., Zhang, N.R., Moss, R.B., Holodniy, M., et al. (2015). Emergence of Hemagglutinin Mutations During the Course of Influenza Infection. Sci Rep *5*, 16178.

Davis, A.R., Hiti, A.L., and Nayak, D.P. (1980). Influenza defective interfering viral RNA is formed by internal deletion of genomic RNA. Proc. Natl. Acad. Sci. U. S. A. *77*, 215–219.

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.

Deyde, V.M., Okomo-Adhiambo, M., Sheu, T.G., Wallis, T.R., Fry, A., Dharan, N., Klimov, A.I., and Gubareva, L. V (2009). Pyrosequencing as a tool to detect molecular markers of resistance to neuraminidase inhibitors in seasonal influenza A viruses. Antiviral Res. *81*, 16–24.

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.

Dong, G., Peng, C., Luo, J., Wang, C., Han, L., Wu, B., Ji, G., and He, H. (2015). Adamantane-resistant influenza A viruses in the world (1902-2013): Frequency and distribution of M2 gene mutations. PLoS One *10*, e0119115.

Dutta, R.N., Rouzine, I.M., Smith, S.D., Wilke, C.O., and Novella, I.S. (2008). Rapid Adaptive Amplification of Preexisting Variation in an RNA Virus. J. Virol. *82*, 4354–4362.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften *58*, 465–523.

Feder, A.F., Rhee, S.-Y., Holmes, S.P., Shafer, R.W., Petrov, D.A., and Pennings, P.S. (2016). More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. Elife *5*, e10670.

Fitch, W.M., Leiter, J.M., Li, X.Q., and Palese, P. (1991). Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. U. S. A. *88*, 4270–4274.

Fonville, J.M., Wilks, S.H., James, S.L., Fox, A., Ventresca, M., Aban, M., Xue, L., Jones, T.C., Le, N.M.H., Pham, Q.T., et al. (2014). Antibody landscapes after influenza virus infection or vaccination. Science (80-.). *346*, 996–1000.

Frensing, T., Heldt, F.S., Pflugmacher, A., Behrendt, I., Jordan, I., Flockerzi, D., Genzel, Y., and Reichl, U. (2013). Continuous influenza virus production in cell culture shows a periodic accumulation of defective interfering particles. PLoS One *8*, e72288.

Frise, R., Bradley, K., van Doremalen, N., Galiano, M., Elderfield, R.A., Stilwell, P., Ashcroft, J.W., Fernandez-Alonso, M., Miah, S., Lackenby, A., et al. (2016). Contact

transmission of influenza virus between ferrets imposes a looser bottleneck than respiratory droplet transmission allowing propagation of antiviral resistance. Sci. Rep. *6*, 29793.

Gallet, R., Fabre, F., Michalakis, Y., and Blanc, S. (2017). The number of target molecules of the amplification step limits accuracy and sensitivity in ultra deep sequencing viral population studies. J. Virol. JVI.00561-17.

Ghedin, E., Sengamalay, N.A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D.J., Sitz, J., Koo, H., Bolotov, P., et al. (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature *437*, 1162–1166.

Ghedin, E., Laplante, J., DePasse, J., Wentworth, D.E., Santos, R.P., Lepow, M.L., Porter, J., Stellrecht, K., Lin, X., Operario, D., et al. (2011). Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. J. Infect. Dis. *203*, 168–174.

Gong, Y.-N., Tsao, K.-C., Chen, G.-W., Wu, C.-J., Chen, Y.-H., Liu, Y.-C., Yang, S.-L., Huang, Y.-C., and Shih, S.-R. (2019). Population dynamics at neuraminidase position 151 of influenza A (H1N1)pdm09 virus in clinical specimens. J. Gen. Virol.

Grenfell, B.T. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science (80-.). *303*, 327–332.

Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., Mumford, J.A., and Holmes, E.C. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science (80-.). *303*.

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. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. *20*, 8.

Gulati, S., Smith, D.F., Cummings, R.D., Couch, R.B., Griesemer, S.B., St George, K., Webster, R.G., and Air, G.M. (2013). Human H3N2 Influenza Viruses Isolated from 1968 To 2012 Show Varying Preference for Receptor Substructures with No Apparent Consequences for Disease or Spread. PLoS One *8*, e66325.

Gulati, U., Wu, W., Gulati, S., Kumari, K., Waner, J.L., and Air, G.M. (2005). Mismatched hemagglutinin and neuraminidase specificities in recent human H3N2 influenza viruses. Virology *339*, 12–20.

Hamada, N., Imamura, Y., Hara, K., Kashiwagi, T., Imamura, Y., Nakazono, Y., Chijiwa, K., and Watanabe, H. (2012). Intrahost emergent dynamics of oseltamivir-resistant virus of pandemic influenza A (H1N1) 2009 in a fatally immunocompromised patient. J. Infect. Chemother. *18*, 865–871.

Hegreness, M., Shoresh, N., Hartl, D., and Kishony, R. (2006). An Equivalence Principle for the Incorporation of Favorable Mutations in Asexual Populations. Science (80-.). *311*, 1615–1617.

Hensley, S.E., Das, S.R., Bailey, A.L., Schmidt, L.M., Hickman, H.D., Jayaraman, A., Viswanathan, K., Raman, R., Sasisekharan, R., Bennink, J.R., et al. (2009). Hemagglutinin Receptor Binding Avidity Drives Influenza A Virus Antigenic Drift. Science (80-.). *326*, 734–736.

Herbeck, J.T., Nickle, D.C., Learn, G.H., Gottlieb, G.S., Curlin, M.E., Heath, L., and Mullins, J.I. (2006). Human immunodeficiency virus type 1 env evolves toward

ancestral states upon transmission to a new host. J. Virol. *80*, 1637–1644.

Hoelzer, K., Murcia, P.R., Baillie, G.J., Wood, J.L.N., Metzger, S.M., Osterrieder, N., Dubovi, E.J., Holmes, E.C., and Parrish, C.R. (2010). Intrahost evolutionary dynamics of canine influenza virus in naive and partially immune dogs. J. Virol. *84*, 5329–5335.

Hoffmann, E., Neumann, G., Kawaoka, Y., Hobom, G., and Webster, R.G. (2000). A DNA transfection system for generation of influenza A virus from eight plasmids. Proc. Natl. Acad. Sci. U. S. A. *97*, 6108–6113.

Hoffmann, E., Stech, J., Guan, Y., Webster, R.G., and Perez, D.R. (2001). Universal primer set for the full-length amplification of all influenza A viruses. Arch. Virol. *146*, 2275–2289.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., and VandePol, S. (1982). Rapid evolution of RNA genomes. Science (80-.). *215*, 1577–1585.

Holmes, E.C. (2003). Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J. Virol. *77*, 11296–11298.

Holmes, E.C. (2010). The RNA virus quasispecies: fact or fiction? J. Mol. Biol. *400*, 271–273.

Hooper, K. a, and Bloom, J.D. (2013). A mutant influenza virus that uses an N1 neuraminidase as the receptor-binding protein. J. Virol. *87*, 12531–12540.

Hughes, J., Allen, R.C., Baguelin, M., Hampson, K., Baillie, G.J., Elton, D., Newton, J.R., Kellam, P., Wood, J.L.N., Holmes, E.C., et al. (2012). Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. PLoS Pathog. *8*, e1003081.

Illingworth, C.J.R. (2015). Fitness inference from short-read data: Within-host evolution

of a reassortant H5N1 influenza virus. Mol. Biol. Evol. *32*, 3012–3026.

Illingworth, C.J.R. (2016). SAMFIRE: Multi-locus variant calling for time-resolved sequence data. Bioinformatics *32*, 2208–2209.

Illingworth, C.J.R., Fischer, A., and Mustonen, V. (2014). Identifying Selection in the Within-Host Evolution of Influenza Using Viral Sequence Data. PLoS Comput. Biol. *10*, e1003755.

Illingworth, C.J.R., Roy, S., Beale, M.A., Tutill, H., Williams, R., and Breuer, J. (2017). On the effective depth of viral sequence data. Virus Evol. *3*, vex030.

Iwasaki, A., and Pillai, P.S. (2014). Innate immunity to influenza virus infection. Nat. Rev. Immunol. *14*, 315–328.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). J. Biosci. Bioeng. *96*, 317–323.

Kao, K.C., and Sherlock, G. (2008). Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. Nat. Genet. *40*, 1499–1504.

Ke, R., Aaskov, J., Holmes, E.C., and Lloyd-Smith, J.O. (2013). Phylodynamic analysis of the emergence and epidemiological impact of transmissible defective dengue viruses. PLoS Pathog. *9*, e1003193.

Koel, B.F., Burke, D.F., Bestebroer, T.M., van der Vliet, S., Zondag, G.C.M., Vervaet, G., Skepner, E., Lewis, N.S., Spronken, M.I.J., Russell, C.A., et al. (2013). Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. Science (80-.). *342*.

Kryazhimskiy, S., and Plotkin, J.B. (2008). The population genetics of dN/dS. PLoS

Genet. *4*, e1000304.

Kryazhimskiy, S., Dushoff, J., Bazykin, G.A., and Plotkin, J.B. (2011). Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet. *7*.

Kugelman, J.R., Wiley, M.R., Nagle, E.R., Reyes, D., Pfeffer, B.P., Kuhn, J.H., Sanchez-Lockhart, M., and Palacios, G.F. (2017). Error baseline rates of five sample preparation methods used to characterize RNA virus populations. PLoS One *12*, e0171333.

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.

Lang, G.I., Rice, D.P., Hickman, M.J., Sodergren, E., Weinstock, G.M., Botstein, D., and Desai, M.M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature *500*, 571–574.

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

Lässig, M., Mustonen, V., and Walczak, A.M. (2017). Predicting evolution. Nat. Ecol. Evol. *1*, 0077.

Lauring, A.S., and Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. PLoS Pathog. *6*, e1001005.

Lee, H.K., Tang, J.W.-T., Kong, D.H.-L., Loh, T.P., Chiang, D.K.-L., Lam, T.T.-Y., and Koay, E.S.-C. (2013). Comparison of mutation patterns in full-genome A/H3N2 influenza sequences obtained directly from clinical samples and the same samples

after a single MDCK passage. PLoS One *8*, e79252.

Lemey, P., Rambaut, A., and Pybus, O.G. (2006). HIV evolutionary dynamics within and among hosts. AIDS Rev. *8*, 125–140.

Leslie, A.J., Pfafferott, K.J., Chetty, P., Draenert, R., Addo, M.M., Feeney, M., Tang, Y., Holmes, E.C., Allen, T., Prado, J.G., et al. (2004). HIV evolution: CTL escape mutation and reversion after transmission. Nat. Med. *10*, 282–289.

Li, Y., Myers, J.L., Bostick, D.L., Sullivan, C.B., Madara, J., Linderman, S.L., Liu, Q., Carter, D.M., Wrammert, J., Esposito, S., et al. (2013). Immune history shapes specificity of pandemic H1N1 influenza antibody responses. J. Exp. Med. *210*.

Lin, Y.P., Gregory, V., Collins, P., Kloess, J., Wharton, S., Cattle, N., Lackenby, A., Daniels, R., and Hay, A. (2010). Neuraminidase receptor binding variants of human influenza A(H3N2) viruses resulting from substitution of aspartic acid 151 in the catalytic site: a role in virus attachment? J. Virol. *84*, 6769–6781.

Lin, Y.P., Xiong, X., Wharton, S.A., Martin, S.R., Coombs, P.J., Vachieri, S.G., Christodoulou, E., Walker, P.A., Liu, J., Skehel, J.J., et al. (2012). Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. Proc. Natl. Acad. Sci. U. S. A. *109*, 21474–21479.

Linderman, S.L., Chambers, B.S., Zost, S.J., Parkhouse, K., Li, Y., Herrmann, C., Ellebedy, A.H., Carter, D.M., Andrews, S.F., Zheng, N.-Y., et al. (2014). Potential antigenic explanation for atypical H1N1 infections among middle-aged adults during the 2013-2014 influenza season. Proc. Natl. Acad. Sci. U. S. A. *111*, 15798–15803.

Liu, C., and Air, G.M. (1993). Selection and characterization of a neuraminidase-minus mutant of influenza virus and its rescue by cloned neuraminidase genes. Virology *194*,

403–407.

Lowen, A.C. (2017). Constraints, Drivers, and Implications of Influenza A Virus Reassortment. Annu. Rev. Virol. *4*, 105–121.

Lowen, A.C., Mubareka, S., Steel, J., and Palese, P. (2007). Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature. PLoS Pathog. *3*, e151.

Łuksza, M., and Lässig, M. (2014). A predictive fitness model for influenza. Nature *507*, 57–61.

Luria, S.E., and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. Genetics *28*, 491–511.

Marshall, N., Priyamvada, L., Ende, Z., Steel, J., and Lowen, A.C. (2013). Influenza Virus Reassortment Occurs with High Frequency in the Absence of Segment Mismatch. PLoS Pathog. *9*, e1003421.

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

Matrosovich, M.N., Matrosovich, T.Y., Gray, T., Roberts, N.A., and Klenk, H.-D. (2004). Neuraminidase is important for the initiation of influenza virus infection in human airway epithelium. J. Virol. *78*, 12665–12667.

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.

(2017). The evolutionary dynamics of influenza A virus within and between human hosts. Doi.Org 176362.

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.

McGinnis, J., Laplante, J., Shudt, M., and George, K.S. (2016). Next generation sequencing for whole genome analysis and surveillance of influenza A viruses. J. Clin. Virol. *79*, 44–50.

McKimm-Breschkin, J.L. (2000). Resistance of influenza viruses to neuraminidase inhibitors — a review. Antiviral Res. *47*, 1–17.

McKimm-Breschkin, J., Trivedi, T., Hampson, A., Hay, A., Klimov, A., Tashiro, M., Hayden, F., and Zambon, M. (2003). Neuraminidase Sequence Analysis and Susceptibilities of Influenza Virus Clinical Isolates to Zanamivir and Oseltamivir. Antimicrob. Agents Chemother. *47*, 2264–2272.

McMinn, P., Carrello, A., Cole, C., Baker, D., and Hampson, A. (1999). Antigenic drift of influenza A (H3N2) virus in a persistently infected immunocompromised host is similar to that occurring in the community. Clin. Infect. Dis. *29*, 456–458.

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*, vew026.

Memoli, M.J., Athota, R., Reed, S., Czajkowski, L., Bristol, T., Proudfoot, K., Hagey, R., Voell, J., Fiorentino, C., Ademposi, A., et al. (2014). The natural history of influenza infection in the severely immunocompromised vs nonimmunocompromised hosts.

Clin. Infect. Dis. *58*, 214–224.

Mishin, V.P., Sleeman, K., Levine, M., Carney, P.J., Stevens, J., and Gubareva, L. V. (2014). The effect of the MDCK cell selected neuraminidase D151G mutation on the drug susceptibility assessment of influenza A(H3N2) viruses. Antiviral Res. *101*, 93–96.

Mitnaul, L.J., Matrosovich, M.N., Castrucci, M.R., Tuzikov, A.B., Bovin, N. V., Kobasa, D., and Kawaoka, Y. (2000). Balanced Hemagglutinin and Neuraminidase Activities Are Critical for Efficient Replication of Influenza A Virus. J. Virol. *74*, 6015–6020.

Mohr, P.G., Deng, Y.-M., and McKimm-Breschkin, J.L. (2015). The neuraminidases of MDCK grown human influenza A(H3N2) viruses isolated since 1994 can demonstrate receptor binding. Virol. J. *12*, 67.

Moncla, L.H., Zhong, G., Nelson, C.W., Dinis, J.M., Mutschler, J., Hughes, A.L., Watanabe, T., Kawaoka, Y., and Friedrich, T.C. (2016). Selective Bottlenecks Shape Evolutionary Pathways Taken during Mammalian Adaptation of a 1918-like Avian Influenza Virus. Cell Host Microbe *19*, 169–180.

Mugal, C.F., Wolf, J.B.W., and Kaj, I. (2014). Why time matters: Codon evolution and the temporal dynamics of dN/dS. Mol. Biol. Evol. *31*, 212–231.

Murcia, P.R., Baillie, G.J., Daly, J., Elton, D., Jervis, C., Mumford, J.A., Newton, R., Parrish, C.R., Hoelzer, K., Dougan, G., et al. (2010). Intra- and interhost evolutionary dynamics of equine influenza virus. J. Virol. *84*, 6943–6954.

Murcia, P.R., Hughes, J., Battista, P., Lloyd, L., Baillie, G.J., Ramirez-Gonzalez, R.H., Ormond, D., Oliver, K., Elton, D., Mumford, J.A., et al. (2012). Evolution of an Eurasian avian-like influenza virus in naïve and vaccinated pigs. PLoS Pathog. *8*,

e1002730.

Murcia, P.R., Baillie, G.J., Stack, J.C., Jervis, C., Elton, D., Mumford, J.A., Daly, J., Kellam, P., Grenfell, B.T., Holmes, E.C., et al. (2013). Evolution of equine influenza virus in vaccinated horses. J. Virol. *87*, 4768–4771.

Najera, I., Holguin, A., Quinones-Mateu, M., Munoz-Fernandez, M., Najera, R., Lopez-Galindez, C., and Domingo, E. (1995). Pol gene quasispecies of human immunodeficiency virus: mutations associated with drug resistance in virus from patients undergoing no drug therapy. J. Virol. *69*, 23–31.

Neher, R.A. (2013). Genetic Draft, Selective Interference, and Population Genetics of Rapid Adaptation. Annu. Rev. Ecol. Evol. Syst. *44*, 195–215.

Neher, R.A., Russell, C.A., and Shraiman, B.I. (2014). Predicting evolution from the shape of genealogical trees. Elife *3*, 3332–3333.

Neher, R.A., Bedford, T., Daniels, R.S., Russell, C.A., and Shraiman, B.I. (2016). Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc. Natl. Acad. Sci. U. S. A. *113*, E1701-9.

Neverov, A.D., Kryazhimskiy, S., Plotkin, J.B., and Bazykin, G.A. (2015). Coordinated Evolution of Influenza A Surface Proteins. PLoS Genet. *11*, e1005404.

Nichols, W.G., Guthrie, K.A., Corey, L., and Boeckh, M. (2004). Influenza infections after hematopoietic stem cell transplantation: risk factors, mortality, and the effect of antiviral therapy. Clin. Infect. Dis. *39*, 1300–1306.

Nobusawa, E., and Sato, K. (2006). Comparison of the mutation rates of human influenza A and B viruses. J. Virol. *80*, 3675–3678.

Novella, I.S., Reissig, D.D., and Wilke, C.O. (2004). Density-dependent selection in

vesicular stomatitis virus. J. Virol. *78*, 5799–5804.

Oh, D.Y., Barr, I.G., Mosse, J.A., and Laurie, K.L. (2008). MDCK-SIAT1 cells show improved isolation rates for recent human influenza viruses compared to conventional MDCK cells. J. Clin. Microbiol. *46*, 2189–2194.

Okomo-Adhiambo, M., Nguyen, H.T., Sleeman, K., Sheu, T.G., Deyde, V.M., Garten, R.J., Xu, X., Shaw, M.W., Klimov, A.I., and Gubareva, L. V (2010). Host cell selection of influenza neuraminidase variants: implications for drug resistance monitoring in A(H1N1) viruses. Antiviral Res. *85*, 381–388.

Park, D.J., Dudas, G., Wohl, S., Goba, A., Whitmer, S.L.M., Andersen, K.G., Sealfon, R.S., Ladner, J.T., Kugelman, J.R., Matranga, C.B., et al. (2015). Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell *161*, 1516–1526.

Pauly, M.D., Procario, M.C., and Lauring, A.S. (2017). A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. Elife *6*, e26437.

Petrova, V.N., and Russell, C.A. (2017). The evolution of seasonal influenza viruses. Nat. Rev. Microbiol. nrmicro.2017.118.

Pfeiffer, J.K., and Kirkegaard, K. (2005). Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. PLoS Pathog. *1*, e11.

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.

Posada-Cespedes, S., Seifert, D., and Beerenwinkel, N. (2017). Recent advances in

inferring viral diversity from high-throughput sequencing data. Virus Res. *239*, 17–32.

Pybus, O.G., and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. *10*, 540–550.

Rambaut, A., Posada, D., Crandall, K.A., and Holmes, E.C. (2004). The causes and consequences of HIV evolution. Nat. Rev. Genet. *5*, 52–61.

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

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

Renaud, C., Kuypers, J., and Englund, J.A. (2011). Emerging oseltamivir resistance in seasonal and pandemic influenza A/H1N1. J. Clin. Virol. *52*, 70–78.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

van Riel, D., Munster, V.J., de Wit, E., Rimmelzwaan, G.F., Fouchier, R.A.M., Osterhaus, A.D.M.E., and Kuiken, T. (2006). H5N1 Virus Attachment to Lower Respiratory Tract. Science *312*, 399.

Rocha, E., Cox, N.J., Black, R.A., Harmon, M.W., Harrison, C.J., and Kendal, A.P. (1991). Antigenic and genetic variation in influenza A (H1N1) virus isolates recovered from a persistently infected immunodeficient child. J. Virol. *65*, 2340–2350.

Rogers, M.B., Song, T., Sebra, R., Greenbaum, B.D., Hamelin, M.-E., Fitch, A., Twaddle, A., Cui, L., Holmes, E.C., Boivin, G., et al. (2015). Intrahost dynamics of

antiviral resistance in influenza A virus reflect complex patterns of segment linkage, reassortment, and natural selection. MBio *6*, e02464-14-.

Saira, K., Lin, X., DePasse, J. V, Halpin, R., Twaddle, A., Stockwell, T., Angus, B., Cozzi-Lepri, A., Delfino, M., Dugan, V., et al. (2013). Sequence analysis of in vivo defective interfering-like RNA of influenza A H1N1 pandemic virus. J. Virol. *87*, 8064–8074.

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

Shinya, K., Ebina, M., Yamada, S., Ono, M., Kasai, N., and Kawaoka, Y. (2006). Avian flu: influenza virus receptors in the human airway. Nature *440*, 435–436.

Shirogane, Y., Watanabe, S., and Yanagi, Y. (2012). Cooperation between different RNA virus genomes produces a new phenotype. Nat. Commun. *3*, 1235.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus--15 years on. J. Gen. Virol. *85*, 3173–3188.

Skowronski, D.M., Janjua, N.Z., De Serres, G., Sabaiduc, S., Eshaghi, A., Dickinson, J.A., Fonseca, K., Winter, A.-L., Gubbay, J.B., Krajden, M., et al. (2014). Low 2012–13 Influenza Vaccine Effectiveness Associated with Mutation in the Egg-Adapted H3N2 Vaccine Strain Not Antigenic Drift in Circulating Viruses. PLoS One *9*, e92153.

Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., and Fouchier, R.A.M. (2004a). Mapping the Antigenic and Genetic Evolution of Influenza Virus. Science (80-.). *305*, 371–376.

Smith, D.J., Lapedes, A.S., De Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., and Fouchier, R.A.M. (2004b). Mapping the antigenic and

genetic evolution of influenza virus. Science (80-.).

Sobel Leonard, A., 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.

Sobel Leonard, A., 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. (2017a). The effective rate of influenza reassortment is limited during human infection. PLOS Pathog. *13*, e1006203.

Sobel Leonard, A., Weissman, D.B., Greenbaum, B., Ghedin, E., and Koelle, K. (2017b). Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. J. Virol. *91*, JVI.00171-17.

Squires, R.B., Noronha, J., Hunt, V., García-Sastre, A., Macken, C., Baumgarth, N., Suarez, D., Pickett, B.E., Zhang, Y., Larsen, C.N., et al. (2012). Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respi. Viruses *6*, 404–416.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Strelkowa, N., and Lässig, M. (2012). Clonal interference in the evolution of influenza. Genetics *192*, 671–682.

Suarez-Lopez, P., and Ortin, J. (1994). An estimation of the nucleotide substitution rate at defined positions in the influenza virus haemagglutinin gene. J. Gen. Virol. *75*, 389–393.

Tamura, D., Nguyen, H.T., Sleeman, K., Levine, M., Mishin, V.P., Yang, H., Guo, Z., Okomo-Adhiambo, M., Xu, X., Stevens, J., et al. (2013). Cell culture-selected substitutions in influenza A(H3N2) neuraminidase affect drug susceptibility assessment. Antimicrob. Agents Chemother. *57*, 6141–6146.

Tao, H., Li, L., White, M.C., Steel, J., and Lowen, A.C. (2015). Influenza A Virus Coinfection through Transmission Can Support High Levels of Reassortment. J. Virol. *89*, 8453–8461.

Thompson, C.I., Barclay, W.S., Zambon, M.C., and Pickles, R.J. (2006). Infection of human airway epithelium by human and avian strains of influenza a virus. J. Virol. *80*, 8060–8068.

Varble, A., Albrecht, R.A., Backes, S., Crumiller, M., Bouvier, N.M., Sachs, D., García-Sastre, A., and tenOever, B.R. (2014). Influenza A virus transmission bottlenecks are defined by infection route and recipient host. Cell Host Microbe *16*, 691–700.

Varghese, J.N., McKimm-Breschkin, J.L., Caldwell, J.B., Kortt, A.A., and Colman, P.M. (1992). The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor. Proteins Struct. Funct. Genet. *14*, 327–332.

Vasilijevic, J., Zamarreño, N., Oliveros, J.C., Rodriguez-Frandsen, A., Gómez, G., Rodriguez, G., Pérez-Ruiz, M., Rey, S., Barba, I., Pozo, F., et al. (2017). Reduced accumulation of defective viral genomes contributes to severe outcome in influenza virus infected patients. PLOS Pathog. *13*, e1006650.

Vigil, K.J., Adachi, J.A., and Chemaly, R.F. (2010). Viral pneumonias in immunocompromised adult hosts. J. Intensive Care Med. *25*, 307–326.

Vignuzzi, M., Stone, J.K., Arnold, J.J., Cameron, C.E., and Andino, R. (2006).

Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature *439*, 344–348.

van der Vries, E., Schutten, M., Fraaij, P., Boucher, C., and Osterhaus, A. (2013a). Chapter Six – Influenza Virus Resistance to Antiviral Therapy. In Advances in Pharmacology, pp. 217–246.

van der Vries, E., Stittelaar, K.J., van Amerongen, G., Veldhuis Kroeze, E.J.B., de Waal, L., Fraaij, P.L.A., Meesters, R.J., Luider, T.M., van der Nagel, B., Koch, B., et al. (2013b). Prolonged influenza virus shedding and emergence of antiviral resistance in immunocompromised patients and ferrets. PLoS Pathog. *9*, e1003343.

Wagner, R., Matrosovich, M., and Klenk, H.D. (2002). Functional balance between haemagglutinin and neuraminidase in influenza virus infections. Rev. Med. Virol. *12*, 159–166.

Weis, W.I., Brünger, A.T., Skehel, J.J., and Wiley, D.C. (1990). Refinement of the influenza virus hemagglutinin by simulated annealing. J. Mol. Biol. *212*, 737–761.

Wilke, C.O., Reissig, D.D., and Novella, I.S. (2004). REPLICATION AT PERIODICALLY CHANGING MULTIPLICITY OF INFECTION PROMOTES STABLE COEXISTENCE OF COMPETING VIRAL POPULATIONS. Evolution (N. Y). *58*, 900–905.

Xiao, Y., Dolan, P.T., Goldstein, E.F., Li, M., Farkov, M., Brodsky, L., and Andino, R. (2017). Poliovirus intrahost evolution is required to overcome tissue-specific innate immune responses. Nat. Commun. *8*, 375.

Xu, R., de Vries, R.P., Zhu, X., Nycholat, C.M., McBride, R., Yu, W., Paulson, J.C., and Wilson, I.A. (2013). Preferential recognition of avian-like receptors in human influenza A H7N9 viruses. Science *342*, 1230–1235.

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

Xue, K.S., Hooper, K.A., Ollodart, A.R., Dingens, A.S., and Bloom, J.D. (2016). Cooperation between distinct viral variants promotes growth of h3n2 influenza in cell culture. Elife *5*.

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. (2018a). Within-Host Evolution of Human Influenza Virus. Trends Microbiol.

Xue, K.S., Greninger, A.L., Pérez-Osorio, A., and Bloom, J.D. (2018b). Cooperating H3N2 Influenza Virus Variants Are Not Detectable in Primary Clinical Samples. MSphere *3*, e00552-17.

Xue, K.S., Greninger, A.L., Pérez-Osorio, A., Bloom, J.D., Koelle, K., and Neher, R. (2018c). Cooperating H3N2 Influenza Virus Variants Are Not Detectable in Primary Clinical Samples. MSphere.

Yang, X., Steukers, L., Forier, K., Xiong, R., Braeckmans, K., Van Reeth, K., and Nauwynck, H. (2014). A beneficiary role for neuraminidase in influenza virus penetration through the respiratory mucus. PLoS One.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. *8*, 28–36.

Zanini, F., Brodin, J., Thebo, L., Lanz, C., Bratt, G., Albert, J., and Neher, R.A. (2015).

Population genomics of intrapatient HIV-1 evolution. Elife *4*, e11282.

Zanini, F., Brodin, J., Albert, J., and Neher, R.A. (2017). Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing. Virus Res. *239*, 106–114.

Zhou, B., Donnelly, M.E., Scholes, D.T., St George, K., Hatta, M., Kawaoka, Y., and Wentworth, D.E. (2009). Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. J. Virol. *83*, 10309–10313.

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.

Zhu, X., McBride, R., Nycholat, C.M., Yu, W., Paulson, J.C., and Wilson, I. a. (2012). Influenza virus neuraminidases with reduced enzymatic activity that avidly bind sialic acid receptors. J. Virol. *86*, 13371–13383.

APPENDIX A. WITHIN-HOST DIVERSITY OF INFLUENZA VIRUSES UNDER NEUTRAL EVOLUTION.

How much genetic diversity is expected to arise as influenza viruses replicate within human hosts? In evolutionary biology, it can be useful to estimate what variation would be observed if all mutations were purely neutral. Simple frameworks that model neutral evolution can establish basic expectations, even though purifying and positive selection clearly affect the mutation frequencies observed in real infections.

In human hosts, influenza virus populations expand exponentially at the beginning of an acute infection. Viral titers peak two to four days after the infection's start, and afterwards, titers decline for three or four days until the virus reaches undetectable levels (Beauchemin and Handel, 2011; Carrat et al., 2008). Mutations that arise early in the exponential expansion can reach high frequencies through Luria-Delbruck dynamics (Luria and Delbrück, 1943).

To estimate how many mutations are expected to reach detectable frequencies under neutral evolution, we use the stochastic birth-death model proposed by Bozic et al. to describe how neutral mutations accumulate as cells expand clonally during cancer evolution (Bozic et al., 2016). In this model, a viral population begins with a single starting genotype, although natural human infections begin with anywhere from one to several hundred initial genotypes (McCrone et al., 2017; Poon et al., 2016). Viruses reproduce at rate b and leave the population at rate d. Neutral mutations occur at a rate of u mutations per genome per replication cycle, and all sites are completely linked. Bozic et al. demonstrate that the expected number of mutations m above frequency α is

$$m = \frac{u(1-\alpha)}{(1-\frac{d}{b})\alpha} \quad (0.1)$$

We estimate *b* and *d* using Beauchemin's and Handel's models of influenza-virus kinetics within human hosts (Beauchemin and Handel, 2011). If influenza viruses expand exponentially with rate *b*-*d* for the first phase of the infection and then decline exponentially with rate *d* after viral titers peak, then we estimate $b \approx 5.7$/day and $d \approx 3.2$/day. Most studies in cell culture estimate mutation rates ranging from 10^{-6} to 10^{-5}/site/generation depending on the type of mutation and exact method of estimation (Bloom, 2014; Nobusawa and Sato, 2006; Sanjuán et al., 2010; Suarez-Lopez and Ortin, 1994), although one recent study estimates a higher rate of 10^{-4}/site/generation (Pauly et al., 2017). These per-site mutation rates correspond to $u \approx 0.013$ to $u \approx 1.3$ across the 1.3kB viral genome. Since the number of expected mutations is directly proportional to the viral mutation rate, this variation has a large effect on estimates of genetic diversity (**Figure A1**).

Future work that refines estimates of mutation rate would help establish more confident expectations about within-host viral diversity. It will also be important to develop models with more realistic assumptions about initial within-host genetic diversity, as well as how purifying and positive selection would affect this variation. By comparing these models with empirical observations, we can improve our understanding of how influenza viruses evolve within human hosts.

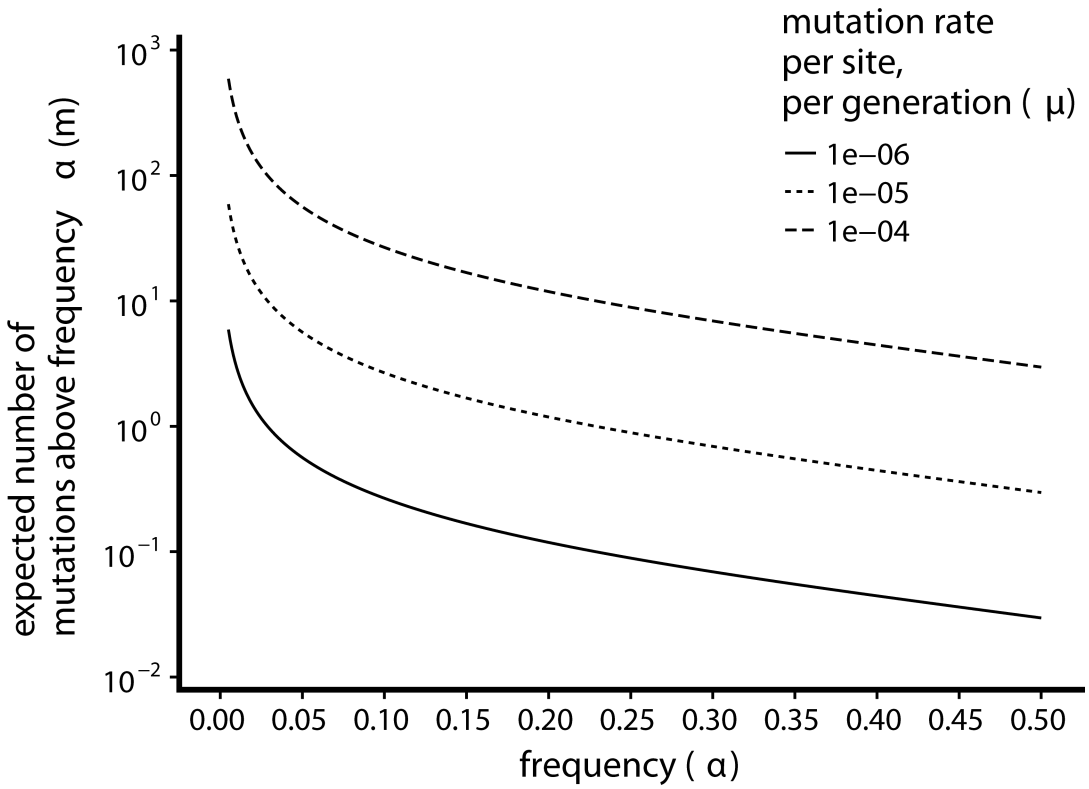

Figure A1. Expected number of within-host variants *m* above a given variant frequency *α* under neutral evolution.

Expectations are displayed for different values of the per-site, per-generation mutation rate *μ*, which is multiplied by the number of base pairs in the genome of influenza virus to obtain the per-genome mutation rate *u*.

VITA

Katherine Xue grew up in Knoxville, Tennessee. She received her A.B. in Chemical and Physical Biology *summa cum laude* from Harvard University in 2013. As an undergraduate, she worked in the lab of Kirsten Bomblies and studied molecular mechanisms of adaptation to whole-genome duplication in *Arabidopsis arenosa.* She started her PhD work in Genome Sciences at the University of Washington in 2014. Her graduate work with Jesse Bloom and Joshua Akey focuses on tracking the evolutionary dynamics of influenza viruses in cell culture and within human hosts. She plans to pursue postdoctoral work studying evolutionary dynamics in microbial communities with David Relman and Dmitri Petrov at Stanford University.