Quantifying progeny production from individual influenza virus-infected cells

David Bacsik

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

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Jesse D. Bloom, Chair

Adam Geballe

Gavin Ha

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Quantifying progeny production from individual
influenza virus-infected cells

David Bacsik

Chair of the Supervisory Committee:
Affiliate Professor Jesse D. Bloom
Genome Sciences

The distribution of progeny virions produced by virus-infected cells is extremely heterogeneous. This trend has been observed in diverse viruses, including bacteriaphage and pathogenic human viruses. To date, it has been difficult to explain why some infected cells generate thousands of progeny virions while others – infected under identical conditions – produce no progeny.

Established methods for quantifying progeny from single cells rely on the isolation of individual cells during infection. Such methods are not compatible with contemporary single-cell assays. With limited information gathered about the host and virus processes that occur during infection, the factors that might influence progeny production at the single-cell level have remained largely inaccessible.

I have developed new methods to quantify the amount of progeny produced by single influenza virus-infected cells; these methods do not require single-cell isolation during infection. Applying these methods, I have simultaneously measured viral transcription, viral genotype, and progeny virion production in the same influenza-infected cells. The correlation between viral transcription and progeny production is surprisingly poor at an early time of influenza infection. Using the viral gene expression information provided by single-cell RNA sequencing, I learned that cells with extremely high viral transcription often lack the influenza non-structural (NS) gene, precluding them from contributing infectious progeny.

While this system was developed to study influenza virus progeny production in single cells, individually-traceable virions may be useful in several areas of virology. The general approach to generate highly-diverse libraries of barcoded virions could likely be applied to other viruses with established reverse genetics systems. In the course of developing these methods, further opportunities for optimization have become apparent. I have outlined potential future applications that could be facilitated by either the current virus libraries or more diverse libraries that are developed in the future.

TABLE OF CONTENTS

Page

LIST OF FIGURES

ACKNOWLEDGMENTS

Thank you to Jesse Bloom. Jesse's close mentorship – particularly in the early years of my PhD – helped me grow immensely as a scientist. The hours we spent designing and revising experiments provided me with a thorough education in virology, genetics, and molecular evolution. I am incredibly grateful for the time we spent together. Jesse has brought together a wonderful group of scientists in the Bloom Lab, and I have been fortunate to work with and learn from this group. I leave the lab with many valuable models for conducting creative and carefully reasoned experiments.

I would like to thank my supervisory committee, Adam Geballe, Kelley Harris, Stephen Tapscott, Cole Trapnell, and Gavin Ha. Thanks to Adam for valuable virology feedback throughout the project, and to Cole for consistent enthusiasm from inception to completion. Special thanks to Adam, Jesse, and Gavin for serving on the reading committee for this dissertation. Additionally, I would like to thank Michael Emerman and the members of the Thursday Morning Virus Meeting for their quality annual discussions of this work.

I have learned that I prefer conducting research in partnership more than executing experiments alone. I would like to thank Bernadeta Dadonaite for her thoughtful guidance on virology questions and for her work sequencing viral genomes for this project. I would also like to thank Andrew Butler for his input, which resolved many lingering questions about how to best analyze the data I had generated. His enthusiasm in carrying this system forward into new applications was particularly gratifying. Thank you to Jason Underwood for sharing his wealth of molecular biology knowledge with me. Thank you to Andrea Loess for brainstorming and troubleshooting throughout my PhD, and for her collaboration on recent projects.

Thank you to Pete Degen, Clara Bien Peek, Amy Clore, and Lauren Fishbein for their mentorship before graduate school that prepared me to begin this research.

For many years, it was a daily joy to see Kate Crawford, Sarah Hilton, and Adam Dingens sitting at the three desks nearest to mine. Our time together in the lab was one of the best parts of this experience. I'm very glad to have their continued friendship. Thanks as well to my friends from medical and graduate school; I'm fortunate that there are too many to list by name.

Thank you to my parents, Edward John Bacsik and Michele Keane Bacsik. They have always encouraged my curiosity, and they continue to give me their unwavering support. Thanks to my siblings for making me feel like a large team is always looking out for me. Finally, thank you to my partner Emily Johnson, who has taught me much about the value of making small progress on hard problems.

Chapter 1

INTRODUCTION

The outcomes of viral infection vary widely from person to person, and even from cell to cell within the same individual. To date, the factors that might influence cell-to-cell variation have been difficult to assess. Recently, advances in single-cell assays have provided unprecedented access to gene expression information within single cells. My graduate work has focused on developing methods that make single-cell assays more useful for virology applications. In this dissertation, I present scalable techniques to quantify the amount of new virions generated by single influenza virus-infected cells. This is an important metric for understanding an infection because progeny virions propogate an infection within a host, and sometimes transmit an infection to other individuals. To demonstrate the utility of this approach, I describe the distribution of progeny produced by single cells infected with pdmH1N1 influenza virus, and use viral gene expression measurements to explain the outcome of some cells that transcribe high levels of viral mRNA but fail to generate progeny virions.

In this first chapter, I introduce relevant background about influenza virus replication, single-cell RNA sequencing, and the heterogeneity observed in viral infection across single cells. In Chapter 2, I describe techniques I developed to generate diverse barcoded libraries of influenza virus. I use these libraries to make simultaneous measurements of viral transcription and progeny production from single infected cells. I learn that influenza gene expression and progeny production correlate poorly in single cells, and that absence of the influenza non-structural (NS) gene explains the outcomes of cells with the highest levels of viral transcription. In Chapter 3, I describe potential avenues to improve the generation of diverse barcoded virus libraries and single-cell sequencing of virus-infected cells. I outline open questions that could be addressed using barcoded virus libraries. Finally, in Chapter 4, I summarize the major themes from each section.

1.1 Heterogeneity across single virus-infected cells

The outcomes of viral infection are extremely heterogeneous when measured in single infected cells [18, 74, 77, 33]. Bacteriaphage provided an early model system amenable to quantitative experiments [12]. To make measurements of progeny production from many individual bacteria, novel methods were required. In 1945, Delbruck developed a method of isolating individual phage-infected bacteria a scale sufficient to describe the distribution of progeny production in a bacterial culture [18]. He found that progeny production varied by two orders of magnitude across single infected cells.

The distribution of progeny production has now been characterized for several pathogenic human viruses [77, 33]. To date, all experiments have cultured infected cells in single-cell isolation. After infection, the supernatant surrounding a single cell is collected and a plaque assay is used to quantify the number of infectious progeny produced. Using this approach, poliovirus progeny production was found to vary over three orders of magnitude: some infected cells produced no detectable progeny, while others produced thousands of progeny virions [77]. Surprisingly, the distribution's variance was not influenced by the initial multiplicity of infection. Influenza virus progeny production has also been shown to vary by three orders of magnitude across individual infected cells [33]. Measurements of influenza progeny production were made in the context of a high multiplicity of infection, where every cell is likely to be infected by multiple virions and a contain complete set of viral genes. Even under these favorable conditions, nearly half of all influenza-infected cells produce no detectable progeny. The most productive infected cells contribute approximately one thousand infectious progeny each. The distribution is highly skewed, with a bias towards low progeny production.

Recent advances in single-cell RNA sequencing have made it possible to assay the expression of all host and viral genes in a population of single virus-infected cells. Single-cell studies of viral transcription have been performed for human pathogens like influenza virus [74, 83, 14], zika virus [101], and SARS-CoV-2 [67, 81]. The transcriptional state of infected cells might explain some of the differences between cells that make many progeny virions and cells that make none. However, standard single-cell RNA sequencing methods do not

provide information about the number of progeny generated by each infected cell. Likewise, standard plaque assays do not provide single-cell resolution. In order to understand the relationship between viral transcription and progeny production in single cells, new methods were required to link these measurements in the same infected cells.

1.2 Influenza virus replication cycle

For an individual influenza virion to contribute progeny to the virus population, it must undergo replication without fatal error. Influenza virus replication begins when a virion binds to the surface of a cell. The influenza viral entry protein haemagglutinin (HA) binds to sialic acid on the surface of mammalian or avian cells [96]. After binding, the influenza virion is endocytosed and trafficked to the host endosome [44]. The acidic environment of the endosome triggers a conformational change in HA which results in fusion of the viral and endosomal membranes [47]. The vRNA genomic segments arrive packaged with essential viral proteins in viral ribonucleoproteins (vRNPs) [22]. Upon fusion, the vRNPs are released into the cytoplasm and trafficked to the nucleus [22].

Viral genome replication and viral transcription are performed by the same viral polymerase complex. Transcription of viral mRNAs is templated directly by the vRNA genomic segments [93]. Transcribed viral mRNA molecules mimic host mRNAs and are translated into proteins using host ribosomes via standard translation pathways [100].

Replication of the vRNA occurs in the nucleus and proceeds through an intermediate species, complementary RNA (cRNA), which is a full-length antisense copy of the vRNA sequence [31]. Newly synthesized vRNA molecules are packaged with nucleoprotein (NP) and a viral polymerase complex into a vRNP [54]. vRNPs are transported out of the nucleus by nuclear export protein (NEP), encoded by a splice isoform of the non-structural (NS) gene [60].

Assembly of progeny virions occurs at the plasma membrane [19]. One vRNP containing each genome segment is typically packaged into a nascent virion. The process by which genome packaging is controlled is multi-factorial and incompletely understood [19]. Unique RNA sequences at the ends of each viral segment encode critical packaging signals which are necessary for incorporation into progeny virions and confer segment identity [27].

In the final step of virion assembly, the plasma membrane containing the vRNPs and viral proteins folds in upon itself, encapsulating the viral contents in a process known as budding [72]. Newly-generated virions remain attached to the cell surface by interactions between HA and sialic acid. The influenza virus protein neuraminidase (NA) cleaves the sialic acid to release the virion from the cell [7]. The newly-generated progeny virion is then liberated into the surrounding environment.

1.3 Influenza genome replication

Influenza virus has a segmented RNA genome, which is replicated by a virus-encoded RNA-dependent RNA polymerase complex [54]. This polymerase complex has a high rate of error [**?**]. Disparate estimates of the inherent error rate of the polymerase have been measured using different techniques, but an approximate "rule of thumb" is that one nucleotide substitution likely occurs in each cycle of genome replication [82, 57]. The result of this high error rate is that influenza virus propagating through a cell culture or tissue accumulates low-frequency variants as replication errors compound [6].

Other genomic abnormalities are commonly found in populations of influenza virions, as well. Virions carrying fewer than eight genome segments have been documented by electron microscopy [30]. Cells infected with influenza virus often fail to express one or more viral genes[74, 85] or proteins [10]. Additionally, large internal deletions are commonly detected by sequencing in both *in vitro* influenza infections [17] and acute human infections [75]. Large internal deletions are deletions that remove most of the coding sequence of a viral gene, while retaining the packaging signals necessary for incorporation into progeny virions. Virions carrying these species are sometimes referred to as "defective interfering particles", since they are preferentially replicated and propagated in the presence of complementary fully-functional virions [17].

1.4 Single-cell RNA sequencing

Single-cell RNA sequencing provides gene expression measurements for individual cells. In this dissertation, a microfluidics device was used to isolate single cells in water-in-oil droplets [102]. Alternative methods of single-cell RNA sequencing isolate single cells in

tissue-culture wells [89] or utilize combinatorial indexing[13], which precludes the need for physical isolation. Regardless of the method employed, cDNA is ultimately generated in a way that encodes a unique barcode sequence for each cell in a sample.

In droplet-based methods, cells are co-encapsulated in a droplet with a primer-coated bead [102]. The primers coating the bead target the polyA tract at the 3' end of mRNA molecules. Influenza virus mRNA transcripts are polyadenylated and are efficiently captured by these primers. Influenza virus vRNA and cRNA species are not polyadenylated, and are not captured by these primers. The primers also contain a unique cell barcode that is embedded in the cDNA transcript. Once the cells and primers are co-encapsulated, reverse transcription is performed.

The emulsions are broken and the cDNA library is recovered and amplified by PCR. The full-length cDNA library is enzymatically fragmented, and a priming sequence is ligated to the fragmented molecules [102]. Another PCR reaction is used to amplify the transcriptome and append sequencing adapters. Fragments derived from the 3' end of the cDNA transcript contain the appropriate priming sequences at each end of the molecule and are amplified; fragments derived from other portions of the cDNA transcript contain zero or one priming sequence, and are not amplified efficiently. High-throughput short-read sequencing is used to read the sequence of each transcript fragment. Alignment to a reference genome provides the identity of each transcript and parsing of the cell barcode links it to a specific single cell [102].

Chapter 2

INFLUENZA VIRUS TRANSCRIPTION AND PROGENY PRODUCTION ARE POORLY CORRELATED IN SINGLE CELLS

A version of this chapter has previously been posted as a pre-print:

David J. Bacsik, Bernadeta Dadonaite, Andrew Butler, Allison J. Greaney, Nicholas S. Heaton, Jesse D. Bloom. Influenza virus transcription and progeny production are poorly correlated in single cells. *bioRxiv* DOI:10.1101/2022.08.30.505828

2.1 Abstract

The ultimate success of a viral infection at the cellular level is determined by the number of progeny virions produced. However, most single-cell studies of infection quantify the expression of viral transcripts and proteins, rather than the amount of progeny virions released from infected cells. Here we overcome this limitation by simultaneously measuring transcription and progeny production from single influenza-virus-infected cells by embedding nucleotide barcodes in the viral genome. We find that viral transcription and progeny production are poorly correlated in single cells. The cells that transcribe the most viral mRNA do not produce any detectable progeny, and represent aberrant infections that fail to express the influenza NS gene. However, only some of the discrepancy between transcription and progeny production can be explained by viral gene absence or mutations: there is also a wide range of progeny production among cells infected by complete unmutated virions. Overall, our results show that viral transcription is a relatively poor predictor of an infected cell's contribution to the progeny population.

2.2 Introduction

Many aspects of viral infection are extremely heterogeneous when measured across single cells. Individual infected cells vary widely in transcription of viral genes [74, 85, 101],

presence of viral mutations [73], expression of viral proteins [10], replication of viral genomes [77], and production of viral progeny [77, 18, 33]. However, it is unclear how variation in these different aspects of infection are related within the same infected cells. For instance, to what degree does the extent of viral transcription in an infected cell determine the number of progeny virions the cell produces? The answer to this question remains elusive because the most common single-cell techniques (flow cytometry and single-cell RNA sequencing) measure the levels of proteins and transcripts, rather than the number of viral progeny produced.

Here, we develop a novel approach to simultaneously measure viral transcription, viral mutations, and viral progeny production in single cells infected with influenza virus. We find that progeny production is even more heterogeneous than viral transcription in single cells. The cells that express the most viral transcripts usually do not generate any detectable viral progeny. Instead, cells with extremely high viral transcription often fail to express the NS gene and represent non-productive infections. Our findings emphasize that different aspects of viral infection are not always correlated at the single cell level, and that many of the cells contributing large amounts of viral mRNA to bulk RNA sequencing studies do not appreciably contribute virions to the progeny population.

2.3 Results

2.3.1 Viral barcoding to measure transcription, progeny production, and viral genotype in single cells

To quantify the progeny virions released from single infected cells, we inserted random nucleotide barcodes [45, 5, 91] into the influenza virus genome so that they are positioned near the 3' end of the viral mRNAs (Fig 2.1A). When cells are infected at a low multiplicity of infection (MOI), cells will usually be infected with no more than one barcoded virion. Standard 3'-end single-cell sequencing of the mRNA in infected cells [74, 85, 73, 94, 83, 14] captures the viral barcode sequence along with host and viral transcripts, enabling determination of which barcoded virion infected each cell (Fig 2.1A). We can sequence the viral barcodes on progeny virions released into the supernatant to quantify the relative

number of physical progeny produced by each cell, and sequence the viral barcodes in cells secondarily infected with an aliquot of the supernatant to quantify the relative number of infectious progeny produced by each cell. Additionally, we can reconstruct the genome of the virion that infected each cell by selectively amplifying viral genes from the single-cell cDNA library and performing long-read sequencing as described previously [73]. This strategy enables simultaneous measurement of transcription, progeny production, and viral genotype in single cells.

2.3.2 Creation of dual-barcoded virus library

To insert barcodes into influenza virus genes, we used a previously described approach to duplicate the packaging signals of the HA and NA genes to create sites where exogenous sequence can be added without disrupting viral genome packaging [32, 24]. This approach allowed us to insert 16-nucleotide random barcodes near the 3' end of the genes, downstream of the stop codon but upstream of the polyadenylation signal (Fig 2.1B). These barcodes are therefore present in both viral mRNAs and genomic RNAs (vRNAs), but do not modify the amino acid sequence of the viral protein. We will refer to these viruses as "dual barcoded" as they have barcodes on two different genes.

We engineered barcodes into the A/California/04/2009 (pdmH1N1) strain of influenza virus with the G155E cell-culture adaptation mutation [15]. Viruses with barcoded HA and NA segments could be generated by reverse genetics, and in cell culture grew to titers comparable to unmodified viruses (Fig 2.1C). To confirm that the sequence of individual barcodes did not affect viral growth in cell culture, we generated virus libraries carrying a small pool of barcodes and verified that the virus titers and barcode frequencies were stable across three passages (Fig A1).

For our single-cell experiments, we generated libraries of virions with a high diversity of barcodes on the HA and NA genes. Our experiments involved infecting 10,000 cells at a MOI of 0.15, so it is important that the barcoded virus libraries be sufficiently diverse that nearly every virion will have a unique barcode in a random sample of 1,500 virions. We used deep sequencing to verify that for both HA and NA, in a sample of 1,500 barcodes

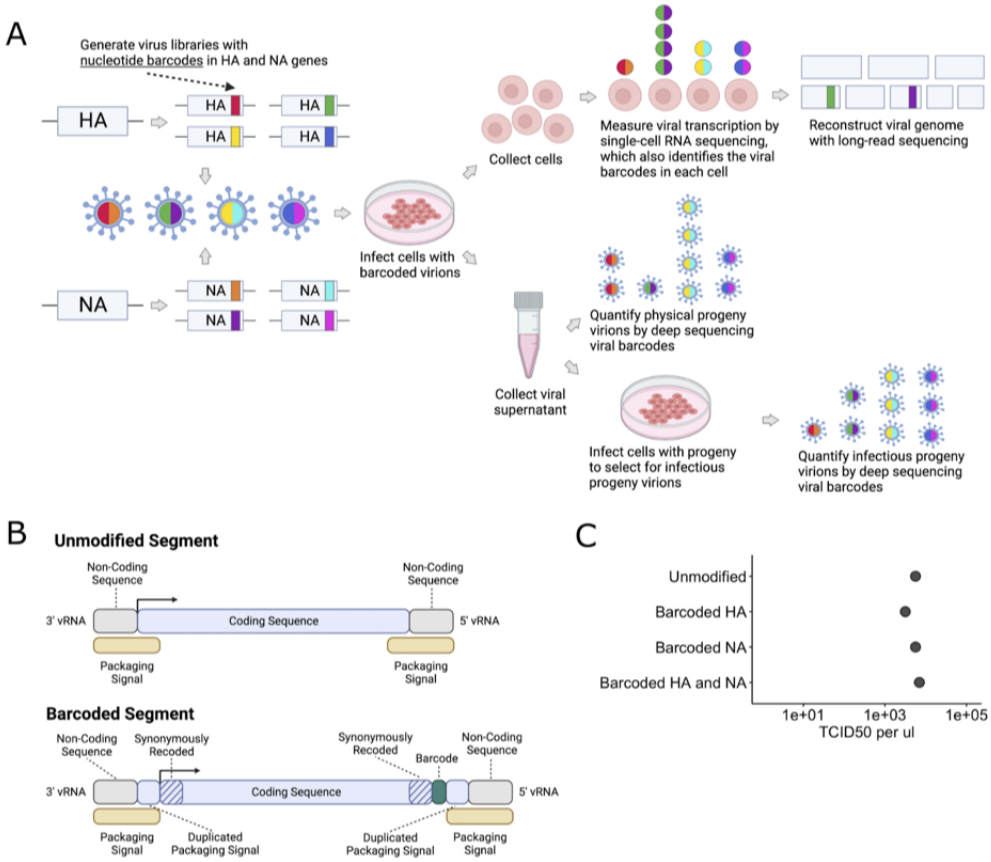


Figure 2.1: Strategy to measure transcription, progeny production, and viral genotype in single cells. (A) Insertion of barcodes in the viral genome makes it possible to quantify the progeny released from single cells, and relate progeny production to viral transcription and viral genotype. (B) Barcodes were inserted near the 3' end of the mRNA sequence between the stop codon and the polyA site, using a duplicated packaging signal scheme to avoid disrupting viral genome packaging. (C) Viruses with one or two barcoded segments grew to similar titers as viruses with unmodified genomes. The titers shown were measured after generating the viruses by transfection.

from our virus library, >96% of barcodes were unique (Fig A2).

2.3.3 We recapitulate prior findings that viral transcription is extremely heterogeneous across single infected cells

We implemented the experiment in Fig 2.1A by infecting approximately 10^4 MDCK-SIAT1-TMPRSS2 cells [46] with the dual-barcoded virus library at low MOI of 0.15 to ensure most cells were infected by at most one virion. To ensure a single round of relatively synchronized infection, we replaced the virus inoculum with fresh medium after one hour and added ammonium chloride, which prevents secondary infection by blocking the endosomal acidification necessary for viral fusion [50, 59]. We collected the cells for single-cell RNA sequencing on a 10X Chromium device at 12 hours post-infection. We then added a control sample to quantify the rate of cell multiplets and PCR strand exchange: this control sample consisted of cells infected with a virus carrying synonymous mutations that can be distinguished by sequencing (see Methods).

We obtained single-cell RNA sequencing data for 254 cells infected with our barcoded virus library, resulting in an empirical MOI of 0.14 that closely corresponds to our target MOI of 0.15. The reason we captured only 254 infected cells after infecting approximately 10^4 cells at a MOI of 0.15 is because many cells are lost during library preparation before loading into the 10X Chromium device [102, 98].

Among the infected cells, there was extremely wide variation in the amount of viral transcription (Fig 2.2A), similar to that observed in prior single-cell studies of influenza infection [74, 85, 73, 94, 83]. In most infected cells, viral transcripts accounted for ¡10% of all transcripts, but in a handful of cells over half of the transcripts were derived from virus (Fig 2.2A). A substantial fraction (40%) of the infected cells also failed to express all eight viral genes (Fig 2.2B), a phenomenon that has been extensively described in prior studies [74, 85, 10]. At a per gene level, each viral gene was not expressed in at least some cells, with HA and NS exhibiting the highest rates of absence (Fig 2.2C). Note that our ability to determine whether a viral gene is absent depends on the total level of viral transcription in a cell (Fig A3 and Methods), which could reduce our ability to detect the absence of the

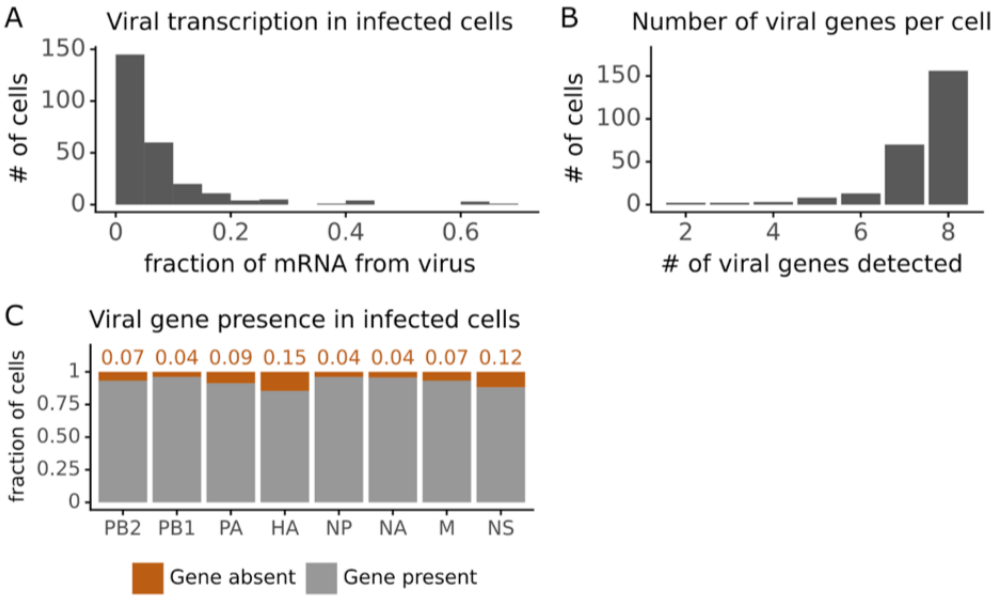


Figure 2.2: Viral transcription is extremely heterogeneous across single infected cells, and some cells fail to express some viral genes. This plot shows single-cell RNA-sequencing data for the 254 cells that were infected. (A) Viral transcription in infected cells is extremely heterogeneous, with viral mRNA composing <1% of total mRNA in some cells, but >50% in others. (B) The number of viral genes detected in each infected cell. More than half of infected cells express mRNA from all 8 viral segments. (C) The fraction of infected cells expressing each viral gene.

four viral genes involved in transcription (PB2, PB1, PA, NP).

2.3.4 Full genome sequences of the virions infecting individual cells

Virions can be defective in two ways: they can fail to express a viral gene, or they can encode mutated viral proteins. To identify cells infected by mutated virions, we used long-read PacBio sequencing of the viral transcripts to reconstruct the consensus genome of virions infecting single cells [73]. Because each transcript in the single-cell RNA sequencing library carries a cell barcode, we could link the sequence of each viral transcript to the cell

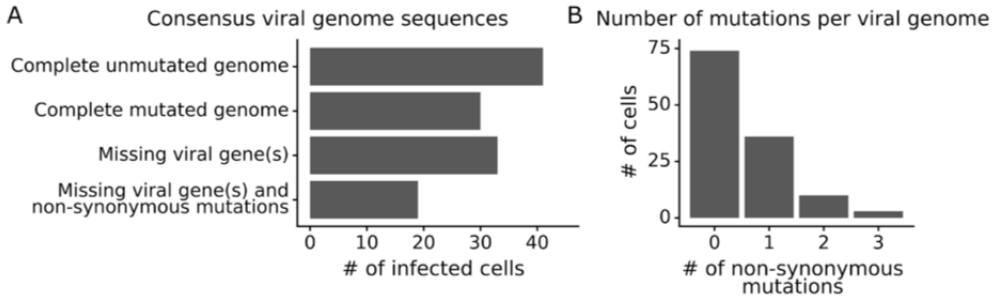


Figure 2.3: Statistics on consensus viral genome sequences from single infected cells. (A) The number of infected cells expressing all eight viral genes without non-synonymous mutations, expressing all eight viral genes with one or more non-synonymous mutation(s), missing one or more viral gene(s), or with both mutated and missing genes. (B) The number of non-synonymous mutations in each viral genome. Deletions are classified as a non-synonymous mutation for these counts. This plot shows only the 123 of 254 single infected cells for which we could determine the sequence of all genes expressed by the infecting virion. See Fig A4A for details on properties of infected cells for which we could obtain full viral sequences, and Fig A5 for the full set of viral mutations in each infected cell.

that produced it.

We obtained complete sequences of all expressed viral genes for 123 of the 254 infected cells in our dataset (Fig A4A). About a third of the infected cells expressed all eight viral genes without any non-synonymous mutations (Fig 2.3A, Fig A5). The remainder of infected cells failed to express a viral gene, expressed a gene with a non-synonymous mutation, or both (Fig 2.3A). Mutated virions most commonly had just one non-synonymous mutation in their genome, but some virions had two or three mutations (Fig 2.3B, Fig A5). Note that some virions had large internal deletions in a gene (Fig A5) as has been previously described [75, 17]; here we have classified deletions as non-synonymous mutations.

2.3.5 Progeny production from single infected cells is more heterogeneous than viral transcription

We measured the amount of physical and infectious progeny virions produced by single infected cells. We quantified physical progeny virions by sequencing viral barcodes from vRNA molecules in the supernatant at 12 hours post infection (Fig 2.1A). We quantified infectious progeny virions by infecting another set of cells with some of the viral supernatant and sequencing barcodes from viral RNA expressed in these newly infected cells (Fig 1A). We analyzed these progeny production measurements for infected cells that met these criteria: both barcoded genes are expressed (allowing us to identify both viral barcodes), and the complete sequences of all expressed viral genes were obtained (Fig A4). We analyzed the progeny contributions of the 92 infected cells with complete progeny measurements, transcriptomes, and viral genomes, and describe their relative contributions to the progeny population. Progeny contributions range from 0-100% of the defined population.

The amount of progeny virions produced per cell was extremely heterogeneous (Fig 2.4A,B). Nearly half of infected cells failed to produce any detectable physical or infectious progeny, and only a few cells produced high levels of progeny (i.e. >10% of all progeny generated by the 92 infected cells analyzed) (Fig 2.4, S6). Progeny production is more heterogeneous than viral transcription across single cells. We quantify this heterogeneity by calculating a Gini coefficient, which can range from zero to one with larger values indicating more uneven distributions [26]. The Gini coefficients are 0.78 and 0.88 for physical and infectious progeny production versus 0.46 for viral transcription (Fig 2.4). Just 6 of the 92 infected cells generated over half the physical progeny—whereas transcription is much less skewed, with 23 cells required to account for over half the viral transcripts (Fig A6).

2.3.6 Cells with the most viral transcription produce no progeny, and represent aberrant infections that fail to express the NS gene

The correlation between viral transcription and progeny production in single cells is surprisingly poor (Fig 2.5A). None of the cells with >25% of their mRNA transcripts derived from virus produced any detectable progeny (Fig 2.5A). Instead, the progeny come from

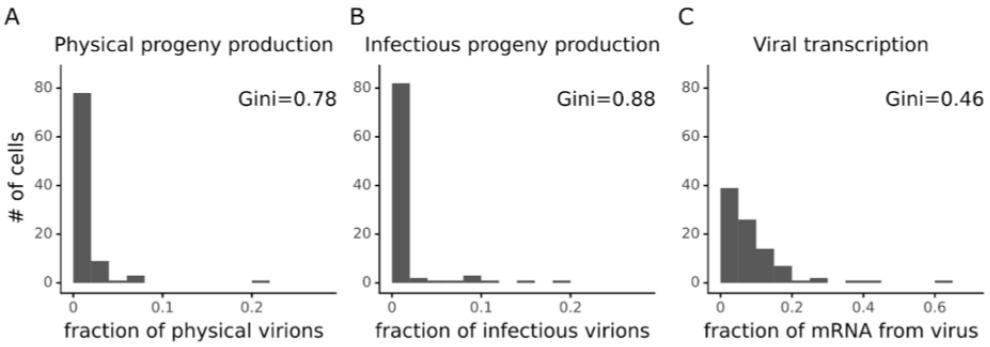


Figure 2.4: Viral progeny production is even more heterogeneous than viral transcription across single infected cells. Heterogeneity across single infected cells in (A) physical progeny production, (B) infectious progeny production, and (C) viral transcription. The Gini coefficient [26] quantifying the extent of cell-to-cell variability is indicated on each panel; a larger Gini coefficient indicates a more uneven distribution. For (A) and (B) the x-axis is the fraction of all barcodes associated with each cell among barcodes assignable to any infected cell; for (C) the x-axis is the fraction of mRNA in each cell that is derived from virus. This plot shows only the 92 of 254 single infected cells for which we could identify the viral barcode on both barcoded genes and determine the sequence of all genes expressed by the infecting virion.

cells where 1% to 20% of mRNA transcripts are derived from virus.

The viral genetic information associated with each cell offers a straightforward explanation for some of these trends. Cells that fail to express a viral gene produce little or no detectable progeny virions, presumably because lack of the encoded protein impairs virion formation (Fig 2.5A; note that our analysis is limited to cells that express HA and NA since those are the barcoded genes). But although cells that fail to express a viral gene produce little or no progeny, the converse is not true: cells that express the full complement of viral genes often still fail to produce progeny (Fig 2.5A). Physical viral progeny are produced both by cells that express viral genes with and without mutations, however, more of the infectious progeny virions come from cells with unmutated viral genomes (Fig 2.5A, Fig A7)—probably because some non-synonymous mutations interfere with protein functions required for infection of new cells. Nonetheless, physical and infectious progeny production are much more correlated among single cells than are transcription and progeny production (Fig 2.5A versus Fig A7; Pearson's R values of -0.14 for correlation of transcription with infectious progeny versus 0.39 for correlation of physical progeny with infectious progeny).

Strikingly, absence of the viral NS gene not only precludes progeny production but is associated with an aberrant state of high viral transcription (Fig 2.5A,B). Specifically, most of the highest transcribing cells fail to express NS, and the mean level of viral transcription is significantly higher ($p < 0.01$) in cells that do not express NS (Fig 2.5B). This observation is consistent with the known functional roles of the NEP protein expressed from the NS gene, which is to export viral ribonucleoprotein complexes from the nucleus [60, 11] and possibly act as a switch from transcription to genome replication [71, 56]. Overall, this result suggests that the cells that contribute the most to the signal observed in transcriptomic studies often represent aberrant non-productive infections that do not contribute viral progeny.

However, none of the viral genetic factors we measure (failure to express a gene or mutated viral proteins) fully explain the extremely wide variation in progeny production. Progeny production is quite variable even across cells expressing unmutated copies of all viral genes, although it is less variable than across all cells (Fig 2.5C,D). This fact suggests other cellular or unknown viral factors must also contribute to cell-to-cell variation in progeny production.

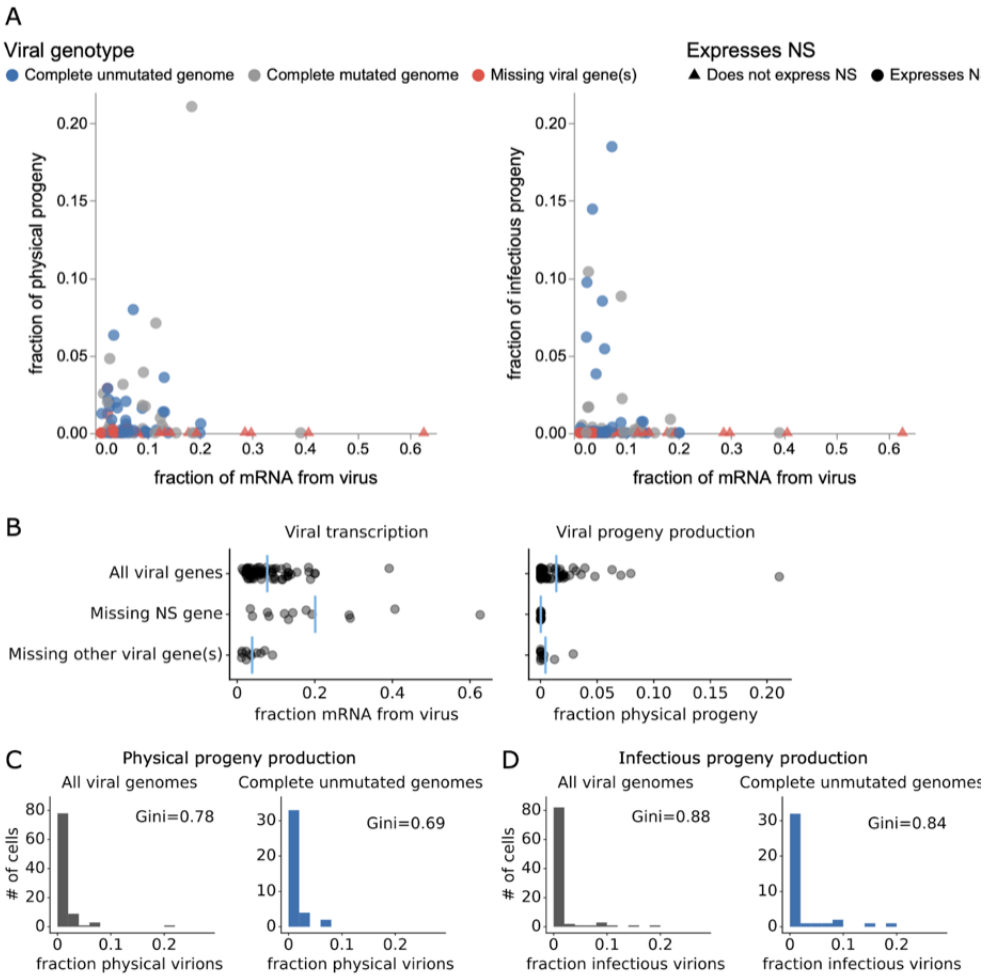


Figure 2.5: Relationship between viral transcription and progeny production in single infected cells. (A) Relationship between viral transcription and physical (left) or infectious (right) progeny virion production. Each point is a different cell, colored according to whether the cell expresses unmutated copies of all eight genes, all genes with one or more non-synonymous mutations, or less than all genes (with or without mutations). Point shapes indicate whether the viral NS gene is expressed. An interactive version of this figure that enables mouse-overs of points with details about individual cells is at https://jbloomlab.github.io/barcoded_flu_pdmH1N1. (B) Cells that do not express NS produce significantly more viral transcripts than other cells ($p<0.01$, Kruskal-Wallis test followed by a Dunn's test with Bonferroni correction). However, cells that do not express NS produce minimal physical progeny. The mean for each group is shown as a blue line. (C) Distribution of physical progeny virions produced by all infected cells (left) or only cells that express all eight viral genes without non-synonymous mutations (right). (D) Like panel (C), but for infectious rather than physical progeny. The plots showing all infected cells are duplicated from Fig 4 to facilitate direct comparison of all cells to those with complete unmutated genomes. This figure shows only the 92 infected cells for which we could identify the viral barcode on both barcoded genes and determine the sequence of all genes expressed by the infecting virion.

2.4 Discussion

In this work, we simultaneously measure several important characteristics of single influenza-virus infected cells. We sequence the genome of the infecting virion, measure viral transcription, and quantify the progeny produced by each infected cell. We observe a poor correlation between viral transcription and progeny production across single cells. Surprisingly, the cells that transcribe the most viral mRNA do not contribute any detectable progeny. Some of the discrepancy between transcription and progeny production can be explained using the viral gene expression and mutation information we captured. Cells that lack expression of the influenza NS gene tend to transcribe very high amounts of viral mRNA, near the extreme of the transcription distribution. However, like all cells that fail to express one or more influenza genes, these cells are not capable of producing viral progeny. Among cells that express all viral genes, contributions to the physical progeny population are similar between cells infected by virions with mutated and unmutated genomes. In contrast, the infectious progeny population has very few virions contributed by cells infected by virions carrying protein-coding mutations. These results have implications for the application of RNA sequencing in virology research and for the understanding of transmission bottlenecks in acute viral infections.

We recapitulate prior findings that viral transcription [74] and progeny production [33] are both highly heterogeneous across single cells when measured independently. However, when measured simultaneously in the same cells infected at low multiplicity of infection, the correlation between these two processes is remarkably poor. Infected cells with the most viral transcription do not contribute the most viral progeny. In fact, in this study, we detect no progeny virions from cells with the highest burden of viral transcripts. The majority of physical and infectious progeny both come from cells near the median of the viral transcription distribution. This result suggests that influenza-infected cells with unusually high viral mRNA content–which are often highlighted in single-cell studies of virus infection [74, 73, 83, 101] are not likely to be contributing progeny at a level commensurate with their extreme viral transcription. If this finding applies broadly, an important implication for virology research is that RNA sequencing, flow cytometry, and fluorescent microscopy studies

may be biased towards "dead-end" infections that express unusually large amounts of viral RNA or protein but cannot produce virions capable of propagating infection. Furthermore, this result raises the question of whether metrics other than progeny produced (e.g. viral transcription or viral protein expression) can serve as correlates of virion production from single cells. It may be the case that many factors influence the amount of progeny an infected cell produces, and multi-modal measurements are required to understand the outcomes of infection in single cells. Quantification of viral particles continues to be addressed best by direct measurement using plaque assays, TCID50 assays, or methods similar to the barcoding strategy we have introduced here.

Failure to express specific influenza virus genes can explain some of the discordance we observe between viral transcription and progeny production in single cells. Using single-cell RNA sequencing, the expression or absence of expression of each influenza gene can be detected in virus-infected cells. The presence and absence of viral genes provides a clear explanation for the outcomes of infection in cells with extremely high viral transcription. Under the conditions of infection tested here, cells at the extreme high end of viral transcription often lack expression of the influenza NS gene. Because each influenza virus gene is essential at some point in the virus' life cycle [19], it is not surprising that we do not detect progeny generated by these cells. However, it is notable that loss of this specific viral gene is associated with high viral transcription; this suggests a role for NS as a negative regulator of viral transcription. This fits with the canonical function of the influenza NS gene, which is to export vRNA templates from the nucleus for assembly into progeny virions [60, 11]. It stands to reason that loss of a vRNA segment from the nucleus terminates transcription from that template molecule. This role is also supported by several studies that explore the effect of influenza NS expression on viral transcription. These studies find that transcription by the influenza virus polymerase is downregulated when NS is expressed [70], and that the effect may be mediated by direct protein-protein interactions between NS and PB2 [56]. The specific discordance driven by loss of NS expression may not occur in a situation where every infected cell expresses a functional copy of the NS gene. Co-infection by virions generated in neighboring regions has been shown to occur commonly in lung tissue in in vivo animal studies of influenza virus infection [79, 39]. Genetic complementation,

where an additional virion provides a functional copy of a missing or mutated viral gene, is common, as well [78]. Under such conditions, it becomes much less likely that any viral gene will be missing from an infected cell. Studying progeny production from cells infected at a high MOI was not possible with the current virus library because the number of infecting virions with identical virus barcodes would be too great to confidently map progeny back to a single infected cell. However, in the future, a more diverse virus library could facilitate such studies by enabling experiments that utilize a greater number of individually traceable virions.

We reconstructed the genome of the virion that infected each cell using long-read virus genome sequencing [73]. This data provides information about viral mutations, which is not accessible with traditional single-cell gene expression profiling. Standard single-cell RNA sequencing techniques capture only a few hundred base pairs from the 3' end of transcripts, and generally do not parse sequence information beyond the identity of gene. For the set of cells with complete long-read sequencing information, we used a simple classification scheme to begin to understand how non-synonymous mutations that arise during viral replication affect progeny production from single cells. We found that non-synonymous mutations are depleted in the population of infectious progeny, relative to their representation in the population of all physical progeny. This result makes sense, given that most mutations to viral genes are deleterious to viral replication [76, 88, 20]. This study was not powered to characterize the effects of individual mutations to the viral genome. In the future, if many more infected cells were profiled and common mutations were observed multiple times, it may be possible to estimate the effects of specific mutations on progeny production in single cells.

Viral gene absence and mutations explain only a portion of the heterogeneity in viral transcription observed across single cells. Even among cells that express all viral genes without any protein-coding mutations (i.e. cells with a complete and umutated viral genome), progeny production remains highly heterogeneous. It is unclear to what extent the amount of progeny produced by a single infected cell is stochastic. It is possible that further characterization of single infected cells could provide explanations for these variable outcomes. Expression of host genes is a particularly rich area for investigation that was not addressed

in this study due to limited statistical power. Protein expression, protein modifications, and non-genetically encoded compounds (e.g. lipids and metabolites) offer other possible targets of study. However, it is possible that even with comprehensive profiling, progeny production varies between infected cells that are, for all intents and purposes, identical.

The persistent heterogeneity in progeny production observed in single cells has implications for understanding viral transmission events. The virions transmitted from donor to recipient during a transmission event are likely to come from a highly-productive subpopulation of infected cells. Recent work on the transmission of influenza virus [97, 53] and SARS-CoV-2 [9, 48] in humans has emphasized the narrow genetic bottleneck, with only a small fraction of all viral variants detected in the donor transmitted to the recipient. Acute influenza virus infections are generally initiated by just one or two unique viral genomes [53]. Physical constraints contribute to this narrow genetic bottleneck. Animal studies utilizing barcoded virus libraries have shown that the route of infection affects the number of viral particles reaching the donor. More particles are transmitted between animals that come into physical contact than between animals that share air, for example [91]. Our result demonstrating that viral progeny production is heterogeneous, even in a cell line with uniform infection conditions, suggests that the genetic bottleneck is narrowed beyond the physical bottleneck in part because the viral particles that are sampled in a transmission event are likely to be descended from the same highly productive cells in the donor, and therefore, share the mutations present in the virions that infected those cells.

We have demonstrated that diverse barcoded virus libraries can be used to trace individual virions through each step of the influenza virus replication cycle. The methods for library generation we present here have increased the complexity and scale of isogenic barcoded virus libraries beyond what has been previously published [91, 5]. By combining these novel barcoded virus libraries with single-cell RNA sequencing and long-read viral genome sequencing, we have begun to evaluate how viral transcription and progeny production are related within the same influenza virus-infected cells. We find that, on average, early viral transcription is a poor predictor of how many new virions an infected cell produces. Future work, building on the concepts of individually-traceable virions and progeny production measurements from single virus-infected cells, appears promising. The methods described

in this chapter could have applications in studies that address the effects of cell type on progeny production and in studies that make more precise measurements of the number of virions passing from donor to recipient in virus transmission events.

Acknowledgements

Thanks to Jason Underwood for providing valuable guidance related to performing long-read sequencing on single-cell cDNA libraries. We thank Will Hannon for assistance with interactive plots. BioRender was used to generate Figs 2.1A and 2.1B. This work was funded in part by the NIH/NIAID under grant R01AI165821 and contract No. 75N93021C00015, as well as using a Burroughs Wellcome Fund Young Investigator in the Pathogenesis of Infectious Diseases grant to JDB. JDB is an Investigator of the Howard Hughes Medical Institute.

Competing interests

JDB consults or has recently consulted with Apriori Bio, Merck, Moderna, or Oncorus on topics related to viruses and their evolution. JDB and AJG are inventors on Fred Hutch licensed patents related to viral deep mutational scanning. The other authors declare no competing interests.

2.5 Methods

2.5.1 Engineering barcodes in the influenza virus HA and NA genes

The HA segment of the A/California/04/2009 (pdmH1N1) strain of influenza virus with the G155E cell-culture adaptation mutation was engineered to carry exogenous sequence by duplicating the packaging signals at each end [32, 24], as schematized in Fig 2.1B. A complete plasmid map of the barcoded HA plasmid is at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/data/flu_sequences/plasmid_maps/pHH_bcHA_G155E_DropSeqR1.gb. We included the G155E mutation as it greatly enhances viral growth in cell culture [15]. Packaging signal length and location was informed by previous studies [27, 95, 49]. The terminal 105 nucleotides of the HA coding sequence were duplicated to provide an authen-

tic packaging signal at the 5' end of the vRNA. The corresponding 105 nucleotides of the HA protein coding sequence were synonymously recoded to remove competing RNA-RNA interactions. A second stop codon (TGA) was added at the end of the coding sequence to reduce the chance of translation read-through. The stop codons were followed by an exogenous sequence containing a priming site, a 16-nucleotide random barcode, a second priming site, and a HindIII restriction site. The 3' end of the vRNA was treated similarly. The first 67 nt of the HA coding sequence were duplicated and the corresponding region of the coding sequence was synonymous recoded. All potential start codons were removed from the duplicated packaging signal using single nucleotide substitutions. A BamHI restriction site was added between the duplicated packaging signal and the start codon.

The NA segment of the A/California/04/2009 strain was engineered using the same strategy except we duplicated 99 nucleotides at the 5' end of the vRNA and 93 nucleotides at the 3' end of the vRNA. A complete map of the barcoded NA plasmid is at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/data/flu_sequences/plasmid_maps/pHH_bcNA_DropSeqR1.gb.

2.5.2 Cloning barcoded plasmid libraries

To facilitate cloning highly-diverse barcoded plasmid libraries, a recipient vector was created for each segment. The recipient vectors contained an eGFP insert flanked by the duplicated packaging signals described above. Recipient vector maps are at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/data/experiment_resources/plasmid_maps/2548_pHH_Haflankpdm-eGFP-DropSeqR1.txt and https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/data/experiment_resources/plasmid_maps/2549_pHH_Naflankpdm-eGFP-DropSeqR1.txt.

Inserts were prepared by amplifying the HA and NA genes from templates with synonymously-recoded terminal regions. Random barcodes were added as a string of 16 nucleotides in the primer that binds near the 3' end of the viral mRNA. PCR was performed using KOD Hot Start Master Mix with 1 ng of plasmid template for 17 cycles. Reactions were treated with DpnI for 1 hour to remove the template plasmid. Barcoded products were gel purified and

cleaned with 1X AmpureXP beads. The recipient vectors were prepared by digestion with BamHI and XbaI for 1 hour to remove the eGFP insert and linearize the backbone. Linear backbones were gel purified and cleaned with 1X AmpureXP beads.

Plasmids were assembled from linear vector and barcoded insert using NEBuilder HiFi Assembly Master Mix. A 2:1 molar ratio of insert to vector was used. 25 µl of NEBuilder Master Mix was combined with 0.27 pmol of barcoded insert and 0.13 pmol of linearized vector in a total volume of 50 µl. Assembly was allowed to proceed for 1 hour. Reactions were cleaned with 0.6X AmpureXP beads and eluted in 26 µl of EB. A small portion of the assembled product (1 µl) was used to transform 20 µl of NEB 10-Beta electrocompetent E. coli cells. Transformation was performed at 1.8 kV for >5 ms per sample. Cells were grown in SOC media for 1 hr at 37C with shaking.

After shaking, transformed E. coli were plated on large LB-ampicillin agar plates and grown at 37C overnight to produce a "lawn" of bacterial colonies. Liquid medium was pipetted onto the plate and a sterile plastic scraper was used to collect all of the bacterial colonies. Bacteria were grown in 200 ml of liquid medium in a 1 liter flask for 4 hours at 37C with shaking. Bacteria were pelleted by centrifugation and frozen at -20C. Plasmid libraries were collected using Qiagen HiSpeed Maxi Prep kit.

2.5.3 Generating a dual-barcoded virus library

We generated a dual-barcoded virus library with all non-HA/NA genes derived from the A/California/04/2009 (pdmH1N1) strain of influenza virus. Virus was generated by reverse genetics in 39 independent transfection reactions. For each transfection reaction, 4e5 293T cells (ATCC CRL-3216) were seeded in a well of a 6-well dish. Cells were grown in D10 medium (DMEM supplemented with 10% heat-inactivated fetal bovine serum, 2 mM L-glutamine, 100 U per mL penicillin, and 100 g per mL streptomycin). After 16 hours, we transfected each well with bidirectional reverse-genetics plasmids based on the pHW2000 vector [34] carrying the six unmodified segments: (PB2, PB1, PA, NP, M, and NS), unidirectional reverse-genetics plasmids based on the pHH21 vector [58] carrying the two barcoded segments (HA and NA), and a plasmid constitutively expressing the TMPRSS protease

(which proteolytically activates HA) [46]. Maps of all plasmids are available at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/tree/main/data/flu_sequences/plasmid_maps. We used 250 ng of each plasmid and 3.4 µl BioT transfection reagent per reaction.

Twenty-four hours after transfection, the medium was replaced with Influenza Growth Medium (Opti-MEM supplemented with 0.1% heat-inactivated FBS, 0.3% bovine serum albumin, 100 g per mL of calcium chloride, 100 U per mL penicillin, and 100 g per mL streptomycin) and 3e5 MDCK-SIAT1-TMPRSS2 cells [46] were added to each well. Viral supernatants were collected at 65 hours post-transfection and centrifuged at 500 RCF for 5 min to remove any cellular material. Aliquots were frozen at -80°C and titered by TCID50 assay.

To ensure a genotype-phenotype link between the viral genome and the proteins displayed on the surface of each virion, the virus library was passaged at low MOI. Infections were done at large scale to maintain library diversity. Four five-layer flasks (Falcon 353144) were seeded with 50 million MDCK-SIAT1-TMPRSS2 cells each [46] in D10 medium, for a total of approximately 200 million cells. After 4 hours, the medium was removed and two million TCID50 units of virus library in IGM were used to infect the cells. Viral supernatants were collected at 38 hours after infection and centrifuged at 500 RCF for 10 min to remove cellular material. Aliquots were frozen at -80°C and titered by TCID50 assay. We obtained titers of 1e4 TCID50/µl (Fig 2.1C).

2.5.4 Estimating the rate of infected cell multiplets and chimeric PCR products using a second control virus library

Immediately prior to performing single-cell RNA sequencing on our sample of interest, we mixed the infected cells with a second control sample of cells. The control cells were infected with an otherwise isogenic influenza virus that carried identifying synonymous mutations on all eight viral genes. The synonymous mutations are detectable by sequencing. They mark each mRNA transcript and genome segment derived from the virus library with a distinct "genetic tag." These synonymous genetic tags allow us to distinguish between viral transcripts from our sample of interest and the control sample, thereby enabling us to

quantify two important sources of technical error.

First, in the single-cell RNA sequencing data, these tags provide a means to detect transcriptomes that are derived from droplets that encapsulated multiple infected cells (multiplets) [8]. Such transcriptomes are marked by high frequencies of both tags among the viral transcripts. The overall rate of multiplets among all cells was calculated, and multiplets bearing both tags (which will be about half of multiplets) were excluded to remove them from the dataset.

Second, the genetic tags are also detectable in the long-read viral sequencing data [73] we used to reconstruct the genotype of infecting virions. In the course of preparing long-read sequencing libraries, a polymerase can move from one template molecule to another in the midst of synthesizing its product–a phenomenon known as "strand exchange" [40]. This phenomenon can be detected in long-read viral sequences that contain discordant genetic tags (see Fig S10 of [73]). We estimated the rate at which this type of error occurs, and sequences bearing both tags were excluded from contributing to the results.

The second dual-barcoded virus library was prepared identically to the first viral library as described above. The second library contains synonymous variants near the 5' and 3' ends of each viral segment. Plasmid maps are available at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/tree/main/data/flu_sequences/plasmid_maps.

2.5.5 Infecting cells with a dual-barcoded virus library

Ten thousand MDCK-SIAT1-TMPRSS2 [46] cells were suspended in D10 medium and plated in a well of 24-well plate. After 5 hours, cells were observed by microscopy and were confirmed to be well-attached. The medium was aspirated and 1500 TCID50 units of dual-barcoded virus library in 100 µl of Influenza Growth Medium was added to the well. The cells were incubated with virus for 1 hour, and the plate was rocked by hand every 15 minutes. After 1 hour, the inoculum was removed and the cells were washed once with 250 µl of phosphate-buffered saline. 500 µl of Influenza Growth Medium supplemented with 20 mM ammonium chloride (to prevent further entry of virions into cells [50, 59]) was added to the well.

At 12 hours post-infection, the supernatant was collected and cells and debris were removed by centrifugation at 300 RCF for 3 min. The supernatant was split into 2 aliquots of 220 µl each and frozen at -80°C. The cells were collected by addition of 100 µl trypsin and a single-cell suspension was generated. The trypsin digestion was stopped by addition of 400 µl of D10 medium. The cells were washed 3 times with phosphate-buffered saline supplemented with 0.8% by volume non-acetylated bovine serum albumin. The cells were counted to confirm that approximately 10,000 cells were present per well.

2.5.6 Single-cell RNA sequencing

Infected cells were prepared and mixed with a second control sample of infected cells to control for technical sources of error (see "Estimating the rate of infected cell multiplets and chimeric PCR products using a second control virus library" above). Approximately 20,000 cells were loaded into the 10X Chromium device. Single-cell RNA sequencing was performed with the 10X Chromium Next GEM Single Cell 3' GEM, Library Gel Bead Kit v3.1. The manufacturer's standard protocol [3] was used with the following modifications. The template-switching oligo was replaced with a modified single-stranded DNA oligo with the sequence 5'-AGAGTGTTTGGGTAGAGCAGCGTGTTGGCATGTrGrGrG-3' at a final concentration of 45 µM in the reaction mix. This change was made to accommodate some of the barcoded influenza segments' exogenous sequence which shares homology with the standard 10X template-switching oligo. The cDNA amplification primer mix was replaced with a pair of primers with the sequences 5'-AGAGTGTTTGGGTAGAGCAGCG-3' (binding to the custom template-switch oligo mentioned above) and 5'-CTACACGACGCTCTTCCGATCT-3' (binding to the standard 10X adapter sequence) at a final concentration of 1 µM in the reaction mix. The cDNA amplification PCR reaction extension time was increased to 20 seconds to encourage the formation of full-length cDNA products. The amplified cDNA product was split in half; one half was used for fragmentation and preparation of the transcriptome sequencing library while the other half was used as template for long-read sequencing of viral transcripts.

2.5.7 Viral long-read sequencing to reconstruct infecting viral genomes

We determined the sequence of the virion that infected each cell. Because the cells were infected at a low MOI, infection was initiated by onevirion in the large majority of infected cells. To capture these sequences, we selectively enriched viral cDNA molecules using a method described previously [73]. In brief, cDNA derived from the 10x Genomics protocol was first amplified in a semi-specific PCR reaction. Each segment was amplified with a primer annealing to the universal TruSeq primer site that is added to all cDNA molecules during the reverse transcription step of 10x Genomics protocol and a segment-specific primer annealing to 5' end of the viral mRNA, which also contains a flanking sequence that is complementary to the TrueSeq primer site (Table S1). Semi-specific PCR reaction conditions were as follows: 12 ng cDNA, 0.5 µM of forward and reverse primer, 10 µl of KOD (EMD Millipore, 71842), 0.1 mg/ml BSA, and final volume adjusted to 20 µl with water. PCR was incubated for 120 s at 95°C followed by 10 cycles of 120 s at 95°C, 20 s at 55°C, 90 s at 70°C, and the final extension step at 70°C for 120 s. Semi-specific PCR reactions were purified using AMPure XP beads at 1.8x beads to sample ratio and eluted in 12 µl of water. Following purification, PCR products were circularized via complementary TrueSeq sequence. For circularization, 10 µl of purified PCR product was used in a 20 µl HiFi assembly reaction (NEB, E2621S). HiFi assembly was performed at 50°C for 1 hour.

Next, HiFi products were used in segment-specific PCR reactions. To amplify viral products of all lengths, primers that anneal to the ends of viral mRNA were used; to preferentially amplify full-length viral segments, primers that anneal to the middle of each viral segment were used (Table S2). Segment-specific PCR conditions were as follows: 9 µl of Hifi reaction, 0.5 µM of forward and reverse primer, 25 µl of KOD, and the final PCR reaction volume adjusted to 50 µl with water. PCR was incubated for 120 s at 95°C followed by cycling 120 s at 95°C, 20 s at 55°C and 90 s at 70°C with a final extension step of at 70°C for 120 s. Cycles were kept to a minimum to reduce strand exchange; since different segments required different yield, different numbers of cycles were employed each segment-specific reaction. For the polymerase segments, 14 cycles of segment-specific PCR were performed;for the HA, NA and NP segments, 10 cycles were performed;, for the M

and NS segments, 7 cycles were performed. PCR reactions were purified using AMPure XP beads at 1.8x beads to sample ratio and eluted in 12 µl of water. All purified PCR products were pooled together and long-read sequencing was performed on a PacBio Sequel II.

We generated CCS sequences of each viral transcript using PacBio long-read sequencing. We measured the rate of strand exchange that occurred during sequencing library preparation (see Fig S10 of [73]), and found that fewer than 1% of sequences were affected, providing high confidence that the sequences we obtained could be assigned to their cell of origin. We generated a consensus sequence for each viral genome (see "Computational analysis of single-cell RNA sequencing, long-read virus sequencing, and progeny production viral barcode data" below) . We counted the number of non-synonymous mutations found in each consensus genome; deletions were considered non-synonymous mutations for this purpose.

2.5.8 Quantifying progeny production

The amount of progeny produced by single infected cells was determined by sequencing the viral barcodes on vRNA molecules. To quantify physical progeny virions, we sequenced the vRNA in the viral supernatant at 12 hours post infection. To quantify infectious progeny virions, we infected a second set of cells to select for virions that could perform viral entry and genome replication (Fig 2.1A) and sequenced the intracellular vRNA molecules at 13 hours post infection.

In detail, we thawed frozen viral supernatants that were collected at 12 hours post infection and split them into four equal volumes. Two volumes were used to isolate supernatant RNA directly. The other two volumes were used to infect MDCK-SIAT1-TMPRSS2 cells [46] at a moderate estimated MOI of 0.25 in two independent replicates. To infect the cells, 60,000 MDCK-SIAT1-TMPRSS2 cells [46] were suspended in D10 medium and plated in a well of 6-well plate. After 7 hours, cells were observed by microscopy and were confirmed to be well-attached. The medium was aspirated and an aliquot of supernatant with an estimated 15,000 TCID50 units was added to the well in 500 µl of Influenza Growth Medium. The cells were incubated with virus for 1 hour, and the plate was rocked by hand every

15 minutes. After 1 hour, the inoculum was removed and the cells were washed once with 500 µl of phosphate-buffered saline. 1600 µl of Influenza Growth Medium supplemented with 20 mM ammonium chloride (to prevent further entry of virions into cells [50, 59] was added to the well. At 13 hours post infection, the cells were collected by aspirating the growth medium and incubating with 300 µl trypsin to detach them from the plate. Trypsin digestion was stopped by the addition of 700 µl of D10 medium. The cells were pelleted by centrifugation at 400 RCF for 3 min. The cell pellet was washed by resuspending in 1 ml of phosphate-buffered saline and pelleting at 400 RCF for 3 min again. The phosphate-buffered saline was aspirated and the cell pellet was flash-frozen on dry ice.

RNA was isolated from the viral supernatant or infected cell pellets using the RNeasy Mini Kit (Qiagen, 74104). Lysis buffer was mixed with the viral supernatant sample and 70% ethanol was added . For the infected cell pellets, the sample was mixed with lysis buffer and homogenized by vortexing at high speed for 20 seconds. The homogenized sample was processed on a gDNA eliminator spin column to remove genomic DNA. The processed sample was combined with 70% ethanol. From this point, both the viral supernatant and infected cell pellets were treated identically and followed the standard RNA purification protocol specified by the manufacturer [66]. The RNA for each sample was eluted in 50 µl of RNase-free water.

Reverse transcription was performed with a segment-specific primer targeted to the HA or NA vRNA (Table S3). Two replicate reactions were performed using RNA from the viral supernatant sample, and two independent reactions were performed using RNA from the two infected cell pellets; these replicates provide technical duplicate measurements of both the physical progeny in the supernatant and the infectious progeny in the cell pellets.

Reverse transcription was performed using the SuperScript III First-Strand Synthesis SuperMix kit according to the manufacturer protocol [87]. For the viral supernatant samples, 12 µl of each RNA sample was used as template for each 40 µl reaction. For the infected cell pellet samples which contain much larger amounts of total RNA due to the host RNA present in the cell, 1000 ng of RNA was used as template for each 40 µl reaction. The low-concentration cDNA generated from the viral supernatant samples was purified and concentrated using 2X Ampure SPRI beads and eluted into 22 µl of elution buffer.

Viral barcodes were amplified in 50 µl PCR reactions using KOD Hot-Start Master Mix (Sigma-Aldrich, 71842). For the viral supernatant samples, 22 µl of concentrated cDNA was used as template. For the high-concentration infected cell pellet samples, 10 µl of unpurified cDNA was used as template. Segment-specific primers (Table S3) were used and reactions were run for 20 cycles. Amplicons were size-selected and purified using a double-sided AmpureXP bead cleanup. Samples were first combined with 0.8X AmpureXP beads and the supernatant was collected. The supernatant was then combined with 1.8X AmpureXP beads and the bound DNA was collected.

Sequencing indices and adapters were attached in a 50 µl PCR reaction using KOD Hot-Start Master Mix. For all samples, 2 ng of purified amplicon DNA was used as template. Sample-specific index primers (Table S3) were used and reactions were run for 20 cycles. The resulting amplicons were gel-purified and pooled for single-end sequencing on an Illumina MiSeq. The progeny contribution of each cell was calculated (see "Computational analysis of single-cell RNA sequencing, long-read virus sequencing, and progeny production viral barcode data" below).

2.5.9 Computational analysis of single-cell RNA sequencing, long-read virus sequencing, and progeny production viral barcode data

A reproducible pipeline that performs all analysis is at https://github.com/jbloomlab/barcoded_flu_pdmH1N1. The pipeline uses Snakemake [43]. The pipeline begins with raw sequencing data and ends by generating the figures shown in this manuscript. Most code in the pipeline is arranged in Jupyter notebooks (https://jupyter.org).

Briefly, the raw sequencing data from the single-cell RNA sequencing was aligned using STARsolo [41] against a composite reference made up of the canine genome CanFam3.1.98 concatenated to the A/California/04/2009 influenza virus genome. Alignment produced a cell-gene matrix containing the gene expression of every canine and virus gene for each single cell. Custom Python code was used to parse the "genetic tag" encoded on viral transcripts which differentiates our library of interest from a second control library. The multiplet rate was calculated and only transcriptomes from our library of interest were used for analysis.

Transcriptomes from the second control library and transcriptomes composed of multiple infected cells were excluded.

The total viral gene expression was calculated for each infected cell. Cells were called as infected if at least 1% of their transcripts came from virus. Individual viral genes were called as expressed if their frequency was greater than the 99th percentile observed in uninfected cells (see Fig A3).

For the statistical test in Fig 2.5B, we classified each cell as expressing all viral genes, missing the NS gene, or missing another influenza gene. Because the data are not normally distributed, we performed the non-parametric Kruskal-Wallis one-way analysis of variance test to determine if all three groups were sampled from distributions with the same median viral transcription. The results indicated a statistically significant difference in viral transcription among the three groups (p ¡ 0.001). To determine which–if any groups–had statistically significant pairwise differences, we performed a post hoc Dunn's test and adjusted for multiple hypotheses using Bonferroni correction. The results indicated that cells missing the NS gene have a statistically significant difference in viral transcription compared to cells that express all viral genes (p ¡ 0.01) and also compared to cells that are missing other viral gene(s) (p ¡ 0.001). The difference in viral transcription between cells expressing all viral genes and cells missing a viral gene other than NS was not statistically significant (p = 0.06) using these methods.

The raw PacBio sequencing data was processed using PacBio's ccs program (https://github.com/PacificBiosciences/ccs). Consensus sequences were generated from the subread files, requiring a minimum accuracy ('rq') of 0.99 for the consensus sequence. The chimera rate was estimated using the "genetic tags". The cell barcode and UMI were parsed from each CCS using custom Python code that utilized the alignparse package [16]. A consensus sequence was called for each cell barcode-viral gene-UMI combination. A mutation was included in the consensus sequence if it was found in ¿50% of the CCS for the cell barcode-viral gene-UMI combination. A consensus sequence was then called for each cell barcode-viral gene combination. A mutation was included in the consensus sequence if it was found in >50% of the UMIs for the cell barcode-viral gene and was found in at least two UMIs.

To parse the viral barcodes sequenced from the supernatant (representing physical

progeny) and from the second infection (representing infectious progeny), we used custom Python code that utilized the dms$_v ariantspackage$(). The viral barcodes were error-corrected using UMI-tools [80]. The technical replicates for each sample were plotted against each other and the limit of detection was set at 1e-5, where viral barcode frequencies fail to correlate (Fig A8), indicating bottlenecked subsampling of the molecules carrying the viral barcodes, and suggesting that frequency measurements below this threshold are not reliable; values below the limit of detection were set to the limit of detection. The mean frequency of both replicates was calculated. A subset of infected cells expressing both barcoded viral genes and with complete long-read sequencing data was used to calculate progeny contributions. To determine the fraction of progeny contributed by each infected cell in this set, we took the geometric mean of the HA and NA barcode frequencies associated with each cell. We normalized the progeny contributions by the total frequencies assignable to any cell in this set. The data were visualized in Jupyter notebook (https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/final_analysis.py.ipynb). We used custom Python code utilizing a combination of plotnine (https://github.com/has2k1/plotnine) and altair (https://github.com/altair-viz/altair). An R script utilizing gggenes (https://github.com/wilkox/gggenes) was used to plot the complete viral genomes of infected cells. The figures generated by this notebook are displayed in this manuscript.

subsectionData availability All data and code are available in the GitHub repository at https://github.com/jbloomlab/barcoded_flu_pdmH1N1. The analysis can be reproduced by running the Snakemake pipeline and final analysis notebook according to the instructions at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/README.md.

Key output files are hosted at the following locations. All raw sequencing files will is available on GEO at accession number GSE214938. An integrated CSV produced by the Snakemake pipeline with cell barcodes, viral gene expression, viral genome sequence, and viral barcode frequencies is available at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/results/viral_fastq10x/scProgenyProduction_trial3_integrate_data.csv. The final CSV file with progeny contribution measurements, viral gene expression, and viral mutations for the 92 infected cells that express both barcoded genes and have sequencing data for all expressed viral genes is available at https://github.com/jbloomlab/barcoded_flu_

pdmH1N1/blob/main/results/viral_fastq10x/scProgenyProduction_trial3_complete_measurements_cells_data.csv.

Chapter 3

BARCODED VIRUS LIBRARIES: IMPROVEMENTS AND APPLICATIONS

In Chapter 2, we used diverse barcoded influenza virus libraries to quantify progeny production from single virus-infected cells. The methods we developed to generate barcoded virus libraries were capable of supplying approximately 1500 uniquely-traceable virions in a single experiment. After single-cell RNA sequencing, we were able to describe the realtionship between viral transcription and progeny production in approximatley 100 infected cells. These experiments provided insight into the effects of viral gene absence on viral transcription and progeny production, but other characteristics of the infected cells were left unstudied due to a lack of statistical power. In the future, it may be possible to increase the diversity of barcoded virus libraries and to improve the methods of sequencing single virus-infected cells. These improvements would allow progeny production measurements to be made from single virus-infected cells at a much larger scale. They would also enable new experiments to be performed that could provide insight into several open questions related to virus replication and transmission.

3.1 Improvements to barcoded virus libraries

To quantify progeny produced from single virus-infected cells, a barcoded virus library must be sufficiently diverse that nearly every virion in an inoculum carries a unique viral barcode. If the virus library is not sufficiently diverse, more than one cell will be infected by virions carrying the same viral barcode, and the progeny generated by these cells will be indistinguishable. To date, we have been able to generate barcoded virus libraries that supply approximately 1500 uniquely-barcoded virions. This enables single-virion tracing in experiments that can be conducted at a relatively small scale; for example, we have infected about 10,000 single cells with influenza virus at a multiplicity of infection (MOI) of 0.15. However, to study a greater number of infected cells or to infect the same number of cells

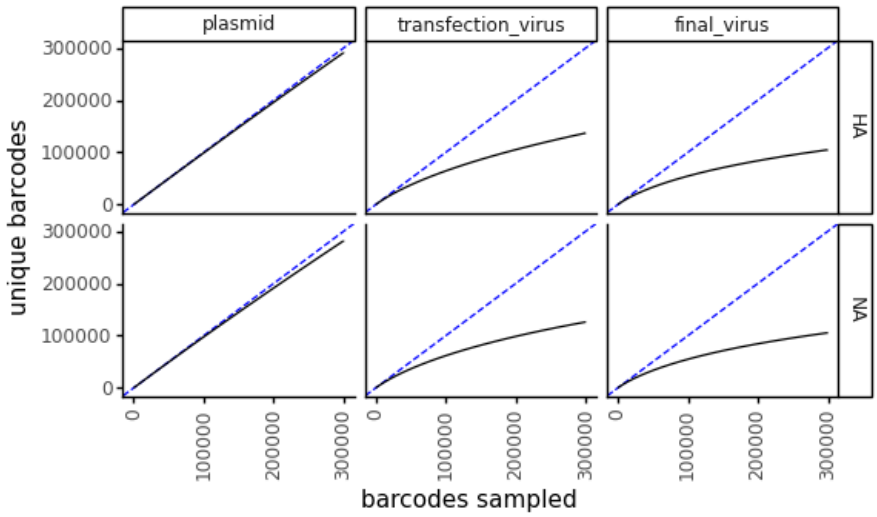


Figure 3.1: Rarefaction curves of barcoded plasmid, virus generated by transfection, and passaged final virus libraries. The dotted blue line represents a hypothetical perfectly diverse library

at a higher MOI, a more diverse barcoded virus library is required.

Rarefaction curves provide a convenient way to visualize and quantify library diversity (Fig 3.1). These curves describe the number of unique barcodes in a sample of a given size. A perfectly diverse library would have a linear rarefaction curve: every barcode sampled would be unique. Less diverse libraries display diminishing returns. At small sample sizes, the number of unique barcodes is close to the size of the sample. As the sample size increases, the number of unique barcodes increases at a rate less than 1:1. When the total number of unique barcodes in the library has been reached, the number of unique barcodes sampled no longer increases.

There are two ways to increase the diversity of a library. First, the number of unique barcodes in the library can be increased. A library with 1 million unique barcodes is inherently more diverse than a library with just one thousand unique barcodes. Second, the frequency of barcodes in a library can be made more homogenous. When randomly sampling a subset of the library, as you would in the types of experiments discussed here,

the maximum number of unique viral barcodes is provided when barcode frequencies are equal. As the frequencies become more variable, it becomes increasingly likely to sample the same barcode twice in a sample of a given size.

Some steps of the protocol used to generate barcoded virus libraries are not likely to benefit from further optimization. Barcoded plasmid pools with extremely high numbers of unique barcodes and homogenous barcode frequencies can be generated using established methods (Fig 3.1 "plasmid"). For example, we have generated plasmid libraries that supply hundreds of thousands of uniquely-barcoded plasmids in a sample of similar size. Similarly, it is straightforward to passage barcoded virus libraries in a way that preserves the diversity present (Fig 3.1 "final virus"). When we passage barcoded influenza virus libraries at a low multiplicity of infection at very large scale (e.g. MOI of 0.01 in 50 million cells) we see only a minimal decrease in diversity.

However, when we generate barcoded influenza virus libraries using the eight-plasmid reverse genetics system [34] (Fig 3.1 "transfection virus"), we see a substantial decrease in diversity compared to the input plasmid library (Fig 3.1 "plasmid"). To generate a more diverse virus library, virus generation needs to be improved. Here, I will present three potential strategies to improve virus generation.

First, it may be possible to improve the efficiency with which individual cells are transfected and produce infectious influenza virus. Generation of infectious virus depends on the concurrent transfection of eight or more different plasmids encoding influenza proteins and vRNA segments [34]. Conditions that increase the efficiency of plasmid uptake or decrease the toxicity of transfection reagents may be beneficial. Optimization of the relative level at which influenza genes and host co-factors are expressed might also improve library diversity.

Second, helper virus offers an alternative to approaches based solely on the transfection of plasmids. Life the first approach, this approach might increase the total number of cells that contribute infectious progeny to the barcoded virus library. Helper virus strategies utilize semi-infectious viral particles that can infect a cell but require the products of a transfected plasmid to complete the viral lifecycle and generate progeny virions. In previous work generating virus libraries for deep mutational scanning (DMS) experiments, influenza helper virus was found to increase the efficiency of virus generation and improve the corre-

lation of variant frequencies between replicate virus libraries [20]. The likely mechanism of improvement is that functional copies of most influenza segments are present in the helper virus, and therefore, efficiently delivered to each cell.

The third strategy is intended to provide more even representation of viral barcodes in the library by reducing variation in the number of virions carrying each barcode. To address the variation in barcode frequencies, the number of independent transfection reactions could be increased to decrease the impact of "jackpot effects", where a highly-productive virion propagates throughout a culture at the expense of other virions. Splitting a reaction into smaller parts limits the extent to which the most productive transfected cells can monopolize resources. A successful example of this approach is the process of cloning barcoded plasmid libraries. Transformed bacteria are grown on a plate of solid agar media, rather than in liquid culture, to limit the growth of individual colonies spatially. The same purpose could be served by performing more transfections at smaller scale, for example in 96-well plate format.

3.2 Improvements to single-cell RNA sequencing of virus-infected cells

To quantify viral transcription in single virus-infected cells, the viral transcripts must be captured, reverse-transcribed, and sequenced efficiently. Each of these processes has a rate of failure [102], and information is lost at each step. These requirements are even more stringent when attempting to identify viral barcodes embedded in infecting viral genomes. The viral barcodes lie at a specific position in the transcript, which must have reliable sequencing coverage in the transcriptome library. Viral barcodes from infected cells with low viral transcription will be rare in the final single-cell RNA sequencing library, biasing the recovered data toward high-transcribing cells. Furthermore, influenza virus fails to express each viral gene in a portion of infected cells [74, 85].

To increase the rate at which viral transcripts are captured, 5'-end single-cell RNA sequencing could be employed. In this approach, the cell barcode is appended to the 5' end of each transcript via template switching [64]. This allows any primer to be used for reverse transcription. For transcriptome studies, primers targeting the polyA tail found at the 3' end of all host mRNA molecules are generally used. The addition of primers targeting viral

sequences may increase the capture and reverse transcription efficiency of viral transcripts specifically. Furthermore, alternative primers could be used to capture RNA molecules from viruses that are not polyadenylated [4]. While this could provide more complete coverage of viral transcripts, it would inflate the relative amount of viral cDNA compared to host cDNA. The absolute fraction of viral mRNA in infected cells could no longer be determined, however, the relative differences between infected cells would be maintained.

To address issues leading to the loss of barcoded transcripts specifically, we previously barcoded two genes in the influenza virus genome, providing redundancy in the event that a barcode was lost for any reason. We barcoded the haemagglutinin gene (HA), which encodes the viral entry protein, and the neuraminidase gene (NA), which encodes the protease that releases newly formed virions from the surface of the cell. We found that the NA gene was expressed at lower average levels and in a smaller fraction of infected cells than the HA gene. For many cells, we recovered viral barcode information from the HA gene but not from the NA gene. Based on this experience, it may be more effective to embed barcodes on a viral gene that is more consistently expressed in infected cells and is expressed at a high level in each infected cell. The influenza gene that fits these criteria the best is the nucleoprotein (NP) gene (Fig A3). It is absent in a smaller fraction of infected cells profiled by single-cell RNA sequencing than other viral genes, and is expressed at a relatively high average level.

Another way to address loss of viral barcode information could be to selectively sequence barcoded viral segments. In our work, we have used a series of PCR and circularization reactions to specifically amplify influenza transcripts for long-read viral genome sequencing [73]. The same strategy could be applied to the barcoded transcripts specifically to increase the rate at which they are sequenced. Alternative methods of enriching specific transcripts from a complex mixture could also be tested. One option is pull down with biotinylated oligonucleotides. The oligonucleotides are designed to bind to a sequence of interest. After binding, streptavidin beads are used to pull down the biotinylated dsDNA. This approach has been used to target the cell barcodes of specific cells in single-cell RNA sequencing cDNA libraries [69].

Finally, single-cell methods that measure multiple molecular species (dubbed "mutli-

omic" approaches) have been found improve the sensitivity and specificity of identifying specific subpopulations of single cells in some settings [29]. It is now fairly trivial to generate DNA-barcoded antibodies that can be used to measure the expression of surface proteins on single cells [84]. Addition of one or more virus-specific antibodies (e.g. targeting the HA protein) to the single-cell sequencing protocol could provide an orthogonal measurement of viral products. Such information might improve discrimination of infected and uninfected cells beyond the current heuristic approach based on a threshold of total viral transcription.

3.3 Future applications of barcoded virus libraries

Libraries of uniquely-traceable barcoded virions have potential applications in several areas of virology. Direct extensions of the work presented in this dissertation would explore factors that may influence progeny production in single virus-infected cells. Further applications might use viral barcodes to examine the cell types infectious virus is produced in and the number of viral particles initiating new infections. Examples of questions that barcoded virus libraries could be utilized to address are:

1. What is the effect of host cell state on progeny production in single virus-infected cells?

2. How does the distribution of single-cell progeny production vary over the course of an infection?

3. How does progeny production vary across cell types represented in the lung?

4. What is the relationship between physical bottlenecks and genetic bottlenecks measured during viral transmission events?

To effectively study host gene expression's influence on virus progeny production, more infected single cells must be profiled. The current study measured progeny production, viral transcription, and host gene expression in only 92 infected cells. Mammalian transcriptomes express thousands of individual genes [102]. Cell states can be defined by relatively small changes to a subset of this repertoire. More infected cell transcriptomes are necessary

to achieve the statistical power required to detect changes in host gene expression. The simplest way to increase the scale of single-cell progeny production studies would be to run several experiments in parallel and pool the resulting data. This approach has been applied successfully to a set of single-cell experiments that quantified progeny production using plaque assays [33]. Further gains in scale would be achieved if a more diverse virus library can be generated. This approach may be preferred because it would ultimately reduce the of labor required to run an individual experiment; this would facilitate larger studies over the long term. Simultaneous profiling of cells from the same experiment also offers the advantage of reduced batch effects, which are an important concern for single-cell RNA sequencing studies [90].

Characterizing the distribution of virus progeny production at multiple time points would provide valuable insights into the dynamics of viral replication. In the current study, we collected viral progeny and virus-infected cells at 12 hours post-infection. This is a relatively early time point in the pdmH1N1 influenza virus replication cycle. Specifically, this was the earliest time point at which we could reliably detect infectious progeny from cells infected at a low multiplicity of infection. We do not know whether the extreme heterogeneity we observe in progeny production – with just a small percentage of cells contributing most of the viral progeny – is maintained over time. It may be the case that only a small percentage of highly-productive infected cells generate virions from the beginning of an infection to its end. Or, it is possible that the distribution of progeny production changes throughout the course of an infection, with less variable or even more variable progeny production observed at various times. Different cells could also generate new virions at different times.

To conduct a time course study of influenza virus progeny production, the current experiments could be repeated at multiple time points. This study design would provide information about the relationship between viral transcription and progeny production at a variety of times after infection. However, because single-cell RNA sequencing is a destructive method, each time point would require a unique sample that is unrelated to the others. Alternatively, the supernatant from a set of influenza-infected cells could be sampled at multiple time points. This study design would provide longitudinal measurements of progeny production from the same infected cells at the expense of characterizing viral

transcription throughout the time course.

Recent studies have used single-cell RNA sequencing to study viral transcription in tissues collected from mice [83] or cells collected from the upper airways of humans [13]. These studies have found that many cell types are infected by influenza virus and that, for both influenza virus [83] and SARS-CoV-2 [81], viral transcription is highest in epithelial cells. These studies are not equipped, however, to determine where transmissible virus is generated in the repsiratory tract.

To begin to address this question, it would be useful to measure progeny production from a variety of cell types in a lung model system. Primary human lung epithelial cells can be cultured in an air-liquid-interface format, which triggers differentiation into multiple epithelial lung cell types [61]. We propose a study infecting a well-differentiated ALI culture with barcoded virus libraries, quantifying the progeny produced by each infected cell, and using single-cell RNA sequencing to determine cell type of each infected cell. This study could determine the relative amount of viral progeny produced by each cell type. It could test for differences in the relationship between viral transcription and progeny production across cell types. It could also provide insight into host factors that promote or restrict viral progeny production, as different cell types will have reproducible differences in host gene expression. Finally, such a study would provide an analytical framework for future in vivo studies of viral progeny production. By infecting a small animal model with diverse barcoded virus libraries, it may be possible to measure progeny production from the respiratory tract directly.

The genetic bottleneck of a virus transmission event is the number of unique viral genomes that pass from donor to recipient. Using deep sequencing of known transmission pairs, the genetic bottleneck has been determined for acute influenza virus infections [53, 97] and for actue SARS-CoV-2 infections [9, 48]. These studies have found that both respiratory viruses have relatively narrow genetic bottlenecks, with just one or a few unique viral variants initiating a new infection. This narrow genetic bottleneck has an important implication for viral evolution. Because fit variants that arise in the donor may not be sampled during transmission, the efficiency of natural selection is limited, and genetic drift dominates viral evolution in acute infections [53].

Because multiple virions can share identical genome sequences, the genetic bottleneck may not represent the number of physical viral particles passing from donor to recipient. Diverse barcoded virus libraries provide a means to distinguish between virions carrying the same viral variants. *In vivo* studies of transmission between small animals using barcoded virus libraries have assessed the physical transmission bottleneck and found that more particles are passed between animals that have physical contact than between animals that share only air [91]. The barcoded virus library used in this study contained approximately 100 unique viral barcodes. With a library of this size, it is possible that multiple virions would share identical barcodes and that estimates of unique virions initiating infection would be artificially reduced. A more diverse barcoded virus library with a greater number of unique barcodes would provide greater resolution to measure the number of unique viral particles initiating an infection in an innoculated animal or transmissiting from one animal to another.

Chapter 4

CONCLUSION

In this dissertation, I present new methods to quantify viral progeny produced by single virus-infected cells. Highly diverse libraries of barcoded virions are generated such that each virion in an experiment carries its own unique viral barcode. Cells are infected with these virus libraries and deep sequencing is used to quantify the number of progeny virions bearing each viral barcode. This approach differs from other methods used to measurement progeny production because it does not require isolation of individual cells during infection to provide single-cell resolution. The pooled format is compatible with modern genomic techniques, including single-cell RNA sequencing.

Leveraging diverse barcoded virus libraries, I simultaneously quantified progeny production and viral transcription in single influenza virus-infected cells. I described the distribution of viral transcription and progeny virion production in the same infected cells. I found that transcription and progeny production are not well correlated in cells infected with pdmH1N1 influenza virus at a low multiplicity of infection at 12 hours post-infection. Single-cell measurements of viral gene expression explain some of the discordance between viral transcription and progeny production. Cells with the highest viral transcription often fail to express the influenza non-structural (NS) gene. Based on published studies [60, 11, 71, 56], it is likely that influenza NS acts as a negative regulator of viral transcription, and its loss permits abnormally high levels of viral transcription. Like nearly all cells missing influenza genes, cells without NS contribute no detectable progeny.

This work was performed at a limited scale and was not powered to assess the effect of host gene expression on virus progeny production. The first experiments conducted with this system serve as proof of the principle that diverse barcoded virus libraries can be used to trace individual virions throughout a single virus replication cycle. Improvements to the virus generation methods may permit more diverse barcoded virus libraries to be produced,

which would further increase their utility. Several avenues for further optimization of virus generation have become apparent in the course of developing these methods, but currently remain untested.

With the current virus libraries, more statistical power could be achieved by simply running multiple experiments in parallel and pooling the resulting data. Similarly, the current virus libraries could be used to conduct time course studies that describe the dynamics of progeny production in single cells over time.

In the future, if more diverse barcoded virus libraries are developed, novel experiments would become possible. Studying progeny production in differentiated human airway cultures would quantify infectious virus generated by diverse cell types, and explore how the relationship between host factors and progeny production varies by cell type. In vivo studies of animals infected with barcoded virus library could provide insight into where in the respiratory tract and in which cell types transmissible virus is generated. Finally, highly-diverse barcoded virus libraries may be able to provide more precise measurements of physical viral particles initiating an acute viral infection. When applied in animal studies of transmission, these measurements might help resolve the relationship between physical transmission bottlenecks and genetic transmission bottlenecks.

In conclusion, I have worked to extend genomic assays to be more useful for virology applications. I have developed new methods to generate extremely diverse barcoded virus libraries. I used these libraries to quantify progeny production from single influenza virus-infected cells. Using single-cell RNA sequencing, I made simultaneous measurements of viral transcription on the same infected cells. Viral gene expression provided a partial explanation for why progeny production varies by three orders of magnitude between cells infected with influenza virus under identical conditions. Improvements to the methods described here are likely possible, and novel experiments utilizing individually traceable virions could contribute to progress on several important questions about viral replication and transmission.

BIBLIOGRAPHY

[1] Comprehensive single-cell transcriptional profiling of a multicellular organism | Science.

[2] dissertation.

[3] Library Construction - Official 10x Genomics Support.

[4] Signe Altmäe, Nerea M. Molina, and Alberto Sola-Leyva. Omission of non-poly(A) viral transcripts from the tissue level atlas of the healthy human virome. *BMC Biology*, 18(1):179, November 2020.

[5] Katherine A. Amato, Luis A. Haddock, Katarina M. Braun, Victoria Meliopoulos, Brandi Livingston, Rebekah Honce, Grace A. Schaack, Emma Boehm, Christina A. Higgins, Gabrielle L. Barry, Katia Koelle, Stacey Schultz-Cherry, Thomas C. Friedrich, and Andrew Mehle. Influenza A virus undergoes compartmentalized replication in vivo dominated by stochastic bottlenecks. *Nature Communications*, 13(1):3416, December 2022.

[6] Raul Andino and Esteban Domingo. Viral quasispecies. *Virology*, 479-480:46–51, May 2015.

[7] L. G. Baum and J. C. Paulson. The N2 neuraminidase of human influenza virus has acquired a substrate specificity complementary to the hemagglutinin receptor specificity. *Virology*, 180(1):10–15, January 1991.

[8] Jesse D. Bloom. Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ*, 6:e5578, September 2018.

[9] Katarina M. Braun, Gage K. Moreno, Cassia Wagner, Molly A. Accola, William M. Rehrauer, David A. Baker, Katia Koelle, David H. O'Connor, Trevor Bedford, Thomas C. Friedrich, and Louise H. Moncla. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS Pathogens*, 17(8):e1009849, August 2021.

[10] Christopher B. Brooke, William L. Ince, Jens Wrammert, Rafi Ahmed, Patrick C. Wilson, Jack R. Bennink, and Jonathan W. Yewdell. Most Influenza A Virions Fail To Express at Least One Essential Viral Protein. *Journal of Virology*, 87(6):3155–3162, March 2013.

[11] Rosario Bullido, Paulino Gómez-Puertas, Marıa José Saiz, and Agustın Portela. Influenza A Virus NEP (NS2 Protein) Downregulates RNA Synthesis of Model Template RNAs. *Journal of Virology*, 75(10):4912–4917, May 2001.

[12] F. M. Burnet. A method for the study of bacteriophage multiplication in broth. *British Jour Exp Path*, 10:109–115, July 1929.

[13] Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, August 2017. Publisher: American Association for the Advancement of Science.

[14] Yuming Cao, Zhiru Guo, Pranitha Vangala, Elisa Donnard, Ping Liu, Patrick McDonel, Jose Ordovas-Montanes, Alex K. Shalek, Robert W. Finberg, Jennifer P. Wang, and Manuel Garber. Single-cell analysis of upper airway cells reveals host-viral dynamics in influenza infected adults. *BioRxiv*, April 2020. Type: article.

[15] Zhongying Chen, Weijia Wang, Helen Zhou, Amorsolo L. Suguitan, Cindy Shambaugh, Lomi Kim, Jackie Zhao, George Kemble, and Hong Jin. Generation of Live Attenuated Novel Influenza Virus A/California/7/09 (H1N1) Vaccines with High Yield in Embryonated Chicken Eggs. *Journal of Virology*, 84(1):44–51, January 2010.

[16] Katharine Crawford and Jesse Bloom. alignparse: A Python package for parsing complex features from high-throughput long-read sequencing. *Journal of Open Source Software*, 4(44):1915, December 2019.

[17] A R Davis, A L Hiti, and D P Nayak. Influenza defective interfering viral RNA is formed by internal deletion of genomic RNA. *Proceedings of the National Academy of Sciences*, 77(1):215–219, January 1980.

[18] M. Delbrück. The Burst Size Distribution in the Growth of Bacterial Viruses (Bacteriophages). *Journal of Bacteriology*, 50(2):131–135, August 1945.

[19] Dan Dou, Rebecca Revol, Henrik Östbye, Hao Wang, and Robert Daniels. Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement. *Frontiers in Immunology*, 9, 2018.

[20] Michael B. Doud and Jesse D. Bloom. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses*, 8(6):155, June 2016. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[21] J. W. Drake. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Annals of the New York Academy of Sciences*, 870:100–107, May 1999.

[22] Amie J. Eisfeld, Gabriele Neumann, and Yoshihiro Kawaoka. At the centre: influenza A virus ribonucleoproteins. *Nature Reviews Microbiology*, 13(1):28–41, January 2015. Number: 1 Publisher: Nature Publishing Group.

[23] Ana M. Falcón, Rosa M. Marión, Thomas Zürcher, Paulino Gómez, Agustín Portela, Amelia Nieto, and Juan Ortín. Defective RNA Replication and Late Gene Expression in Temperature-Sensitive Influenza Viruses Expressing Deleted Forms of the NS1 Protein. *Journal of Virology*, 78(8):3880–3888, April 2004.

[24] Qinshan Gao and Peter Palese. Rewiring the RNAs of influenza virus to prevent reassortment. *Proceedings of the National Academy of Sciences*, 106(37):15891–15896, September 2009.

[25] Todd M Gierahn, Marc H Wadsworth, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, 14(4):395–398, April 2017.

[26] Corrado Gini. Measurement of Inequality of Incomes. *The Economic Journal*, 31(121):124, March 1921.

[27] Julia R. Gog, Emmanuel Dos Santos Afonso, Rosa M. Dalton, India Leclercq, Laurence Tiley, Debra Elton, Johann C. von Kirchbach, Nadia Naffakh, Nicolas Escriou, and Paul Digard. Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic Acids Research*, 35(6):1897–1907, March 2007.

[28] Cait E. Hamele, Alistair B. Russell, and Nicholas S. Heaton. *In Vivo* Profiling of Individual Multiciliated Cells during Acute Influenza A Virus Infection. *Journal of Virology*, 96(14):e00505–22, July 2022.

[29] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

[30] Audray Harris, Giovanni Cardone, Dennis C. Winkler, J. Bernard Heymann, Matthew Brecher, Judith M. White, and Alasdair C. Steven. Influenza virus pleiomorphy characterized by cryoelectron tomography. *Proceedings of the National Academy of Sciences*, 103(50):19123–19127, December 2006. Publisher: Proceedings of the National Academy of Sciences.

[31] A. J. Hay, B. Lomniczi, A. R. Bellamy, and J. J. Skehel. Transcription of the influenza virus genome. *Virology*, 83(2):337–355, December 1977.

[32] Nicholas S. Heaton, Victor H. Leyva-Grado, Gene S. Tan, Dirk Eggink, Rong Hai, and Peter Palese. *In Vivo* Bioluminescent Imaging of Influenza A Virus Infection and Characterization of Novel Cross-Protective Monoclonal Antibodies. *Journal of Virology*, 87(15):8272–8281, August 2013.

[33] Frank S. Heldt, Sascha Y. Kupke, Sebastian Dorl, Udo Reichl, and Timo Frensing. Single-cell analysis and stochastic modelling unveil large cell-to-cell variability in influenza A virus infection. *Nature Communications*, 6(1):8938, December 2015.

[34] Erich Hoffmann, Gabriele Neumann, Gerd Hobom, Robert G. Webster, and Yoshihiro Kawaoka. "Ambisense" Approach for the Generation of Influenza A Virus: vRNA and mRNA Synthesis from One Template. *Virology*, 267(2):310–317, February 2000.

[35] T S Huang, P Palese, and M Krystal. Determination of influenza virus proteins required for genome replication. *Journal of Virology*, 64(11):5669–5673, November 1990. Publisher: American Society for Microbiology.

[36] T S Huang, P Palese, and M Krystal. Determination of influenza virus proteins required for genome replication. *Journal of Virology*, 64(11):5669–5673, November 1990.

[37] Kenrie P. Y. Hui, Rachel H. H. Ching, Stan K. H. Chan, John M. Nicholls, Norman Sachs, Hans Clevers, J. S. Malik Peiris, and Michael C. W. Chan. Tropism, replication competence, and innate immune responses of influenza virus: an analysis of human airway organoids and ex-vivo bronchus cultures. *The Lancet. Respiratory Medicine*, 6(11):846–854, November 2018.

[38] Aida Ibricevic, Andrew Pekosz, Michael J. Walter, Celeste Newby, John T. Battaile, Earl G. Brown, Michael J. Holtzman, and Steven L. Brody. Influenza Virus Receptor Specificity and Cell Tropism in Mouse and Human Airway Epithelial Cells. *Journal of Virology*, 80(15):7469–7480, August 2006.

[39] Nathan T. Jacobs, Nina O. Onuoha, Alice Antia, John Steel, Rustom Antia, and Anice C. Lowen. Incomplete influenza A virus genomes occur frequently but are readily complemented during localized viral spread. *Nature Communications*, 10(1):3526, December 2019.

[40] Michael S. B. Judo, Andrew B. Wedel, and Charles Wilson. Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Research*, 26(7):1819–1825, April 1998.

[41] Benjamin Kaminow, Dinar Yunusov, and Alexander Dobin. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. May 2021.

[42] Jenna N. Kelly, Laura Laloli, Philip V'kovski, Melle Holwerda, Jasmine Portmann, Volker Thiel, and Ronald Dijkman. Comprehensive single cell analysis of pandemic influenza A virus infection in the human airways uncovers cell-type specific host transcriptional signatures relevant for disease progression and pathogenesis. *BioRxiv*, April 2020.

[43] J. Koster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, October 2012.

[44] Melike Lakadamyali, Michael J. Rust, and Xiaowei Zhuang. Endocytosis of influenza viruses. *Microbes and Infection*, 6(10):929–936, August 2004.

[45] Adam S. Lauring and Raul Andino. Exploring the Fitness Landscape of an RNA Virus by Using a Universal Barcode Microarray. *Journal of Virology*, 85(8):3780–3791, April 2011.

[46] Juhye M. Lee, John Huddleston, Michael B. Doud, Kathryn A. Hooper, Nicholas C. Wu, Trevor Bedford, and Jesse D. Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35), August 2018.

[47] Eugenia Leikina, Corinne Ramos, Ingrid Markovic, Joshua Zimmerberg, and Leonid V. Chernomordik. Reversible stages of the low-pH-triggered conformational change in influenza virus hemagglutinin. *The EMBO Journal*, 21(21):5701–5710, November 2002. Publisher: John Wiley & Sons, Ltd.

[48] Katrina A. Lythgoe, Matthew Hall, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-Cockett, Amy Trebes, Monique Andersson, Newton Otecko, Emma L. Wise, Nathan Moore, Jessica Lynch, Stephen Kidd, Nicholas Cortes, Matilde Mori, Rebecca Williams, Gabrielle Vernet, Anita Justice, Angie Green, Samuel M. Nicholls, M. Azim Ansari, Lucie Abeler-Dörner, Catrin E. Moore, Timothy E. A. Peto, David W. Eyre, Robert Shaw, Peter Simmonds, David Buck, John A. Todd, on behalf of the Oxford Virus Sequencing Analysis Group (OVSG), Thomas R. Connor, Shirin Ashraf, Ana da Silva Filipe, James Shepherd, Emma C. Thomson, The COVID-19 Genomics UK (COG-UK) Consortium, David Bonsall, Christophe Fraser, and Tanya Golubchik. SARS-CoV-2 within-host diversity and transmission. *Science*, 372(6539):eabg0821, April 2021.

[49] Glenn A. Marsh, Raheleh Hatami, and Peter Palese. Specific Residues of the Influenza A Virus Hemagglutinin Viral RNA Are Important for Efficient Packaging into Budding Virions. *Journal of Virology*, 81(18):9727–9736, September 2007.

[50] Kelsey Martin and Ari Heleniust. Nuclear transport of influenza virus ribonucleoproteins: The viral matrix protein (M1) promotes export and inhibits import. *Cell*, 67(1):117–130, October 1991.

[51] Michael A. Martin and Katia Koelle. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". *Science Translational Medicine*, 13(617):eabh1803, October 2021.

[52] Mikhail N. Matrosovich, Tatyana Y. Matrosovich, Thomas Gray, Noel A. Roberts, and Hans-Dieter Klenk. Human and avian influenza viruses target different cell types in cultures of human airway epithelium. *Proceedings of the National Academy of Sciences*, 101(13):4620–4624, March 2004.

[53] John T McCrone, Robert J Woods, Emily T Martin, Ryan E Malosh, Arnold S Monto, and Adam S Lauring. Stochastic processes constrain the within and between host evolution of influenza virus. *eLife*, 7:e35962, May 2018.

[54] Andrea Mikulasova, Eva Vareckova, and E Fodor. Transcription and replication of the influenza A virus genome. *Acta virologica*, 44:273–82, November 2000.

[55] Ji-Young Min, Shoudong Li, Ganes C. Sen, and Robert M. Krug. A site on the influenza A virus NS1 protein mediates both inhibition of PKR activation and temporal regulation of viral RNA synthesis. *Virology*, 363(1):236–243, June 2007.

[56] Benjamin Mänz, Linda Brunotte, Peter Reuther, and Martin Schwemmle. Adaptive mutations in NEP compensate for defective H5N1 RNA replication in cultured human cells. *Nature Communications*, 3(1):802, January 2012.

[57] Tadasuke Naito, Kotaro Mori, Hiroshi Ushirogawa, Naoki Takizawa, Eri Nobusawa, Takato Odagiri, Masato Tashiro, Ryosuke L. Ohniwa, Kyosuke Nagata, and Mineki Saito. Generation of a Genetically Stable High-Fidelity Influenza Vaccine Strain. *Journal of Virology*, 91(6):e01073–16, February 2017. Publisher: American Society for Microbiology.

[58] Gabriele Neumann, Tokiko Watanabe, Hiroshi Ito, Shinji Watanabe, Hideo Goto, Peng Gao, Mark Hughes, Daniel R. Perez, Ruben Donis, Erich Hoffmann, Gerd Hobom, and Yoshihiro Kawaoka. Generation of influenza A viruses entirely from cloned cDNAs. *Proceedings of the National Academy of Sciences*, 96(16):9345–9350, August 1999.

[59] S Ohkuma and B Poole. Fluorescence probe measurement of the intralysosomal pH in living cells and the perturbation of pH by various agents. *Proceedings of the National Academy of Sciences*, 75(7):3327–3331, July 1978.

[60] R. E. O'Neill. The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins. *The EMBO Journal*, 17(1):288–296, January 1998.

[61] Alejandro A. Pezzulo, Timothy D. Starner, Todd E. Scheetz, Geri L. Traver, Ann E. Tilley, Ben-Gary Harvey, Ronald G. Crystal, Paul B. McCray, and Joseph Zabner. The air-liquid interface and use of primary cell cultures are important to recapitulate the transcriptional profile of in vivo airway epithelia. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 300(1):L25–L31, January 2011. Publisher: American Physiological Society.

[62] Thu Phan, Elizabeth J. Fay, Zion Lee, Stephanie Aron, Wei-Shou Hu, and Ryan A. Langlois. Segment-Specific Kinetics of mRNA, cRNA, and vRNA Accumulation during Influenza Virus Infection. *Journal of Virology*, 95(10):e02102–20, April 2021.

[63] Kara L. Phipps, Ketaki Ganti, Nathan T. Jacobs, Chung-Young Lee, Silvia Carnaccini, Maria C. White, Miglena Manandhar, Brett E. Pickett, Gene S. Tan, Lucas M. Ferreri, Daniel R. Perez, and Anice C. Lowen. Collective interactions augment influenza A virus replication in a host-dependent manner. *Nature Microbiology*, 5(9):1158–1169, September 2020.

[64] Simone Picelli. Full-Length Single-Cell RNA Sequencing with Smart-seq2. In Valentina Proserpio, editor, *Single Cell Methods: Sequencing and Proteomics*, Methods in Molecular Biology, pages 25–44. Springer, New York, NY, 2019.

[65] S Pleschka, R Jaskunas, O G Engelhardt, T Zürcher, P Palese, and A García-Sastre. A plasmid-based reverse genetics system for influenza A virus. *Journal of Virology*, 70(6):4188–4192, June 1996.

[66] Qiagen. RNeasy Mini Handbook.

[67] Neal G. Ravindra, Mia Madel Alfajaro, Victor Gasque, Victoria Habet, Jin Wei, Renata B. Filler, Nicholas C. Huston, Han Wan, Klara Szigeti-Buck, Bao Wang, Guilin Wang, Ruth R. Montgomery, Stephanie C. Eisenbarth, Adam Williams, Anna Marie Pyle, Akiko Iwasaki, Tamas L. Horvath, Ellen F. Foxman, Richard W. Pierce, David van Dijk, and Craig B. Wilen. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium, July 2020. Pages: 2020.05.06.081695 Section: New Results.

[68] Neal G. Ravindra, Mia Madel Alfajaro, Victor Gasque, Victoria Habet, Jin Wei, Renata B. Filler, Nicholas C. Huston, Han Wan, Klara Szigeti-Buck, Bao Wang, Guilin Wang, Ruth R. Montgomery, Stephanie C. Eisenbarth, Adam Williams, Anna Marie Pyle, Akiko Iwasaki, Tamas L. Horvath, Ellen F. Foxman, Richard W. Pierce, David van Dijk, and Craig B. Wilen. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium. May 2020.

[69] Kent A Riemondy, Monica Ransom, Christopher Alderman, Austin E Gillen, Rui Fu, Jessica Finlay-Schultz, Gregory D Kirkpatrick, Jorge Di Paola, Peter Kabos, Carol A Sartorius, and Jay R Hesselberth. Recovery and analysis of transcriptome subsets from pooled single-cell RNA-seq libraries. *Nucleic Acids Research*, 47(4):e20, February 2019.

[70] Nicole C. Robb, Geoffrey Chase, Katja Bier, Frank T. Vreede, Pang-Chui Shaw, Nadia Naffakh, Martin Schwemmle, and Ervin Fodor. The Influenza A Virus NS1 Protein Interacts with the Nucleoprotein of Viral Ribonucleoprotein Complexes. *Journal of Virology*, 85(10):5228–5231, May 2011.

[71] Nicole C. Robb, Matt Smith, Frank T. Vreede, and Ervin Fodor. NS2/NEP protein regulates transcription and replication of the influenza virus RNA genome. *Journal of General Virology*, 90(6):1398–1407, June 2009.

[72] Jeremy S. Rossman and Robert A. Lamb. Influenza virus assembly and budding. *Virology*, 411(2):229–236, March 2011.

[73] Alistair B. Russell, Elizaveta Elshina, Jacob R. Kowalsky, Aartjan J. W. te Velthuis, and Jesse D. Bloom. Single-Cell Virus Sequencing of Influenza Infections That Trigger Innate Immunity. *Journal of Virology*, 93(14):e00500–19, July 2019.

[74] Alistair B Russell, Cole Trapnell, and Jesse D Bloom. Extreme heterogeneity of influenza virus infection in single cells. *eLife*, 7:e32303, February 2018.

[75] Kazima Saira, Xudong Lin, Jay V. DePasse, Rebecca Halpin, Alan Twaddle, Timothy Stockwell, Brian Angus, Alessandro Cozzi-Lepri, Marina Delfino, Vivien Dugan, Dominic E. Dwyer, Matthew Freiberg, Andrzej Horban, Marcelo Losso, Ruth Lynfield, Deborah N. Wentworth, Edward C. Holmes, Richard Davey, David E. Wentworth, and Elodie Ghedin. Sequence Analysis of *In Vivo* Defective Interfering-Like RNA of Influenza A H1N1 Pandemic Virus. *Journal of Virology*, 87(14):8064–8074, July 2013.

[76] Rafael Sanjuán. Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1548):1975–1982, June 2010. Publisher: Royal Society.

[77] Michael B. Schulte and Raul Andino. Single-Cell Analysis Uncovers Extensive Biological Noise in Poliovirus Replication. *Journal of Virology*, 88(11):6205–6212, June 2014.

[78] Jessica R. Shartouny, Chung-Young Lee, Gabrielle K. Delima, and Anice C. Lowen. Beneficial effects of cellular coinfection resolve inefficiency in influenza A virus transcription. *PLOS Pathogens*, 18(9):e1010865, September 2022. Publisher: Public Library of Science.

[79] Anna Sims, Laura Burgess Tornaletti, Seema Jasim, Chiara Pirillo, Ryan Devlin, Jack Hirst, Colin Loney, Joanna Wojtus, Elizabeth Sloan, Luke Thorley, Chris Boutell, Edward Roberts, and Edward Hutchinson. Superinfection exclusion creates spatially distinct influenza virus populations. June 2022.

[80] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3):491–499, March 2017.

[81] Emily Speranza, Brandi N. Williamson, Friederike Feldmann, Gail L. Sturdevant, Lizzette Pérez-Pérez, Kimberly Meade-White, Brian J. Smith, Jamie Lovaglio, Craig Martens, Vincent J. Munster, Atsushi Okumura, Carl Shaia, Heinz Feldmann, Sonja M. Best, and Emmie de Wit. Single-cell RNA sequencing reveals SARS-CoV-2 infection dynamics in lungs of African green monkeys. *Science Translational Medicine*, 13(578):eabe8146, January 2021. Publisher: American Association for the Advancement of Science.

[82] David A. Steinhauer, Esteban Domingo, and John J. Holland. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene*, 122(2):281–288, December 1992.

[83] Yael Steuerman, Merav Cohen, Naama Peshes-Yaloz, Liran Valadarsky, Ofir Cohn, Eyal David, Amit Frishberg, Lior Mayo, Eran Bacharach, Ido Amit, and Irit Gat-Viks. Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing. *Cell Systems*, 6(6):679–691.e4, June 2018.

[84] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, September 2017. Number: 9 Publisher: Nature Publishing Group.

[85] Jiayi Sun, J. Cristobal Vera, Jenny Drnevich, Yen Ting Lin, Ruian Ke, and Christopher B. Brooke. Single cell heterogeneity in influenza A virus gene expression shapes the innate antiviral response to infection. *PLOS Pathogens*, 16(7):e1008671, July 2020.

[86] Aartjan J. W. te Velthuis and Ervin Fodor. Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis. *Nature Reviews Microbiology*, 14(8):479–493, August 2016.

[87] ThermoFisher. SuperScript™ III First-Strand Synthesis SuperMix for qRT-PCR.

[88] Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3:e03300, July 2014. Publisher: eLife Sciences Publications, Ltd.

[89] John J. Trombetta, David Gennert, Diana Lu, Rahul Satija, Alex K. Shalek, and Aviv Regev. Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Current Protocols in Molecular Biology*, 107(1):4.22.1–4.22.17, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb0422s107.

[90] Po-Yuan Tung, John D. Blischak, Chiaowen Joyce Hsiao, David A. Knowles, Jonathan E. Burnett, Jonathan K. Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, 7(1):39921, January 2017. Number: 1 Publisher: Nature Publishing Group.

[91] Andrew Varble, Randy A. Albrecht, Simone Backes, Marshall Crumiller, Nicole M. Bouvier, David Sachs, Adolfo García-Sastre, and Benjamin R. tenOever. Influenza A Virus Transmission Bottlenecks Are Defined by Infection Route and Recipient Host. *Cell Host & Microbe*, 16(5):691–700, November 2014.

[92] Preben von Magnus. Incomplete Forms of Influenza Virus. In *Advances in Virus Research*, volume 2, pages 59–79. Elsevier, 1954.

[93] Joanna M. Wandzik, Tomas Kouba, Manikandan Karuppasamy, Alexander Pflug, Petra Drncova, Jan Provaznik, Nayara Azevedo, and Stephen Cusack. A Structure-Based Model for the Complete Transcription Cycle of Influenza Polymerase. *Cell*, 181(4):877–893.e21, May 2020.

[94] Chang Wang, Christian V. Forst, Tsui-wen Chou, Adam Geber, Minghui Wang, Wissam Hamou, Melissa Smith, Robert Sebra, Bin Zhang, Bin Zhou, and Elodie Ghedin. Cell-to-Cell Variation in Defective Virus Expression and Effects on Host Responses during Influenza Virus Infection. *mBio*, 11(1):e02880–19, February 2020.

[95] Tokiko Watanabe, Shinji Watanabe, Takeshi Noda, Yutaka Fujii, and Yoshihiro Kawaoka. Exploitation of Nucleic Acid Packaging Signals To Generate a Novel Influenza Virus-Based Vector Stably Expressing Two Foreign Genes. *Journal of Virology*, 77(19):10575–10583, October 2003.

[96] W. Weis, J. H. Brown, S. Cusack, J. C. Paulson, J. J. Skehel, and D. C. Wiley. Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature*, 333(6172):426–431, June 1988. Number: 6172 Publisher: Nature Publishing Group.

[97] Katherine S. Xue and Jesse D. Bloom. Reconciling disparate estimates of viral genetic diversity during human influenza infections. *Nature Genetics*, 51(9):1298–1301, September 2019.

[98] Tracy M. Yamawaki, Daniel R. Lu, Daniel C. Ellwanger, Dev Bhatt, Paolo Manzanillo, Vanessa Arias, Hong Zhou, Oh Kyu Yoon, Oliver Homann, Songli Wang, and Chi-Ming Li. Systematic comparison of high-throughput single-cell RNA-seq methods for immune cell profiling. *BMC Genomics*, 22(1):66, January 2021.

[99] Matthew D Young and Sam Behjati. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, 9(12), November 2020.

[100] Emilio Yángüez and Amelia Nieto. So similar, yet so different: Selective translation of capped and polyadenylated viral mRNAs in the influenza virus infected cell. *Virus Research*, 156(1):1–12, March 2011.

[101] Fabio Zanini, Szu-Yuan Pu, Elena Bekerman, Shirit Einav, and Stephen R Quake. Single-cell transcriptional dynamics of flavivirus infection. *eLife*, 7:e32942, February 2018.

[102] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, January 2017.

Appendix A

SUPPLEMENTARY FIGURES

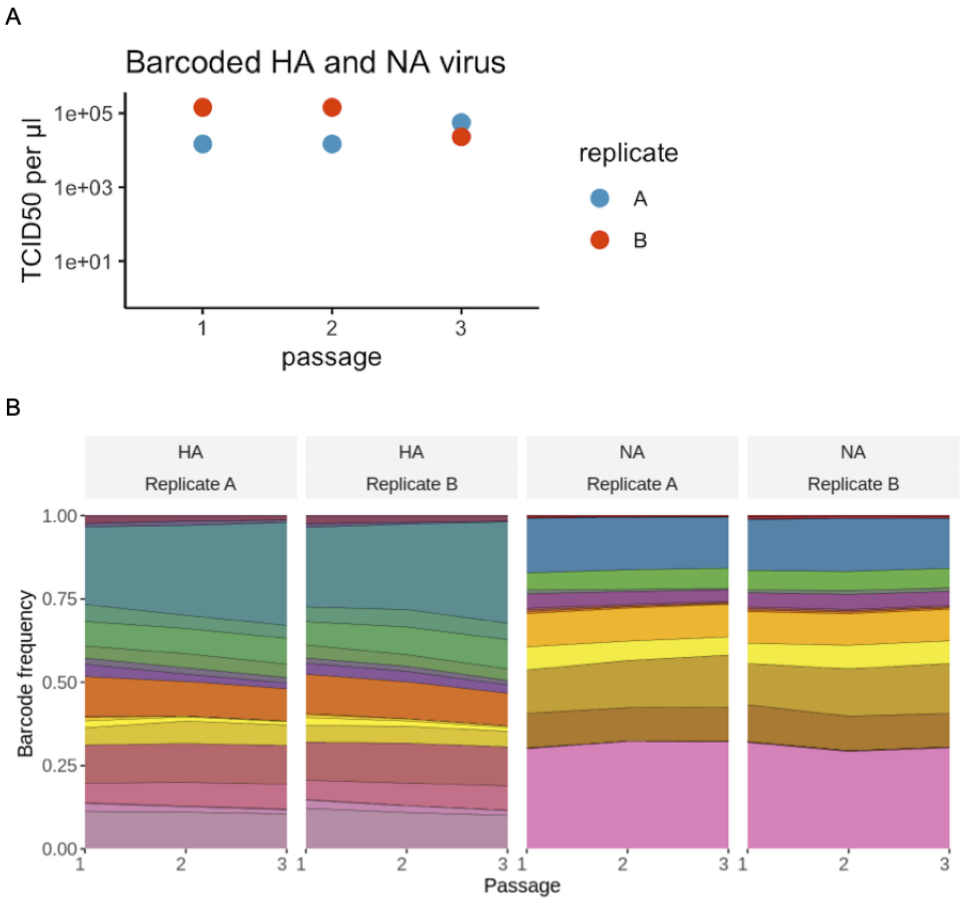


Figure A.1: Viral barcode sequences are selectively neutral. Influenza virus carrying a pool of HA and NA barcodes was generated by reverse genetics and passaged 3 times at low MOI. (A) The titers were measured after each growth step by TCID50. (B) The frequency of each barcode in the viral population was measured by deep sequencing after each growth step. Each color represents a unique viral barcode. The frequencies of viral barcodes were fairly consistent across passages, indicating a lack of selection for any particular barcode sequence. The viral barcode frequencies were calculated using the code at https://github.com/dbacsik/barcode_neutrality.

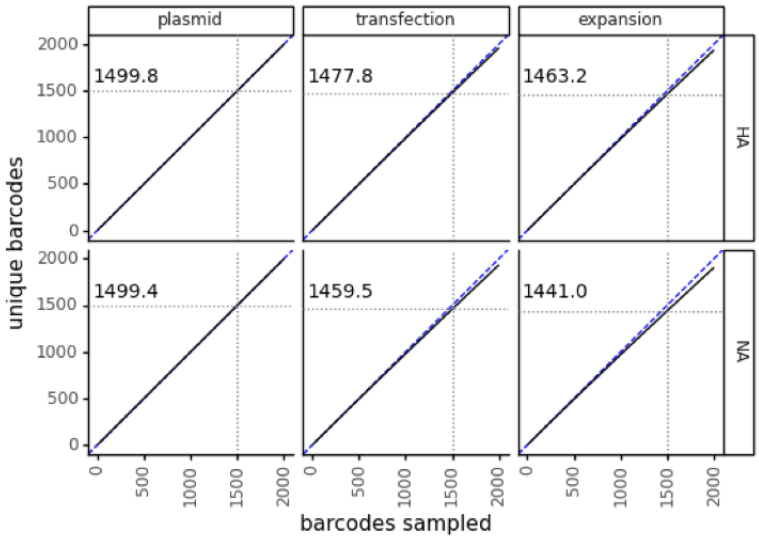


Figure A.2: Extremely diverse barcoded virus libraries. Rarefaction curves show the diversity of the viral barcodes. The x-axis represents the number of barcodes sampled. The y-axis represents the number of sampled barcodes which are unique. A theoretical perfect library where every barcode is unique appears as a straight line with formula x=y and is shown here with a blue dashed line. Our experiments used approximately 1500 virions per sample. The number of unique barcodes in a sample of 1500 is annotated in the top left of each facet. The rarefaction curves were calculated using https://jbloomlab.github.io/dms_variants/dms_variants.barcodes.html?highlight=rarefybarcodes#dms_variants.barcodes.rarefyBarcodes.

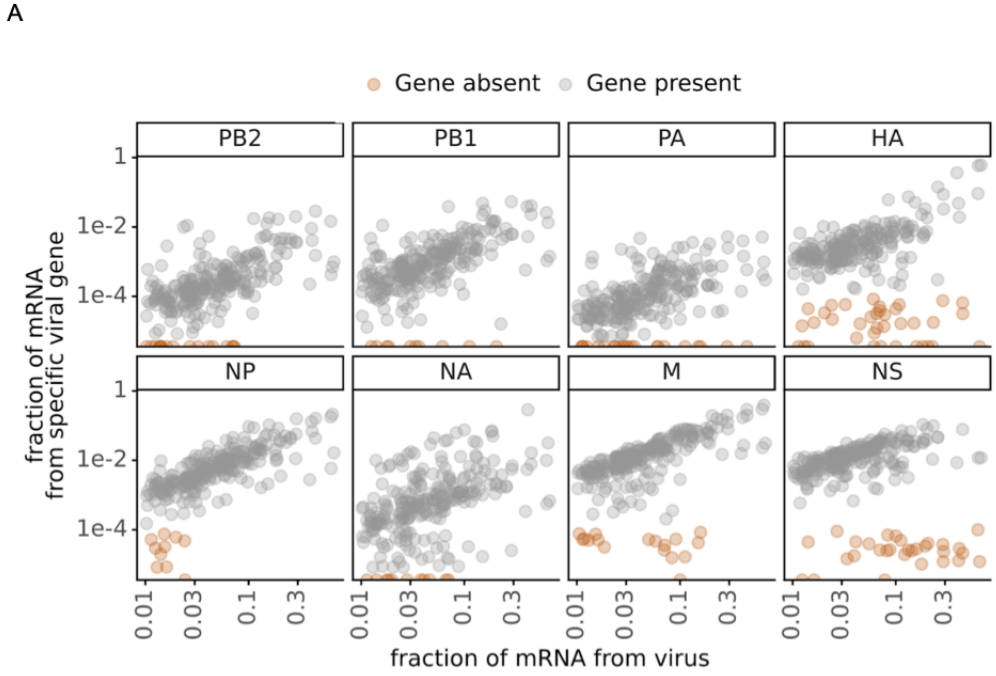


Figure A.3: Expression of viral genes in infected cells. This plot shows single-cell RNA-sequencing data for 254 infected cells. (A) Total viral transcription and expression of each viral gene in single infected cells. Genes with low average transcript counts in the single-cell RNA sequencing data (PB2, PB1, PA, and NA) are called as absent if there are zero transcripts detected in a cell. Genes with higher average transcript counts in this data (HA, NP, M, and NS) are called as absent if their abundance falls at or below the 99th percentile observed in uninfected cells. Low, non-zero transcript counts for these genes most likely result from transcripts leaking from one oil droplet to another during single-cell RNA sequencing [99]

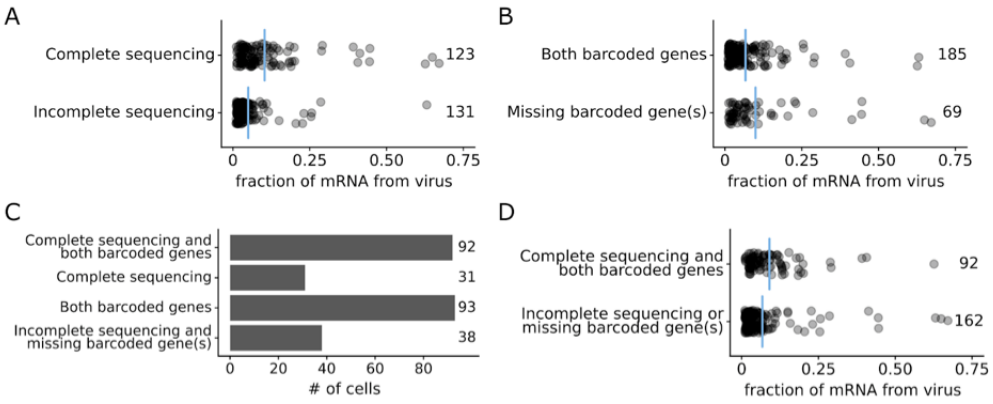
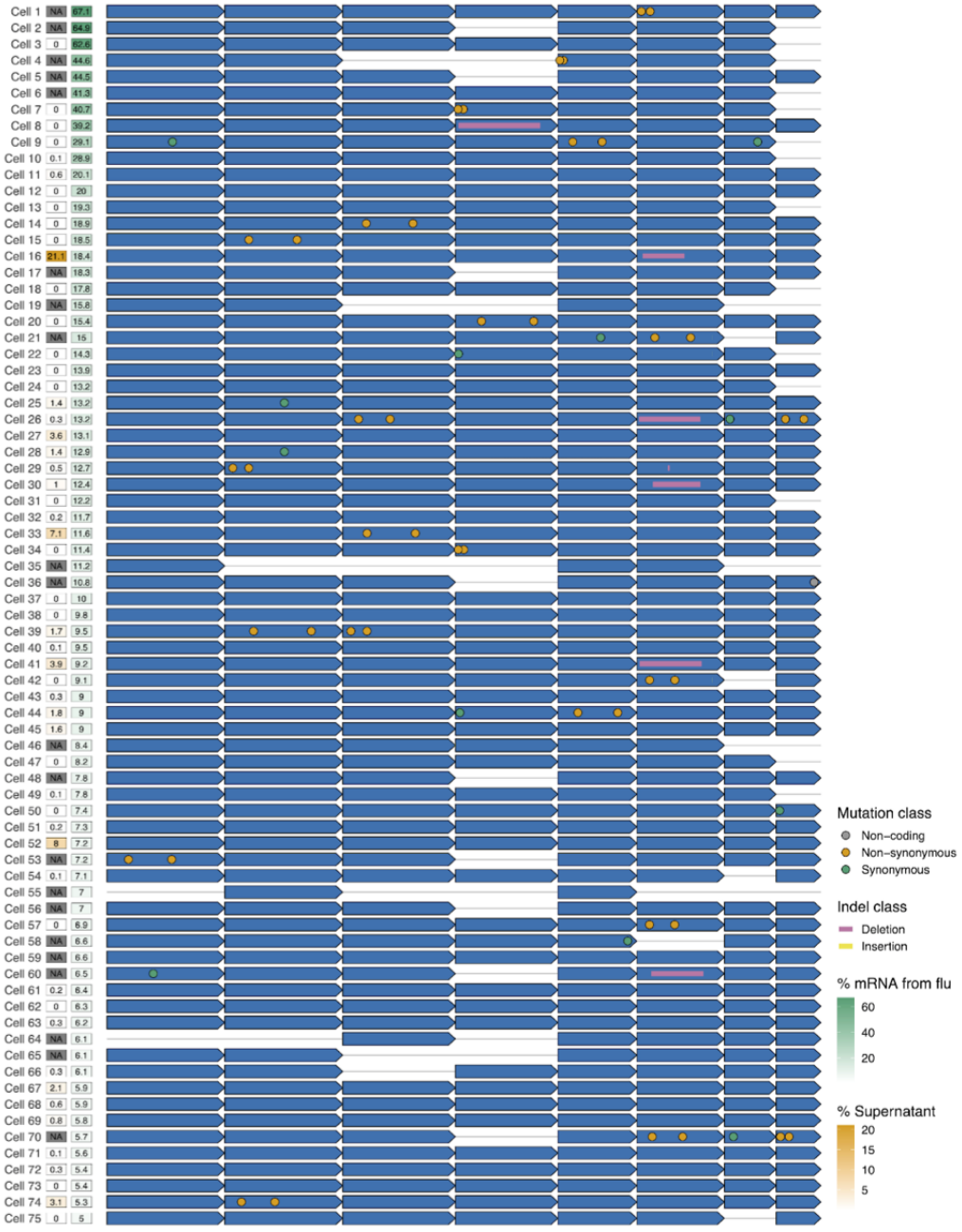


Figure A.4: Statistics of viral genome sequencing and progeny measurements from infected cells. In our dataset, 254 infected cells were identified by single-cell RNA sequencing. (A) Of these, 123 infected cells had complete PacBio long-read sequencing data for every expressed viral gene. On average, cells with complete sequencing coverage had higher viral transcription than cells without complete sequencing coverage. Each point indicates a cell, and blue lines indicate the mean. (B) 185 infected cells expressed both barcoded viral genes (HA and NA). Cells that expressed both barcoded genes had lower viral transcription than cells that were missing one or more barcoded gene(s). (C) The number of cells with complete sequencing and both barcoded viral genes, only complete sequencing, only both barcoded viral genes, or neither complete sequencing nor both barcoded viral genes. (D) 92 infected cells had long-read sequencing of all expressed viral genes and expressed both barcoded viral genes. On average, cells with all measurements had slightly higher viral transcription than cells without all measurements.


Cell 1 NA 67.1
Cell 2 NA 64.8
Cell 3 0 62.6
Cell 4 NA 44.6
Cell 5 NA 44.5
Cell 6 NA 41.3
Cell 7 0 40.7
Cell 8 0 39.2
Cell 9 0 29.1
Cell 10 0.1 28.9
Cell 11 0.6 20.1
Cell 12 0 20
Cell 13 0 19.3
Cell 14 0 18.9
Cell 15 0 18.5
Cell 16 21.1 18.4
Cell 17 NA 18.3
Cell 18 0 17.8
Cell 19 NA 15.8
Cell 20 0 15.4
Cell 21 NA 15
Cell 22 0 14.3
Cell 23 0 13.9
Cell 24 0 13.2
Cell 25 1.4 13.2
Cell 26 0.3 13.2
Cell 27 3.6 13.1
Cell 28 1.4 12.9
Cell 29 0.5 12.7
Cell 30 1 12.4
Cell 31 0 12.2
Cell 32 0.2 11.7
Cell 33 7.1 11.6
Cell 34 0 11.4
Cell 35 NA 11.2
Cell 36 NA 10.8
Cell 37 0 10
Cell 38 0 9.8
Cell 39 1.7 9.5
Cell 40 0.1 9.5
Cell 41 3.9 9.2
Cell 42 0 9.1
Cell 43 0.3 9
Cell 44 1.8 9
Cell 45 1.6 9
Cell 46 NA 8.4
Cell 47 0 8.2
Cell 48 NA 7.8
Cell 49 0.1 7.8
Cell 50 0 7.4
Cell 51 0.2 7.3
Cell 52 8 7.2
Cell 53 NA 7.2
Cell 54 0.1 7.1
Cell 55 NA 7
Cell 56 NA 7
Cell 57 0 6.9
Cell 58 NA 6.6
Cell 59 NA 6.6
Cell 60 NA 6.5
Cell 61 0.2 6.4
Cell 62 0 6.3
Cell 63 0.3 6.2
Cell 64 NA 6.1
Cell 65 NA 6.1
Cell 66 0.3 6.1
Cell 67 2.1 5.9
Cell 68 0.6 5.9
Cell 69 0.8 5.8
Cell 70 NA 5.7
Cell 71 0.1 5.6
Cell 72 0.3 5.4
Cell 73 0 5.4
Cell 74 3.1 5.3
Cell 75 0 5
Mutation class
Non−coding
Non−synonymous
Synonymous
Indel class
Deletion
Insertion
% mRNA from flu
60
40
20
% Supernatant
20
15
10
5

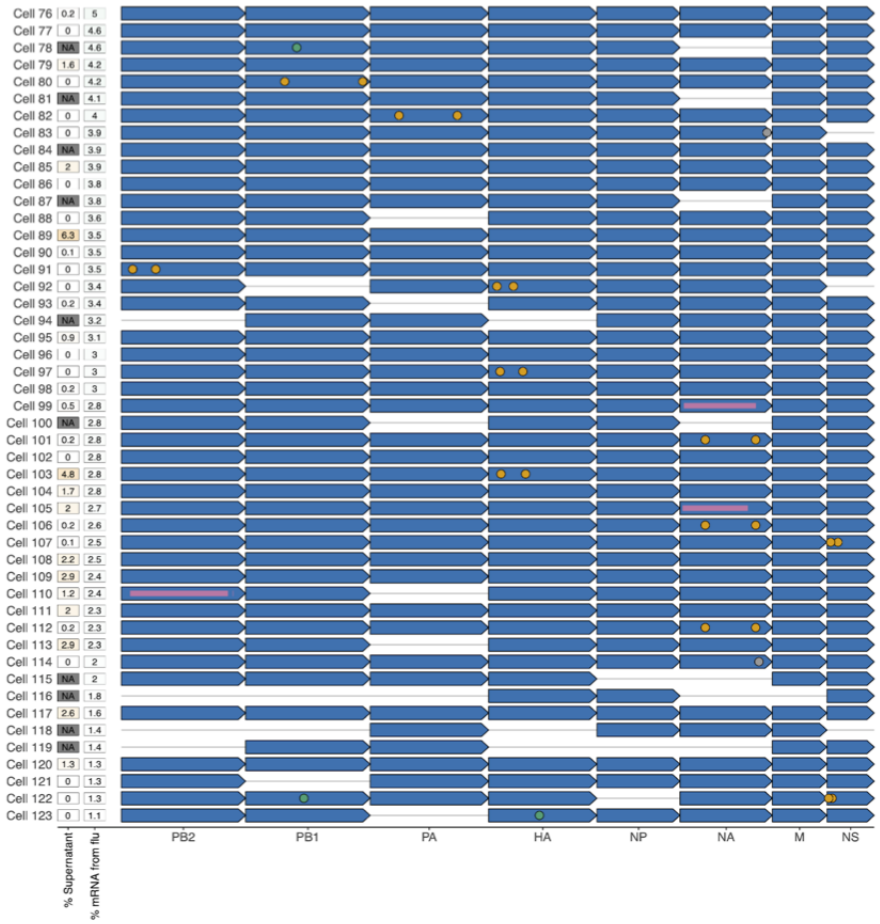


Figure A.5: Viral genotypes in infected cells. The sequence of the infecting virion for the 123 infected cells for which we could determine the sequence of all expressed viral genes. Each infected cell is represented as a row and each viral transcript is represented as an arrow. Missing viral genes, insertions, deletions, and mutations are annotated on the arrows. Viral transcription (as a fraction of UMIs in the cell), and viral progeny production (as a fraction of the physical progeny virions in the supernatant) are shown for each infected cell. Cells with one or more missing barcoded viral genes have "NA" values listed for progeny production. A high-resolution version of this figure is available at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/results/figures/viral_genomes_plot.pdf.

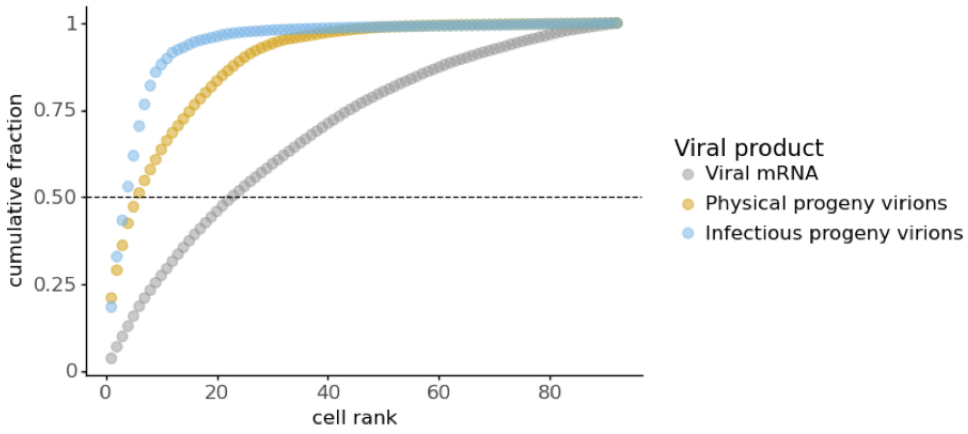


Figure A.6: Cumulative fraction of viral products produced by single infected cells. For the viral mRNA values, the y-axis represents each cell's contribution to the total viral mRNA transcripts across all cells. For the progeny values, the y-axis represents each cell's contribution to the barcodes in the supernatant or second infection that are assignable to one of the infected cells. A horizontal line is drawn at y=0.5 to indicate the minimum number of cells that generated half of the total amount of each viral product. This plot shows the 92 single infected cells for which we could identify the viral barcode on both barcoded genes and determine the sequence of all genes expressed by the infecting virion.

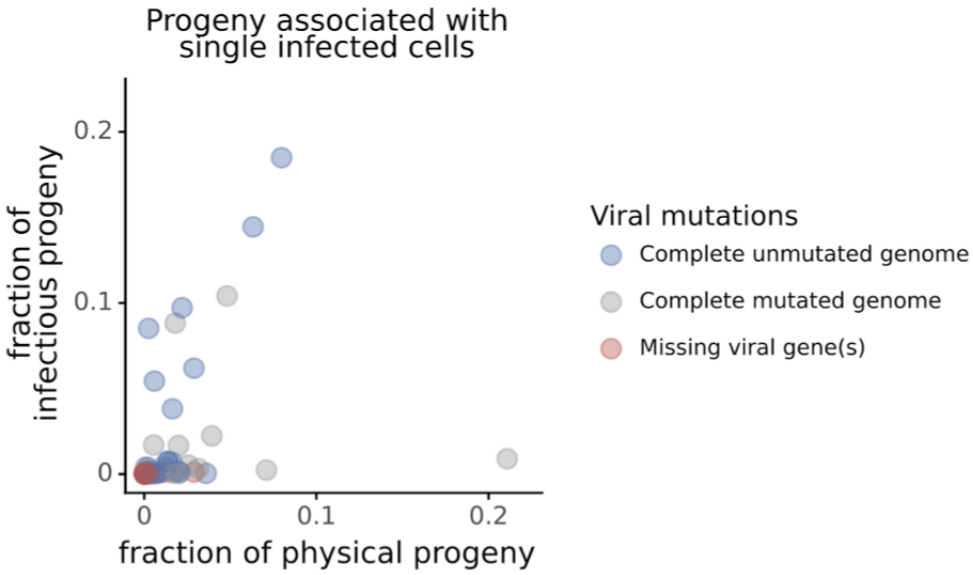


Figure A.7: Frequency of physical progeny and infectious progeny from single infected cells. Each point represents a single infected cell. The x-axis represents the fraction of physical progeny generated by each cell. The y-axis represents the fraction of infectious progeny generated by each cell. This plot shows the 92 cells for which we could identify the viral barcode on both barcoded genes and determine the sequence of all genes expressed by the infecting virion.