Influenza virus transcription and progeny production are poorly correlated in single cells

David J. Bacsik[1,2], Bernadeta Dadonaite[1], Andrew Butler[1], Allison J. Greaney[1,2], Nicholas S. Heaton[3,4], Jesse D. Bloom[1,5,#]

Affiliations:

[1]Basic Sciences Division and Computational Biology Program, Fred Hutchinson Cancer Center, Seattle, Washington, United States of America

[2]Department of Genome Sciences & Medical Scientist Training Program, University of Washington, Seattle, Washington, United States of America

[3]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, North Carolina, United States of America

[4]Duke Human Vaccine Institute, Duke University School of Medicine, Durham, North Carolina, United States of America

[5]Howard Hughes Medical Institute, Chevy Chase, Maryland, United States of America

[#]Lead Contact jbloom@fredhutch.org

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.

Introduction

Many aspects of viral infection are extremely heterogeneous when measured across single cells. Individual infected cells vary widely in transcription of viral genes [1–3], presence of viral mutations [4], expression of viral proteins [5], replication of viral genomes [6], and production of viral progeny [7,8]. 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.

Results

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 [9–11] into the influenza virus genome so that they are positioned near the 3′ end of the viral mRNAs (Fig 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 [1,2,4,12–14] captures the viral barcode sequence along with host and viral transcripts, enabling determination of which barcoded virion infected each cell (Fig 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 [4]. This strategy enables simultaneous measurement of transcription, progeny production, and viral genotype in single cells.

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 [15,16]. 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 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 [17]. 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 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 S1).

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 from our virus library, >96% of barcodes were unique (Fig S2).

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

We implemented the experiment in Fig 1A by infecting ~10^4 MDCK-SIAT1-TMPRSS2 cells [18] 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 [19,20]. 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 ~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 [21,22].

Among the infected cells, there was extremely wide variation in the amount of viral transcription (Fig 2A), similar to that observed in prior single-cell studies of influenza infection [1,2,4,12,13]. 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 2A). A substantial fraction (~40%) of the infected cells also failed to express all eight viral genes (Fig 2B), a phenomenon that has been extensively described in prior studies [1,2,5]. 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 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 S3 and Methods), which could reduce our ability to detect the absence of the four viral genes involved in transcription (PB2, PB1, PA, NP).

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 [4]. 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 that produced it.

We obtained complete sequences of all expressed viral genes for 123 of the 254 infected cells in our dataset (Fig S4A). About a third of the infected cells expressed all eight viral genes without any non-synonymous mutations (Fig 3A, Fig S5). The remainder of infected cells failed to express a viral gene, expressed a gene with a non-synonymous mutation, or both (Fig 3A). Mutated virions most commonly had just one non-synonymous mutation in their genome, but some virions had two or three mutations (Fig 3B, Fig S5). Note that some virions had large internal deletions in a gene (Fig S5) as has been previously described [23,24]; here we have classified deletions as non-synonymous mutations.

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 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 the 92 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 S4).

The number of progeny virions produced per cell was extremely heterogeneous (Fig 4A,B). Nearly half of infected cells failed to produce any detectable physical or infectious progeny, and just a few cells produced >10% of all progeny generated by the 92 infected cells analyzed (Fig 4, S6). Progeny production is notably more heterogeneous across single cells than viral transcription. We can quantify this heterogeneity by calculating a Gini coefficient, which can range from zero to one with larger values indicating more uneven distributions [25]. The Gini coefficients are 0.78 and 0.88 for physical and infectious progeny production versus 0.46 for viral transcription (Fig 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 S6).

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 5A). None of the cells with >25% of their mRNA transcripts derived from virus produced any detectable progeny (Fig 5A). Instead, the progeny come from 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 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 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 5A, Fig S7)—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 5A versus Fig S7; 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 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 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 [26,27] and possibly act as a switch from transcription to genome replication [28,29]. 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 5C,D). This fact suggests other cellular or unknown viral factors must also contribute to cell-to-cell variation in progeny production.

Discussion

Most prior single-cell studies of infection have examined intracellular viral products, like mRNA transcripts or proteins [1–5]. However, the most important outcome for multi-cycle viral growth is how many progeny virions are produced by an infected cell. The only prior studies of progeny production from single cells have relied on isolating individual infected cells in small volumes [6–8]. These studies have shown progeny production is highly heterogeneous, but have not provided measurements of most other properties of the infected cells—and have therefore lacked explanatory power to understand the basis for the variation in progeny production. Here, we have overcome these limitations with a new method to simultaneously measure physical and infectious progeny production alongside transcription of all viral and host genes, as well as sequencing of the genome of the virion that infected each cell.

The most striking finding from our work is that while viral transcription and progeny production are both highly heterogeneous across influenza-infected cells, they are not well correlated. In particular, the cells that transcribe the most viral mRNA generate no detectable progeny virions. Part of this discrepancy between these two single-cell measurements is due to the fact that cells that fail to express the influenza NS gene tend to transcribe very high levels of mRNA but produce no progeny. This result makes biological sense: one of the proteins encoded by the NS gene is NEP, which exports viral ribonucleoproteins from the nucleus [26], which terminates their transcriptional activity [27]. NEP may also mediate a switch from transcription to genome replication by the viral polymerase [28,29]. However, absence of NS and other viral genetic defects only explain part of the discordance between viral transcription and progeny production in single cells, since these two properties are often discordant even in cells expressing unmutated copies of all viral genes. We suggest that further study of both viral and cellular factors that promote transcription versus progeny production is an interesting area for future work.

Our study has several limitations. Our experiments used a cell line rather than the differentiated airway cells that are the actual target cells during human influenza infections. We were also able to profile all relevant single-cell properties for only a relatively modest number of infected cells. However, the method we have described should be extensible to other cell-culture systems and larger numbers of cells in future work.

Despite these limitations, our results suggest several important implications for broader thinking about viral infections. First, several recent studies have examined the distribution of influenza virus transcription or protein expression across differentiated airway cells *ex vivo* [22,30,31] or *in vivo* [13,14,32]. Our results suggest it is also important to study progeny production across airway cells, as the cell types expressing the most viral transcripts or proteins may not be the ones producing the most viral progeny. Second, our results suggest failure of some cells to express specific viral genes contributes to the discordance between viral

transcription and progeny production. It is interesting to speculate if this phenomenon could be reduced under high MOI conditions that promote genetic complementation [33–35]. Third, recent work on the transmission of influenza virus [36,37] and SARS-CoV-2 [38–40] in humans has emphasized the narrow genetic bottleneck, with only a small fraction of viral diversity in the donor transmitted to the recipient. Our results suggest physical bottlenecks in how many virions reach the recipient may be further narrowed by the fact that only a small fraction of the initially infected cells will produce most of the progeny that continue the infection.

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 1A and 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.

Methods

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 [15,16], as schematized in Fig 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 [17]. Packaging signal length and location was informed by previous studies [41–43]. The terminal 105 nucleotides of the HA coding sequence were duplicated to provide an authentic 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.

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.

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 µg per mL streptomycin). After ~16 hours, we transfected each well with bidirectional reverse-genetics plasmids based on the pHW2000 vector [44] carrying the six unmodified segments: (PB2, PB1, PA, NP, M, and NS), unidirectional reverse-genetics plasmids based on the pHH21 vector [45] carrying the two barcoded segments (HA and NA), and a plasmid constitutively expressing the TMPRSS protease (which proteolytically activates HA) [18]. 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 µg per mL of calcium chloride, 100 U per mL penicillin, and 100 µg per mL streptomycin) and 3e5 MDCK-SIAT1-TMPRSS2 cells [18] 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 [18] 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 1C).

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) [46]. 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 [4] 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" [47]. This phenomenon can be detected in long-read viral sequences that contain discordant genetic tags (see Fig S10 of [4]). 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.

Infecting cells with a dual-barcoded virus library

Ten thousand MDCK-SIAT1-TMPRSS2 [18] 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 [19,20]) 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.

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 [48] 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.

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 [4]. 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 [4]), 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.

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 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 [18] at a moderate estimated MOI of ~0.25 in two independent replicates. To infect the cells, 60,000 MDCK-SIAT1-TMPRSS2 cells [18] 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 [19,20] 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 [49]. 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 [50]. 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).

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 [51]. 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 [52] 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 S3).

For the statistical test in Fig 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 [53]. 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_variants package (https://jbloomlab.github.io/dms_variants/). The viral barcodes were error-corrected using UMI-tools [54]. 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 S8), 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.

Data 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 be available on GEO. The single-cell RNA sequencing cell-gene matrix will also be available on GEO. GEO submission in progress; accession will be added to GitHub README at https://github.com/jbloomlab/barcoded_flu_pdmH1N1/blob/main/README.md as soon as it's available. 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.

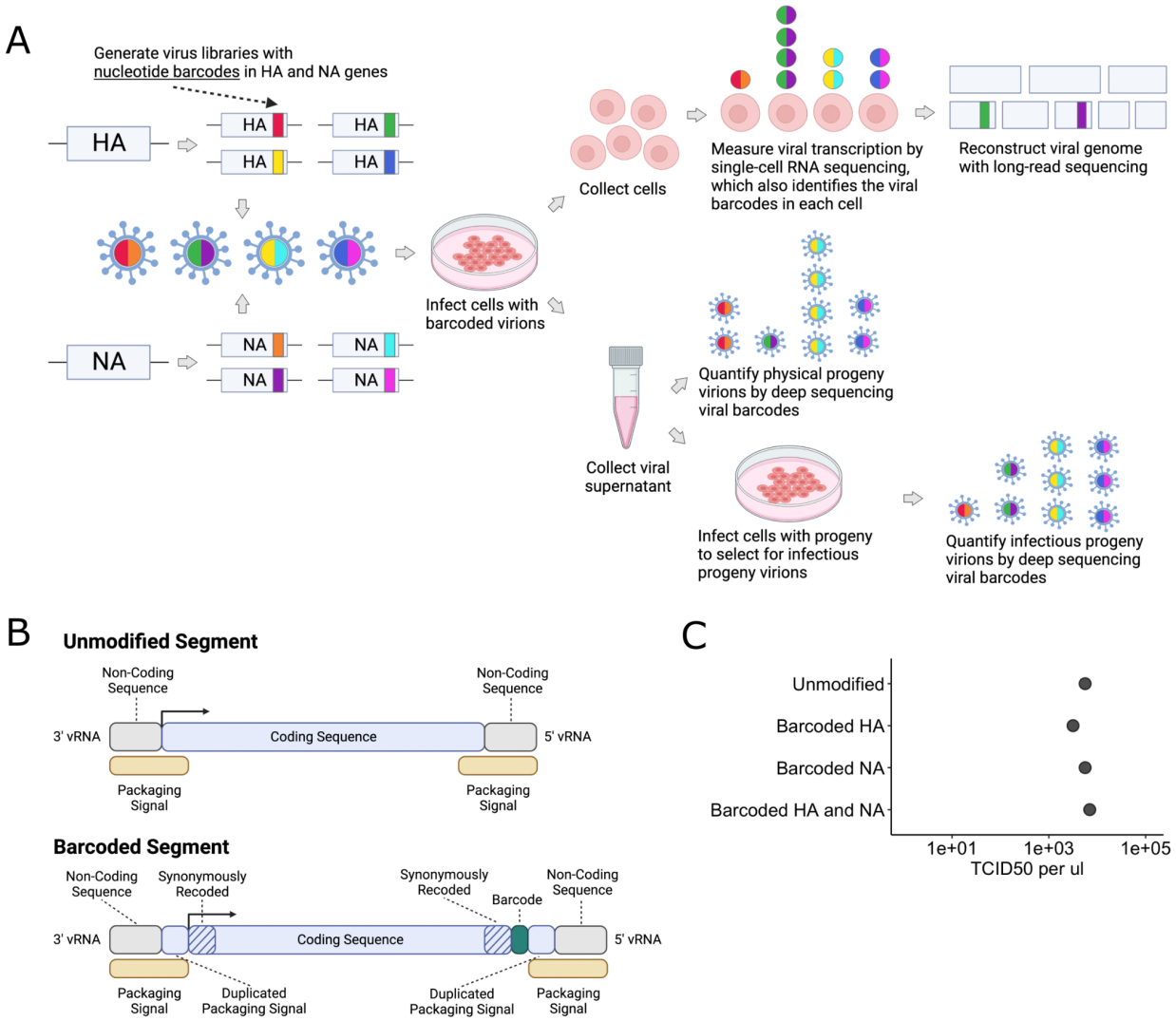


Fig. 1. Strategy to measure transcription, progeny production, and viral genotype in single cells. (A) Insertion of barcodes in the viral genome makes it possible to quantify the progeny released from single cells, and relate progeny production to viral transcription and viral genotype. **(B)** Barcodes 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.

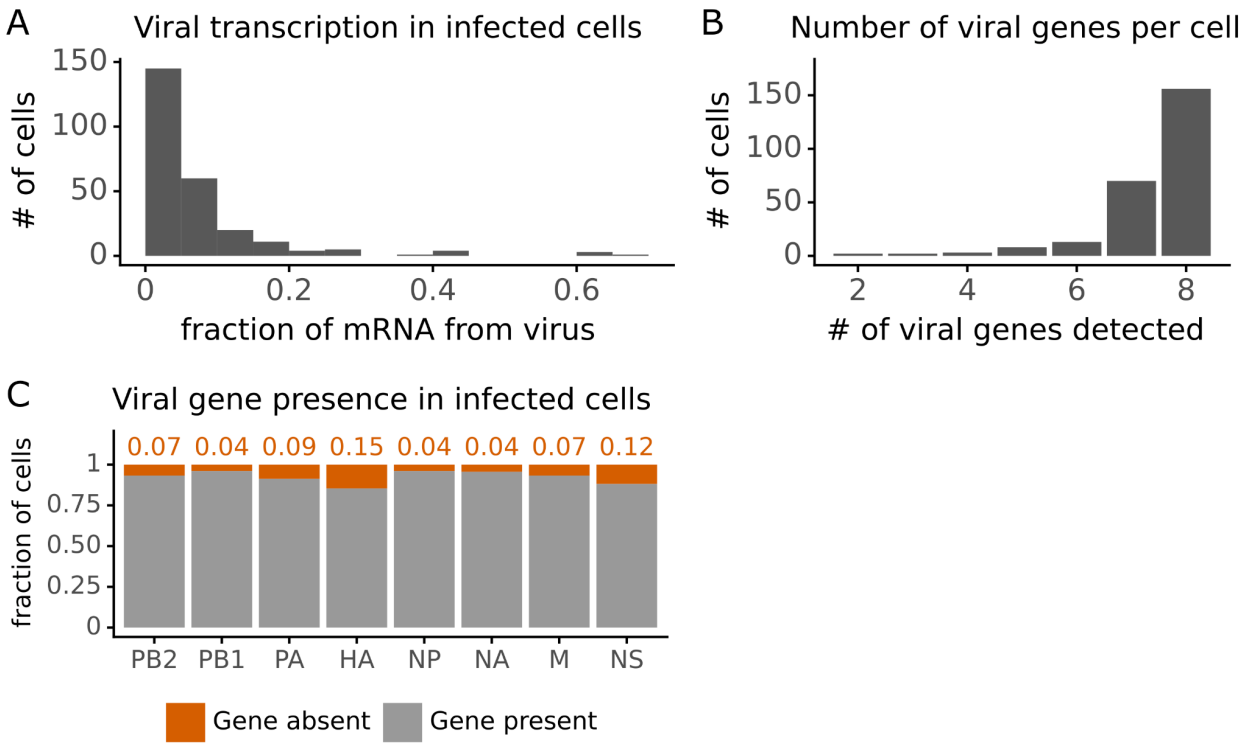


Fig. 2. Viral transcription is extremely heterogeneous across single infected cells, and some cells fail to express some viral genes. This plot shows single-cell RNA-sequencing data for the 254 cells that were infected. **(A)** Viral transcription in infected cells is extremely heterogeneous, with viral 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.

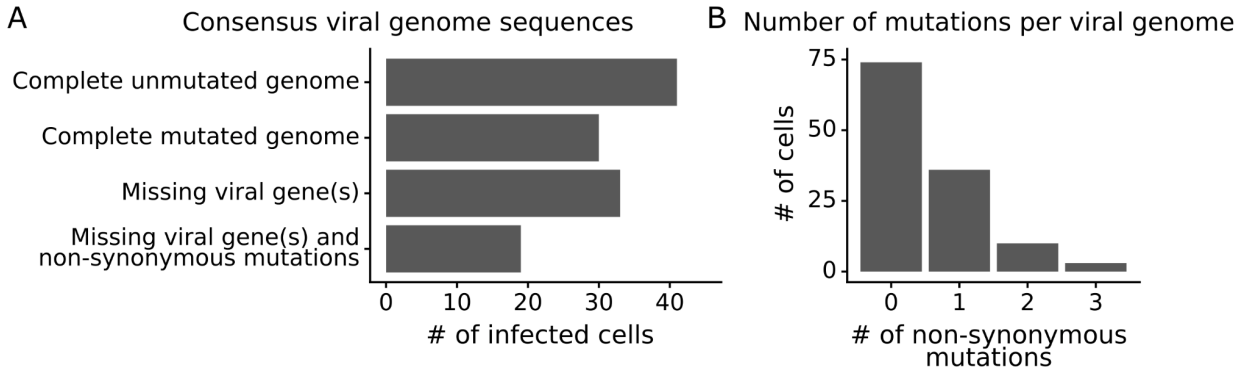


Fig. 3. Statistics on consensus viral genome sequences from single infected cells. (A) The number of infected cells expressing all eight viral genes without non-synonymous mutations, expressing all eight viral genes with one or more non-synonymous mutation(s), missing one or more viral gene(s), or with 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 S4A for details on properties of infected cells for which we could obtain full viral sequences, and Fig S5 for the full set of viral mutations in each infected cell.

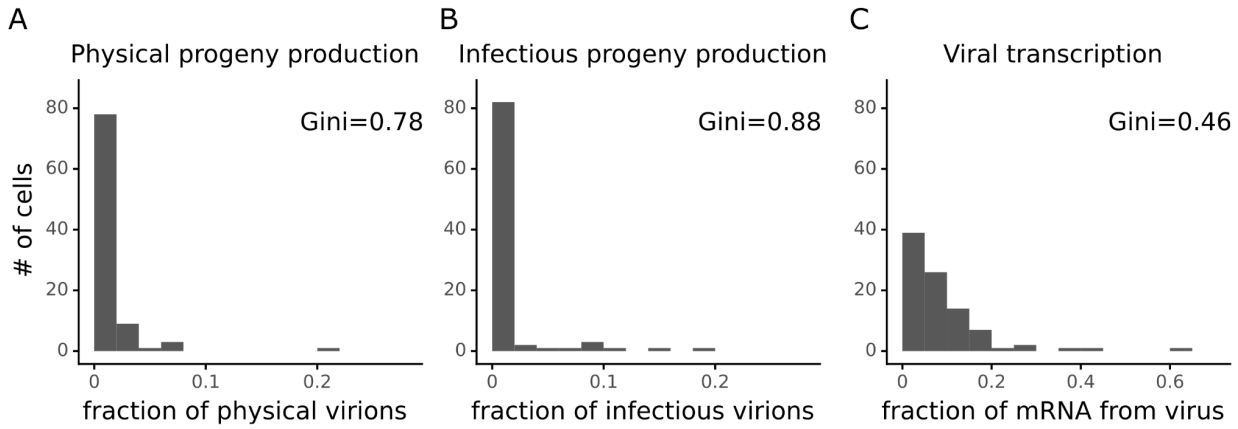


Fig. 4. Viral progeny production is even more heterogeneous than viral transcription across single infected cells. Heterogeneity across single infected cells in **(A)** physical progeny production, **(B)** infectious progeny production, and **(C)** viral transcription. The Gini coefficient [25] quantifying 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.

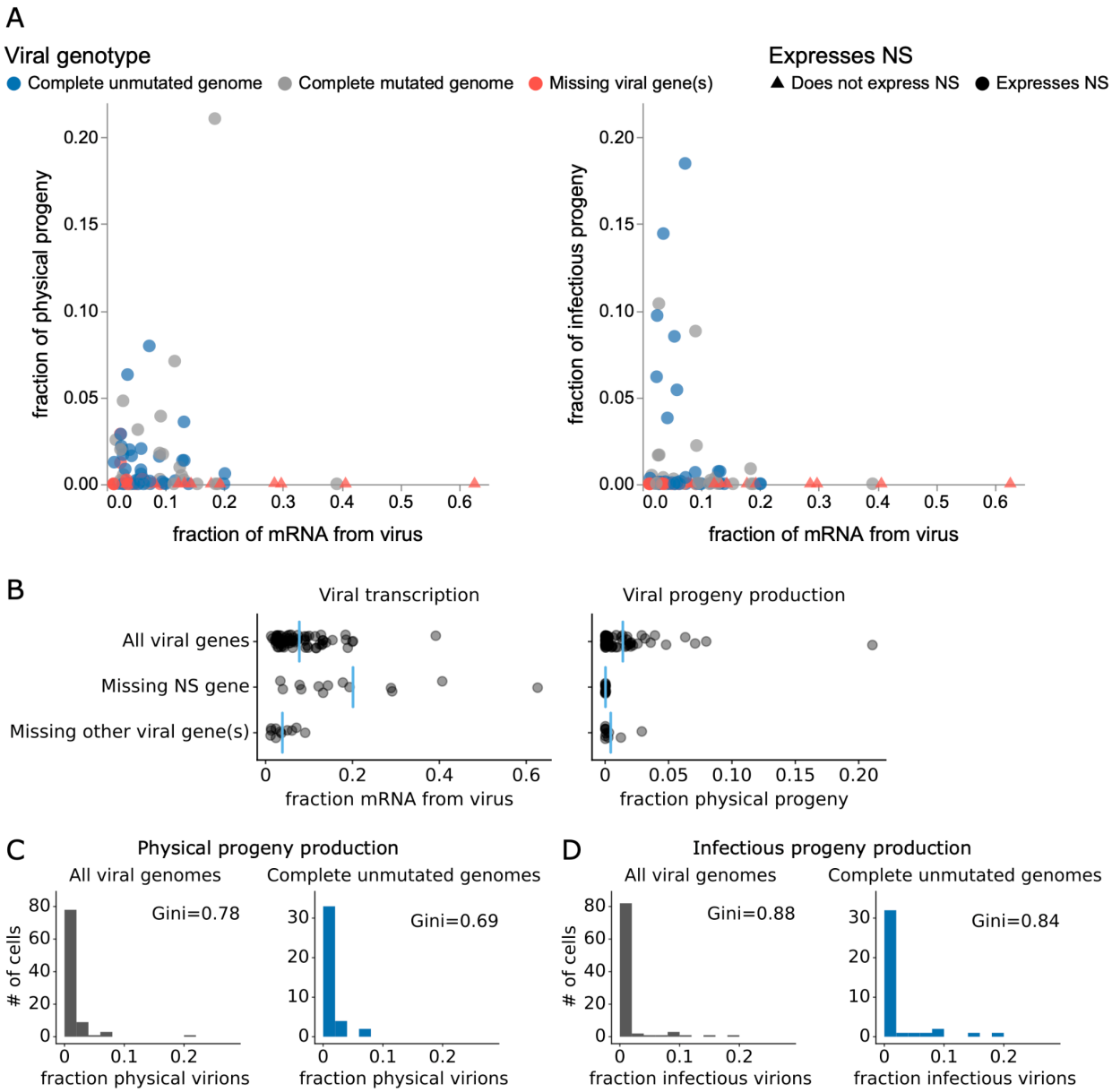
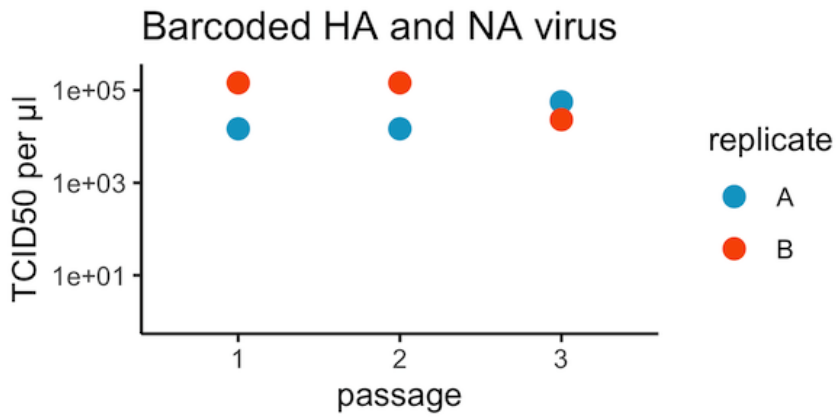


Fig. 5. Relationship between viral transcription and progeny production in single infected cells. (A) Relationship between viral transcription and physical (left) or infectious (right) progeny virion production. Each point is a different cell, colored according to whether the cell expresses unmutated 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.

A



B

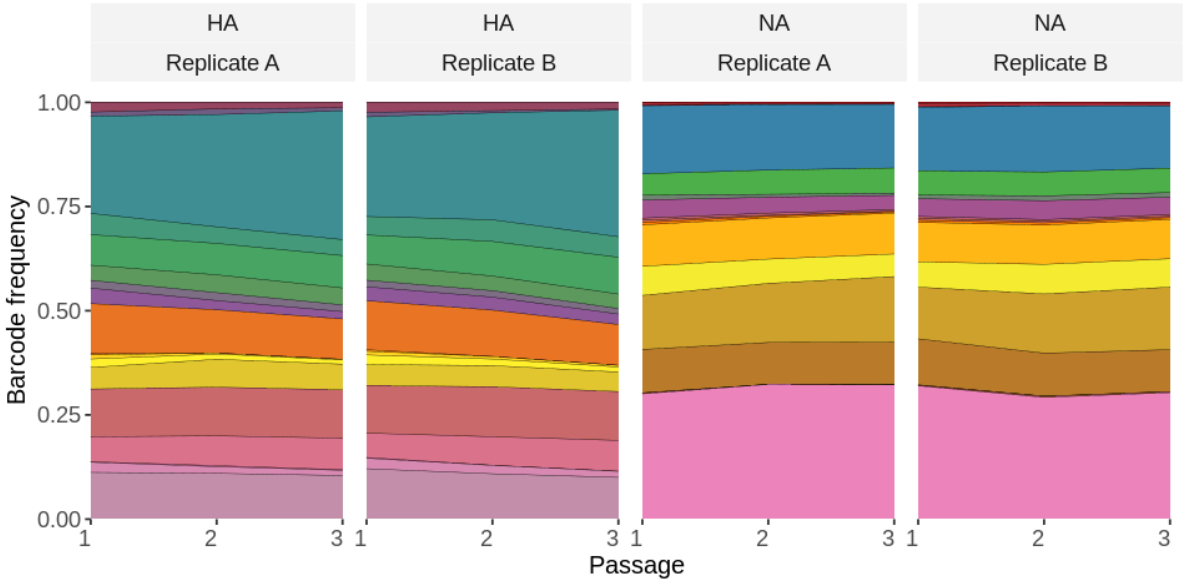


Fig. S1. Viral barcode sequences are selectively neutral. Influenza virus carrying a pool of HA and NA barcodes was generated by reverse genetics and passaged 3 times at low MOI. **(A)** The titers were measured after each growth step by TCID50. (B) The frequency of each barcode in the viral population 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.

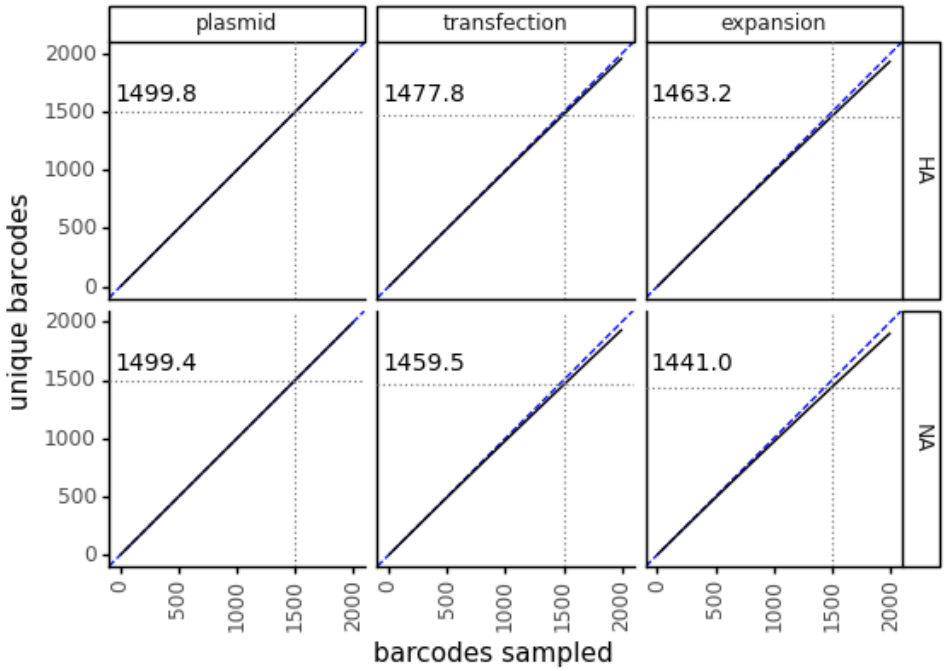


Fig. S2. Extremely diverse barcoded virus libraries. Rarefaction curves show the diversity of the viral barcodes. The x-axis represents the number of barcodes sampled. The y-axis represents the number of sampled barcodes which are unique. A theoretical perfect library where every barcode is unique appears 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.

A

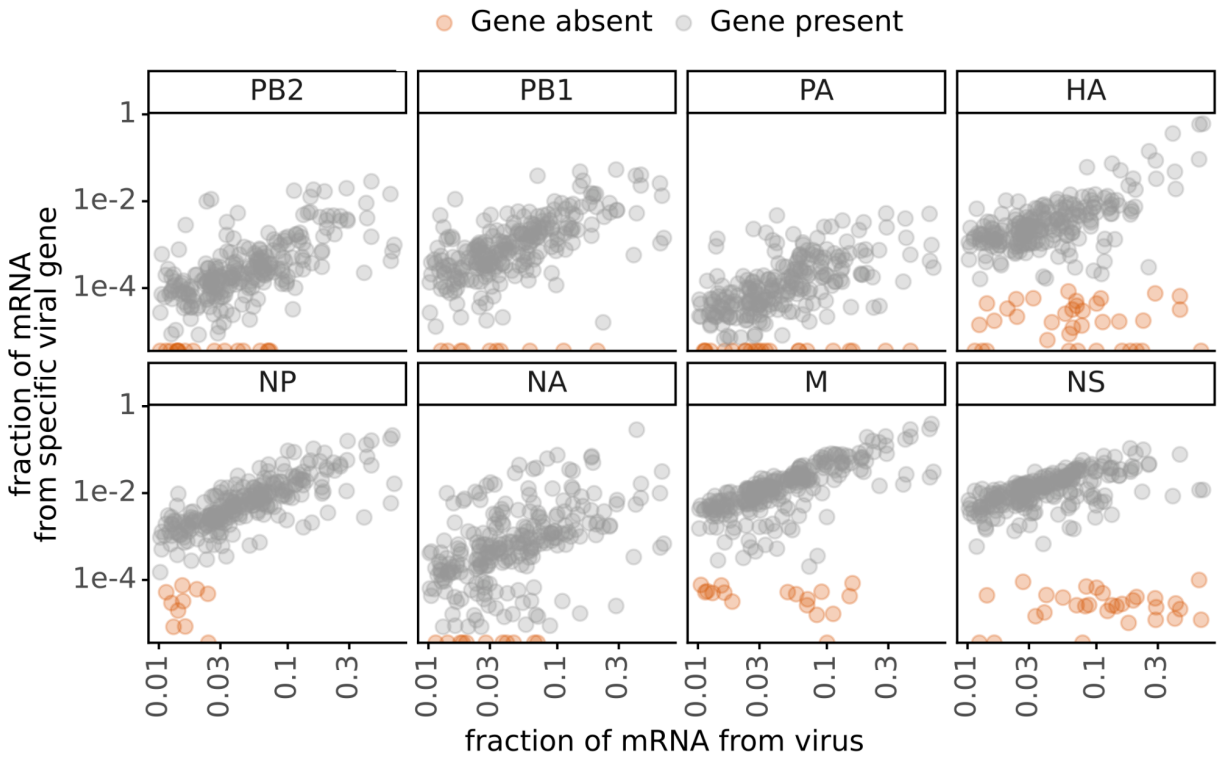


Fig. S3. Expression of viral genes in infected cells. This plot shows single-cell RNA-sequencing data for 254 infected cells. **(A)** Total viral transcription and expression of each viral gene in single infected cells. Genes with low average transcript counts in the single-cell RNA sequencing data (PB2, 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 [55].

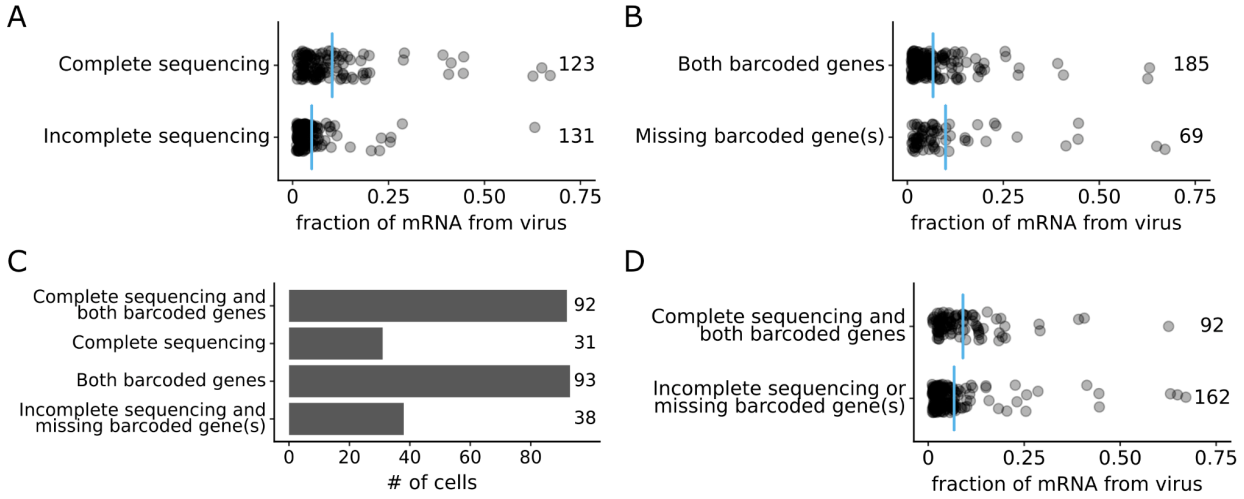
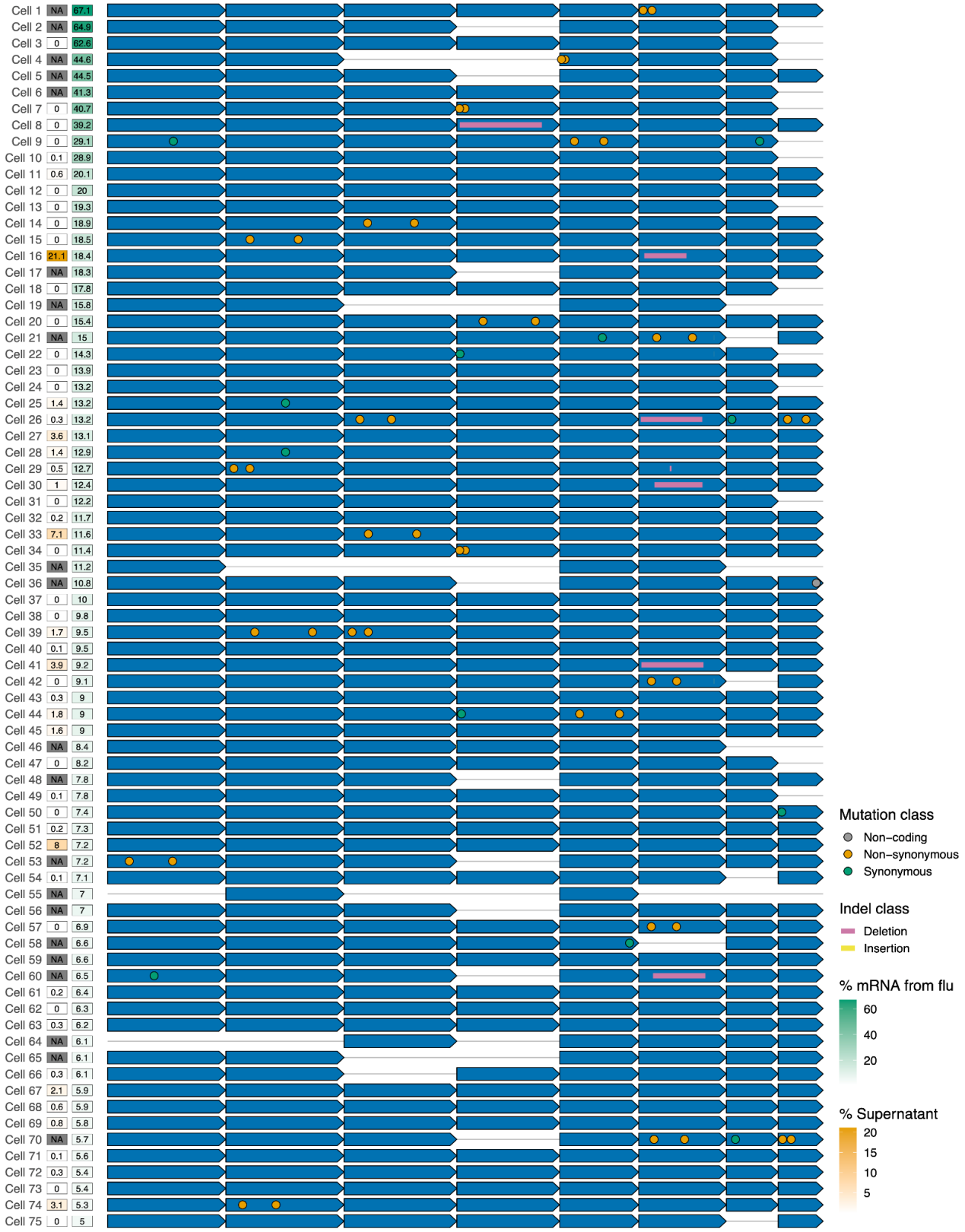


Fig. S4. Statistics of viral genome sequencing and progeny measurements from infected cells. In our dataset, 254 infected cells were identified by single-cell RNA sequencing. **(A)** Of these, 123 infected cells had complete PacBio long-read sequencing data for every expressed viral gene. On average, 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 64.9
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

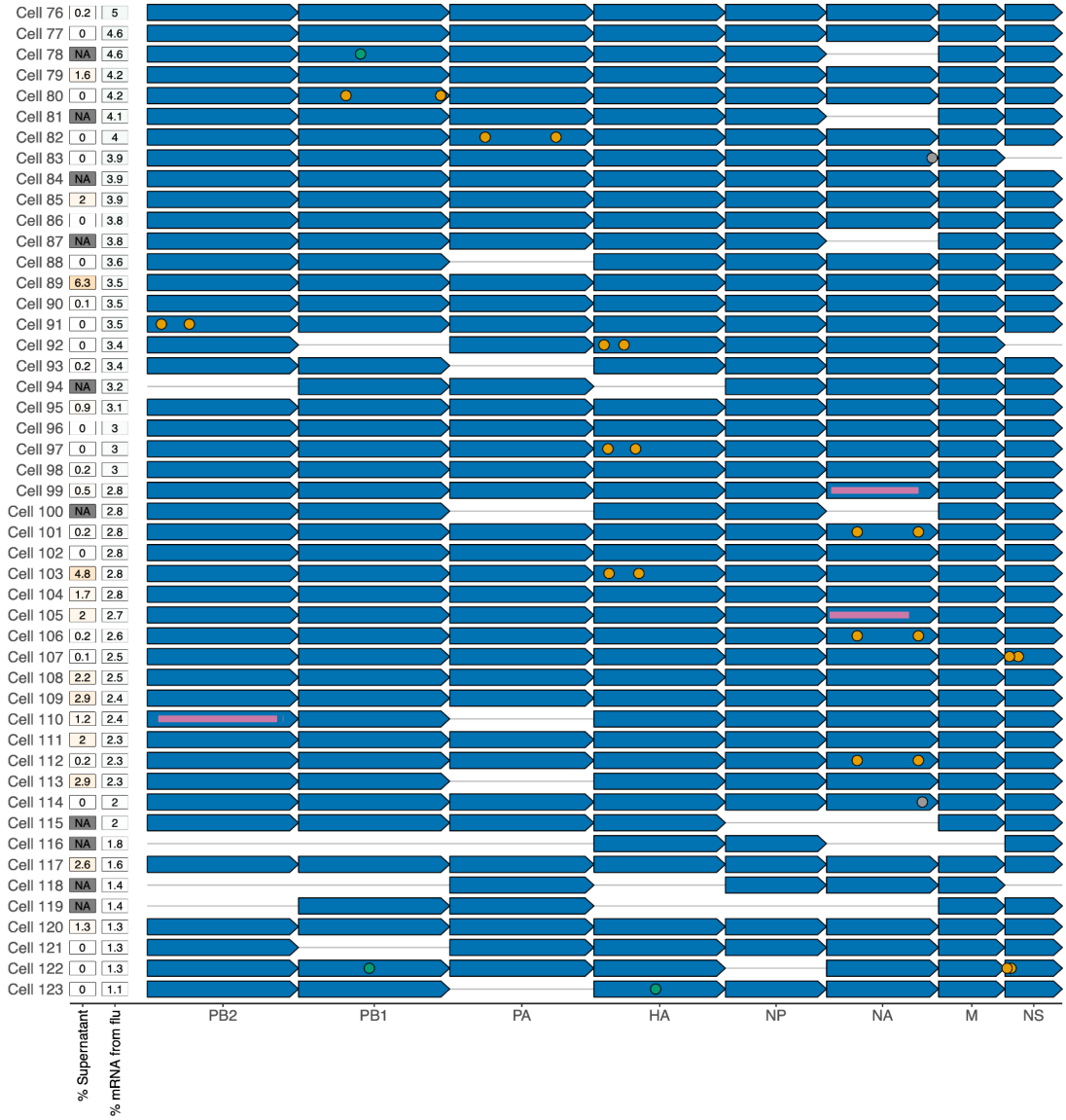


Fig. S5. Viral genotypes in infected cells. The sequence of the infecting virion for the 123 infected cells for which we could determine the sequence of all expressed viral genes. Each infected cell is represented as a row and each viral transcript is represented as an arrow. Missing viral genes, insertions, 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.

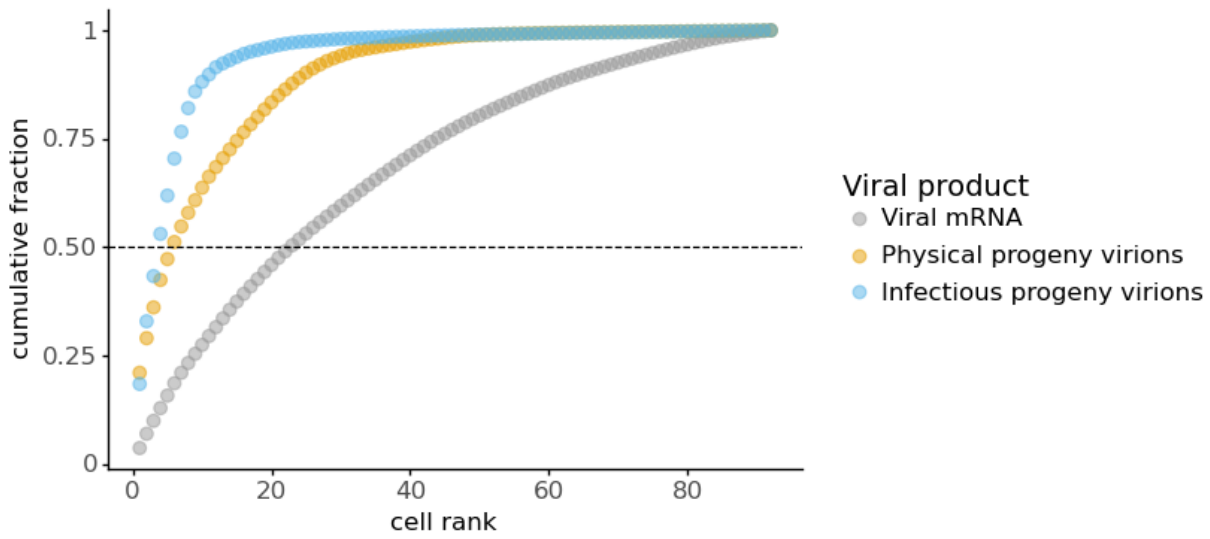


Fig. S6. Cumulative fraction of viral products produced by single infected cells. For the viral mRNA values, the y-axis represents each cell's contribution to the total viral mRNA transcripts across all cells. For the progeny values, the y-axis represents each cell's contribution to the barcodes in the 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.

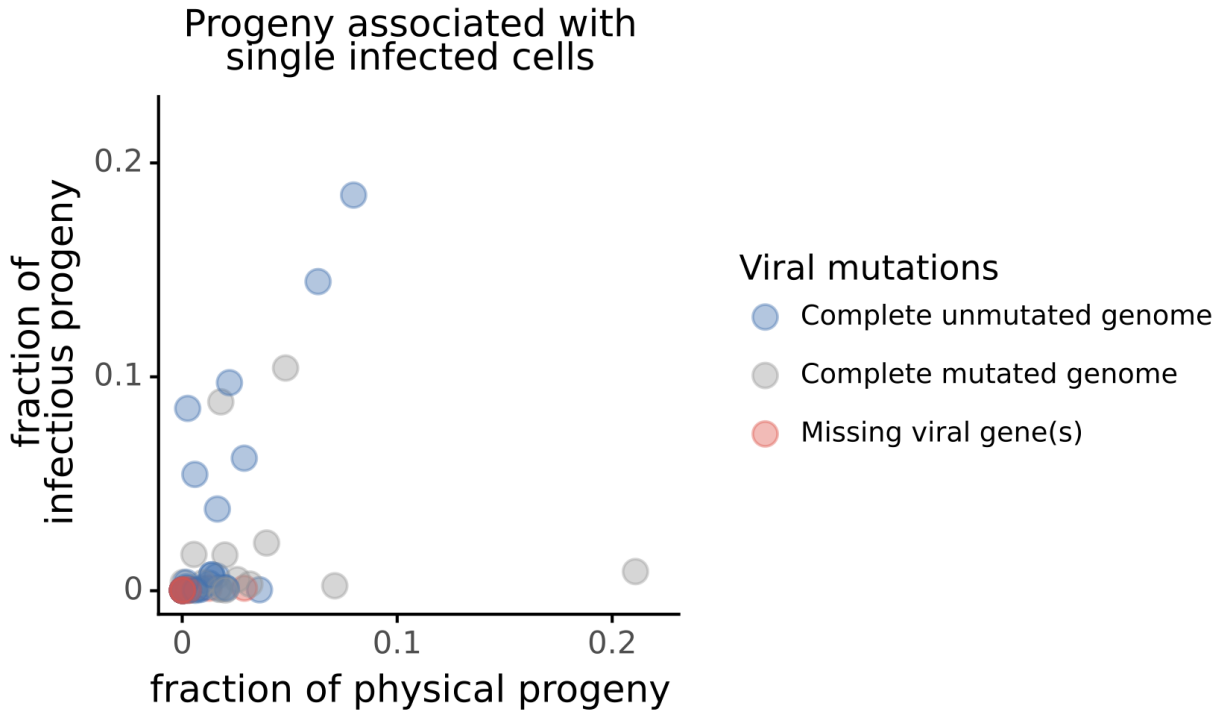


Fig. S7. Frequency of physical progeny and infectious progeny from single infected cells. Each point represents a single infected cell. The x-axis represents the fraction of physical progeny generated by each cell. The y-axis represents the fraction of infectious progeny generated by each cell. This 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.

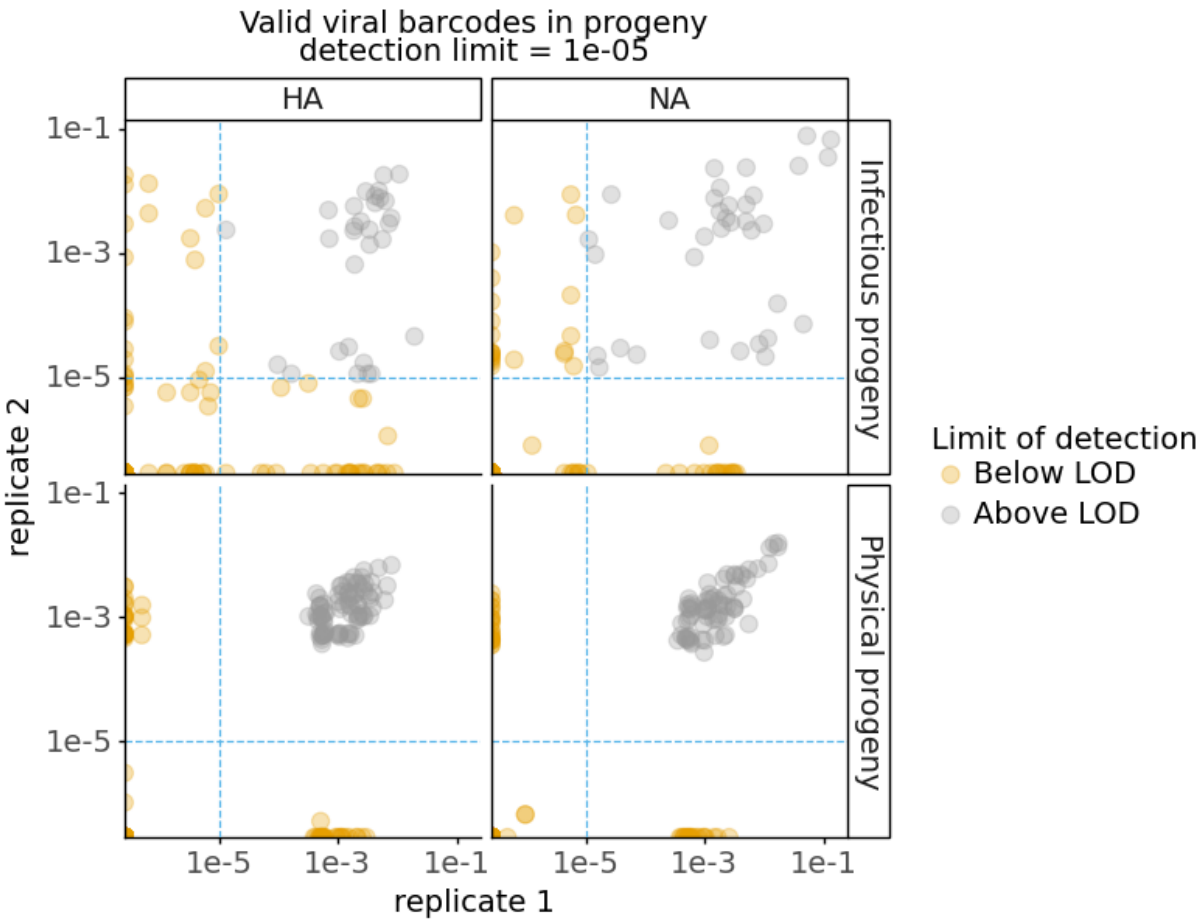


Fig. S8. Technical replicates of progeny measurements. This plot shows the frequency of each viral barcode as measured in two technical replicates of RNA extraction and sequencing. Imperfect correlations indicate that bottlenecking in recovery of molecules contributes noise to the measurements. As described in Fig 1A, physical progeny were measured by sequencing viral barcodes in RNA extracted from viral supernatants, and infectious progeny were measured by sequencing viral barcodes in RNA extracted from cells infected with progeny viral supernatants. Each point represents the fraction of all viral barcodes that could be associated with that cell. The limit of detection (dashed blue line) is set as the point below which replicate frequencies correlate poorly. To calculate the progeny contribution, we averaged the two replicates and then took the geometric mean of the values for HA and NA for each cell.

Primer name	Primer sequence
PB2_CA9_5	ATCGGAAGAGCGTCGTGTAGCGAAAGCAGGTCAATTATATTCAATATGG
PB1_CA9_5	ATCGGAAGAGCGTCGTGTAGCGAAAGCAGGCAAACCATTTG
PA_CA9_5	ATCGGAAGAGCGTCGTGTAGCGAAAGCAGGTACTGATCC
NP_CA9_5	ATCGGAAGAGCGTCGTGTAGCAAAAGCAGGGTAGATAATCACTCAATG
NA_CA9_5	ATCGGAAGAGCGTCGTGTAGCGAAAGCAGGAGTTTAAACCG
HA_CA9_5	ATCGGAAGAGCGTCGTGTAGCAAAAGCAGGGGAAAACAAAAGC
M_CA9_5	ATCGGAAGAGCGTCGTGTAGCAAAAGCAGGTAGATATTTAAAGATGAGTC
NS_CA9_5	ATCGGAAGAGCGTCGTGTAGCAAAAGCAGGGTGACAAAGAC
Read1_TrueSeq	CTACACGACGCTCTTCCGATCT

Table S1. Semi-specific primers for amplification of influenza transcripts from full-length cDNA library. These primers were used to amplify each influenza gene from the full-length single-cell RNA sequencing library. For each reaction, a segment-specific primer was paired with the "Read1_TruSeq" primer, which binds to the Illumina sequencing primer found on all transcripts in the library.

Primer name	Primer sequence
end_PB2_ca_fw	GAGAATAAAAGAACTGAGAGATCTAATGTCGC
end_PB1_ca_fw	GGATGTCAATCCGACTCTACTTTTCC
end_PA_ca_fw	ACTTTGTGCGACAATGCTTCAATCC
end_HA_ca_fw	CGATTTTGGTTGTATTATTGTACACGTTCGC
end_NP_ca_fw	CATCGAAGCCATGGCGTCC
end_NA_ca_fw	GGTAGTCGGAATAATTAGCCTAATATTGCAAATAGG
end_M_ca_fw	CTAACCGAGGTCGAAACGTACG
end_NS_CA_fw	GGACTCCAACACCATGTCAAGC
mid_PB2_ca_fw	CAGCTATTCTCAGAAAGGCAACCAGG
mid_PB1_ca_fw	GCAGAAATGCTAGCAAGCATTGACC
mid_PA_ca_fw	GCTACAGGACATTGAAAATGAAGAGAAGATCC
mid_HA_fw	GCTGGATCTGGTATTATCATTTCAGATACACC
mid_NP_fw	GAGAAAGTCGAAACCCAGGAAACG
mid_NA_fw	TGAGCTGTCCTATTGGTGAAGTTCC
mid_M_fw	GCTATTCAACTGGTGCACTTGCC
mid_NS_fw	GGGAGCAATAGTTGGAGAAATTTCACC
end_PB2_ca_rv	GCGACATTAGATCTCTCAGTTCTTTTATTCTC
end_PB1_ca_rv	GGAAAAGTAGAGTCGGATTGACATCC
end_PA_ca_rv	GGATTGAAGCATTGTCGCACAAAGT
end_HA_ca_rv	GCGAACGTGTACAATAATACAACCAAAATCG
end_NP_ca_rv	GGACGCCATGGCTTCGATG
end_NA_ca_rv	CCTATTTGCAATATTAGGCTAATTATTCCGACTACC
end_M_ca_rv	CGTACGTTTCGACCTCGGTTAG
end_NS_CA_rv	GCTTGACATGGTGTTGGAGTCC
mid_PB2_ca_rv	CCTGGTTGCCTTTCTGAGAATAGCTG
mid_PB1_ca_rv	GGTCAATGCTTGCTAGCATTTCTGC
mid_PA_ca_rv	GGATCTTCTCTTCATTTTCAATGTCCTGTAGC
mid_HA_rv	GGTGTATCTGAAATGATAATACCAGATCCAGC
mid_NP_rv	CGTTTCCTGGGTTTCGACTTTCTC
mid_NA_rv	GGAACTTCACCAATAGGACAGCTCA
mid_NS_rv	GGTGAAATTTCTCCAACTATTGCTCCC
mid_M_rv	GGCAAGTGCACCAGTTGAATAGC

Table S2. Segment-specific primers for amplification of influenza transcripts. These primers were used to amplify the influenza genes from circularized templates. For each gene, two reactions were performed. One set of reactions targeted templates without large deletions and

bound near the middle of the open reading frame; these reactions utilized the primers with "mid" in their name. The other reactions targeted all templates (with and without deletions) and bound near the ends of the open reading frame; these reactions utilized the primers with "end" in their name.

Primer name	Primer sequence
RT_HA	ATGTGTTCCAACGGCTCCTTGCAATGCCGG
RT_NA	CGGATGGTCATGGCCTGATGGAGCAG
PCR_HA_F	gtgactggagttcagacgtgtgctcttccgatctATGTGTTCCAACGGCTCCTTGCAATGCCGG
PCR_NA_F	gtgactggagttcagacgtgtgctcttccgatctCGGATGGTCATGGCCTGATGGAGCAG
PCR_Universal_R	GCCTGTCCGCGGAAGCAGTGG
SampleIndexXX_F	caagcagaagacggcatacgagatXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Adapter_Universal_R	aatgatacggcgaccaccgagatctacacGCCTGTCCGCGGAAGCAGTGG

Table S3. Primers for the reverse transcription, amplification, and indexing of viral barcode sequencing samples. Binding sequences are shown in uppercase and overhangs are shown in lowercase. Primers with "RT" in their name" were used to reverse transcribe the HA or NA viral RNA in supernatants or infected cells. The viral barcodes were amplified from the cDNA using an HA-specific or NA-specific primer (primers with "PCR" in their name) paired with a primer that binds exogenous sequence embedded in the barcoded segments ("PCR_Universal_R"). The samples were prepared for pooling and Illumina sequencing by attaching a sample index ("SampleIndexXX_F") and sequencing adapters ("Adapter_Universal_R") in a final PCR reaction.

References

1. Russell AB, Trapnell C, Bloom JD. Extreme heterogeneity of influenza virus infection in single cells. eLife. 2018;7: e32303. doi:10.7554/eLife.32303
2. Sun J, Vera JC, Drnevich J, Lin YT, Ke R, Brooke CB. Single cell heterogeneity in influenza A virus gene expression shapes the innate antiviral response to infection. PLOS Pathog. 2020;16: e1008671. doi:10.1371/journal.ppat.1008671
3. Zanini F, Pu S-Y, Bekerman E, Einav S, Quake SR. Single-cell transcriptional dynamics of flavivirus infection. eLife. 2018;7: e32942. doi:10.7554/eLife.32942
4. Russell AB, Elshina E, Kowalsky JR, te Velthuis AJW, Bloom JD. Single-Cell Virus Sequencing of Influenza Infections That Trigger Innate Immunity. J Virol. 2019;93: e00500-19. doi:10.1128/JVI.00500-19
5. Brooke CB, Ince WL, Wrammert J, Ahmed R, Wilson PC, Bennink JR, et al. Most Influenza A Virions Fail To Express at Least One Essential Viral Protein. J Virol. 2013;87: 3155–3162. doi:10.1128/JVI.02284-12
6. Schulte MB, Andino R. Single-Cell Analysis Uncovers Extensive Biological Noise in Poliovirus Replication. J Virol. 2014;88: 6205–6212. doi:10.1128/JVI.03539-13
7. Delbrück M. The Burst Size Distribution in the Growth of Bacterial Viruses (Bacteriophages). J Bacteriol. 1945;50: 131–135. doi:10.1128/jb.50.2.131-135.1945
8. Heldt FS, Kupke SY, Dorl S, Reichl U, Frensing T. Single-cell analysis and stochastic modelling unveil large cell-to-cell variability in influenza A virus infection. Nat Commun. 2015;6: 8938. doi:10.1038/ncomms9938
9. Lauring AS, Andino R. Exploring the Fitness Landscape of an RNA Virus by Using a Universal Barcode Microarray. J Virol. 2011;85: 3780–3791. doi:10.1128/JVI.02217-10
10. Amato KA, Haddock LA, Braun KM, Meliopoulos V, Livingston B, Honce R, et al. Influenza A virus undergoes compartmentalized replication in vivo dominated by stochastic bottlenecks. Nat Commun. 2022;13: 3416. doi:10.1038/s41467-022-31147-0
11. Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, Sachs D, et al. Influenza A Virus Transmission Bottlenecks Are Defined by Infection Route and Recipient Host. Cell Host Microbe. 2014;16: 691–700. doi:10.1016/j.chom.2014.09.020
12. Wang C, Forst CV, Chou T, Geber A, Wang M, Hamou W, et al. Cell-to-Cell Variation in Defective Virus Expression and Effects on Host Responses during Influenza Virus Infection. mBio. 2020;11: e02880-19. doi:10.1128/mBio.02880-19
13. Steuerman Y, Cohen M, Peshes-Yaloz N, Valadarsky L, Cohn O, David E, et al. Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing. Cell Syst. 2018;6: 679-691.e4. doi:10.1016/j.cels.2018.05.008
14. Cao Y, Guo Z, Vangala P, Donnard E, Liu P, McDonel P, et al. Single-cell analysis of upper airway cells reveals host-viral dynamics in influenza infected adults. BioRxiv. 2020. doi:10.1101/2020.04.15.042978
15. Heaton NS, Leyva-Grado VH, Tan GS, Eggink D, Hai R, Palese P. *In Vivo* Bioluminescent Imaging of Influenza A Virus Infection and Characterization of Novel Cross-Protective Monoclonal Antibodies. J Virol. 2013;87: 8272–8281. doi:10.1128/JVI.00969-13
16. Gao Q, Palese P. Rewiring the RNAs of influenza virus to prevent reassortment. Proc Natl Acad Sci. 2009;106: 15891–15896. doi:10.1073/pnas.0908897106
17. Chen Z, Wang W, Zhou H, Suguitan AL, Shambaugh C, Kim L, et al. Generation of Live Attenuated Novel Influenza Virus A/California/7/09 (H1N1) Vaccines with High Yield in Embryonated Chicken Eggs. J Virol. 2010;84: 44–51. doi:10.1128/JVI.02106-09
18. Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, et al. Deep mutational scanning

of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proc Natl Acad Sci. 2018;115. doi:10.1073/pnas.1806133115
19. Martin K, Heleniust A. Nuclear transport of influenza virus ribonucleoproteins: The viral matrix protein (M1) promotes export and inhibits import. Cell. 1991;67: 117–130. doi:10.1016/0092-8674(91)90576-K
20. Ohkuma S, Poole B. Fluorescence probe measurement of the intralysosomal pH in living cells and the perturbation of pH by various agents. Proc Natl Acad Sci. 1978;75: 3327–3331. doi:10.1073/pnas.75.7.3327
21. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049. doi:10.1038/ncomms14049
22. Yamawaki TM, Lu DR, Ellwanger DC, Bhatt D, Manzanillo P, Arias V, et al. Systematic comparison of high-throughput single-cell RNA-seq methods for immune cell profiling. BMC Genomics. 2021;22: 66. doi:10.1186/s12864-020-07358-4
23. Saira K, Lin X, DePasse JV, Halpin R, Twaddle A, Stockwell T, et al. Sequence Analysis of *In Vivo* Defective Interfering-Like RNA of Influenza A H1N1 Pandemic Virus. J Virol. 2013;87: 8064–8074. doi:10.1128/JVI.00240-13
24. Davis AR, Hiti AL, Nayak DP. Influenza defective interfering viral RNA is formed by internal deletion of genomic RNA. Proc Natl Acad Sci. 1980;77: 215–219. doi:10.1073/pnas.77.1.215
25. Gini C. Measurement of Inequality of Incomes. Econ J. 1921;31: 124. doi:10.2307/2223319
26. O'Neill RE. The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins. EMBO J. 1998;17: 288–296. doi:10.1093/emboj/17.1.288
27. Bullido R, Gómez-Puertas P, Saiz MJ, Portela A. Influenza A Virus NEP (NS2 Protein) Downregulates RNA Synthesis of Model Template RNAs. J Virol. 2001;75: 4912–4917. doi:10.1128/JVI.75.10.4912-4917.2001
28. Robb NC, Smith M, Vreede FT, Fodor E. NS2/NEP protein regulates transcription and replication of the influenza virus RNA genome. J Gen Virol. 2009;90: 1398–1407. doi:10.1099/vir.0.009639-0
29. Mänz B, Brunotte L, Reuther P, Schwemmle M. Adaptive mutations in NEP compensate for defective H5N1 RNA replication in cultured human cells. Nat Commun. 2012;3: 802. doi:10.1038/ncomms1804
30. Kelly JN, Laloli L, V'kovski P, Holwerda M, Portmann J, Thiel V, et al. 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. 2020. doi:10.1101/2020.04.03.014282
31. Wang C, Forst CV, Chou T, Geber A, Wang M, Hamou W, et al. Cell-to-Cell Variation in Defective Virus Expression and Effects on Host Responses during Influenza Virus Infection. mBio. 2020;11: e02880-19. doi:10.1128/mBio.02880-19
32. Hamele CE, Russell AB, Heaton NS. *In Vivo* Profiling of Individual Multiciliated Cells during Acute Influenza A Virus Infection. J Virol. 2022;96: e00505-22. doi:10.1128/jvi.00505-22
33. Jacobs NT, Onuoha NO, Antia A, Steel J, Antia R, Lowen AC. Incomplete influenza A virus genomes occur frequently but are readily complemented during localized viral spread. Nat Commun. 2019;10: 3526. doi:10.1038/s41467-019-11428-x
34. Sims A, Tornaletti LB, Jasim S, Pirillo C, Devlin R, Hirst J, et al. Superinfection exclusion creates spatially distinct influenza virus populations. 2022. doi:10.1101/2022.06.06.494939
35. Phipps KL, Ganti K, Jacobs NT, Lee C-Y, Carnaccini S, White MC, et al. Collective interactions augment influenza A virus replication in a host-dependent manner. Nat Microbiol. 2020;5: 1158–1169. doi:10.1038/s41564-020-0749-2
36. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS. Stochastic processes constrain the within and between host evolution of influenza virus. eLife. 2018;7: e35962.

doi:10.7554/eLife.35962
37. Xue KS, Bloom JD. Reconciling disparate estimates of viral genetic diversity during human influenza infections. Nat Genet. 2019;51: 1298–1301. doi:10.1038/s41588-019-0349-3
38. Braun KM, Moreno GK, Wagner C, Accola MA, Rehrauer WM, Baker DA, et al. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. PLOS Pathog. 2021;17: e1009849. doi:10.1371/journal.ppat.1009849
39. Martin MA, Koelle K. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2." Sci Transl Med. 2021;13: eabh1803. doi:10.1126/scitranslmed.abh1803
40. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al. SARS-CoV-2 within-host diversity and transmission. Science. 2021;372: eabg0821. doi:10.1126/science.abg0821
41. Gog JR, Afonso EDS, Dalton RM, Leclercq I, Tiley L, Elton D, et al. Codon conservation in the influenza A virus genome defines RNA packaging signals. Nucleic Acids Res. 2007;35: 1897–1907. doi:10.1093/nar/gkm087
42. Watanabe T, Watanabe S, Noda T, Fujii Y, Kawaoka Y. Exploitation of Nucleic Acid Packaging Signals To Generate a Novel Influenza Virus-Based Vector Stably Expressing Two Foreign Genes. J Virol. 2003;77: 10575–10583. doi:10.1128/JVI.77.19.10575-10583.2003
43. Marsh GA, Hatami R, Palese P. Specific Residues of the Influenza A Virus Hemagglutinin Viral RNA Are Important for Efficient Packaging into Budding Virions. J Virol. 2007;81: 9727–9736. doi:10.1128/JVI.01144-07
44. Hoffmann E, Neumann G, Hobom G, Webster RG, Kawaoka Y. "Ambisense" Approach for the Generation of Influenza A Virus: vRNA and mRNA Synthesis from One Template. Virology. 2000;267: 310–317. doi:10.1006/viro.1999.0140
45. Neumann G, Watanabe T, Ito H, Watanabe S, Goto H, Gao P, et al. Generation of influenza A viruses entirely from cloned cDNAs. Proc Natl Acad Sci. 1999;96: 9345–9350. doi:10.1073/pnas.96.16.9345
46. Bloom JD. Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. PeerJ. 2018;6: e5578. doi:10.7717/peerj.5578
47. Judo MSB, Wedel AB, Wilson C. Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res. 1998;26: 1819–1825. doi:10.1093/nar/26.7.1819
48. Library Construction - Official 10x Genomics Support. Available: https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/library-prep/chromium-single-cell-3-reagent-kits-user-guide-v-3-1-chemistry
49. Qiagen. RNeasy Mini Handbook. Available: https://www.qiagen.com/us/resources/resourcedetail?id=14e7cf6e-521a-4cf7-8cbc-bf9f6fa33e24&lang=en
50. ThermoFisher. SuperScript™ III First-Strand Synthesis SuperMix for qRT-PCR. Available: https://www.thermofisher.com/document-connect/document-connect.html?url=https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fmanuals%2Fsuperscript_firststrand_qrtpcr_man.pdf
51. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522. doi:10.1093/bioinformatics/bts480
52. Kaminow B, Yunusov D, Dobin A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. 2021. doi:10.1101/2021.05.05.442755
53. Crawford K, Bloom J. alignparse: A Python package for parsing complex features from high-throughput long-read sequencing. J Open Source Softw. 2019;4: 1915. doi:10.21105/joss.01915
54. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017;27: 491–499. doi:10.1101/gr.209601.116

55. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience. 2020;9. doi:10.1093/gigascience/giaa151