Title:

Heterogeneity of Genetic Sequence within Quasi-species of Influenza Virus Revealed by Single-Molecule Sequencing

Kenji Tamao[1], Hiroyuki Noji[1,2], Kazuhito V. Tabata[1,2*]

1 Department of Applied Chemistry, Graduate School of Engineering, the University of Tokyo, Tokyo, Japan

2 Research Institute of Planetary Health, The University of Tokyo, Japan.

*Corresponding author: Kazuhito V.Tabata

Email: tabatak@g.ecc.u-tokyo.ac.jp

Author Contributions: K.T. and K. V. T. designed the study and wrote the paper. K. T. performed the experiments and analyzed the data. H. N. provided critical discussion and comments.

Competing Interest Statement: The authors declare no competing interests.

Classification: Biological Sciences: Genetics; Biological Sciences: Microbiology

Keywords: Influenza, Genomics, Sequencing, Quasi-species

This PDF file includes:

Main Text
Figures 1 to 5

Abstract

Influenza viruses are characterized by high mutation rates and extensive genetic diversity, which hinder effective vaccine development and facilitate immune evasion (Taubenberger & Morens, 2006; Barr et al., 2010). These mutations primarily arise from the error-prone activity of the viral RNA-dependent RNA polymerase, generating highly heterogeneous viral populations within individual hosts. This phenomenon aligns with the quasi-species model, in which a cloud of related viral genomes evolves under selective pressures (Domingo et al., 2012). Accurate characterization of this intra-host diversity is crucial for understanding viral evolution and informing future vaccine design. However, conventional RNA sequencing technologies often fail to reliably detect low-frequency variants due to technical errors introduced during sample preparation and sequencing steps.

In this study, we implemented a single unique molecular identifier (sUMI) approach to minimize sequencing artifacts and achieve an error rate of approximately 10^{-5}. This high-resolution method enabled precise quantification of quasi-species diversity from influenza virus populations isolated at the single-particle level. Comparative analyses revealed mutation frequencies well above background error levels, confirming that the observed variation was of biological origin. Furthermore, application of information-theoretic metrics such as Shannon entropy and Jensen-Shannon divergence demonstrated that the mutation distribution was non-random, suggesting the presence of selective constraints.

Our findings establish a robust framework for studying intra-host viral evolution and provide critical insights that may enhance AI-driven prediction of mutational trajectories and support more effective influenza vaccine strategies.

Main Text

Introduction

Influenza is a seasonal illness caused by the influenza virus(1–3). The influenza virus has been observed to exhibit a high mutation rate and antigenic diversity, which has the potential to trigger pandemics and exert significant social and economic impacts(4, 5).

In the context of influenza, vaccine development has been a prominent preventive strategy. However, the rapid mutation of the virus can lead to fluctuations in vaccine efficacy on an annual basis, thereby complicating precise prediction and comprehensive prevention of infection difficult(5, 6).

Mutations in influenza viruses are primarily introduced by RNA-dependent RNA polymerase (RdRp)(7). The high error rate characteristic of RdRp enables the virus to undergo rapid evolution, thereby acquiring the capacity to evade existing immune responses(8). It has been suggested that mutant strains that evade antibody-mediated immunity while retaining infectivity tend to exhibit specific patterns(9, 10). Further research is necessary to elucidate these predictable mutation pathways(11). Recently, bioinformatics analyses that employ artificial intelligence (AI) have played a key role in investigating viral evolution mechanisms and exploring strategies for future vaccine design(12–15). However, enhancing the precision of AI-driven mutation prediction necessitates the consolidation of extensive global data. This effort is hampered by spatial and temporal limitations, which restrict the amount of data that can be collected and act as a bottleneck for enhancing the accuracy of predictions(16).

Currently, the genome sequences that are registered in viral genome databases represent the mean sequences of viral particle populations. However, due to the error-prone nature of viral RNA polymerase, viral populations exhibit high genetic diversity within a single host(17, 18). This genetic diversity within viral populations can be described by the quasi-species model, which

envisions a cloud of related viral genomes centered around a dominant sequence(19). Quasi-species formation plays a crucial role in viral survival strategies, allowing the population to rapidly adapt to environmental pressures such as host immunity or antiviral drugs through collective evolution(20). The development of technology capable of accurately measuring the distribution of genome sequences within a virus population would facilitate a comprehensive evaluation of the types and frequencies of such precursor variants. It is anticipated that these data will play a significant role in the field of AI-based mutation prediction. Accordingly, the quasi-species concept serves as a fundamental framework for understanding viral evolution(21). The accurate quantification of quasi-species diversity is pivotal in facilitating a more profound comprehension of mutation dynamics, selective pressures, and evolutionary potential within a viral population(22)-crucial not only for anticipating future dominant variants but also for detecting low-frequency mutations associated with drug resistance or antigenic shifts.

In current RNA sequencing technology, is plagued by the presence of errors that are introduced during reverse transcription process. The error rate has been reported to be 10^{-4} to 10^{-5} per nucleotide when using SuperScript IV reverse transcriptase, as utilized in the present study(23, 24). Accordingly, a target sequencing error rate of 10^{-5} was established, and the single UMI method was employed. This method has been shown to minimize sample loss and streamline procedures, thereby facilitating the quantification of genomic diversity within influenza virus particle populations(25).

In this study, we demonstrated that the sUMI method can effectively reduce sequencing error rates to the 10^{-5} level. This method involves the ligation of UMIs directly to DNA amplified by cloning, followed by PCR and sequencing. Subsequently, an analysis of quasi-species was conducted on a population of influenza virus particles that had proliferated from a single particle and quantified the variation in their genome sequences was quantified (Fig.1). Comparison of this variation with the mutation rate of RNA synthesized by *in vitro* transcription showed that viral populations harbored significantly more mutations, indicating they arose during viral replication. Furthermore, we employed information-theoretic metrics, including Jensen–Shannon divergence(26) and Shannon entropy(27), to analyze the distribution of mutations. Our analysis revealed that specific positions within the viral genome demonstrate selective propagation of mutations.

These results provide a novel approach for high-accuracy detection of minor genetic mutations within influenza virus populations and offer valuable insights for understanding viral evolution and designing future vaccines.

Results

Confirmation of Reduction of Error Rate by Sequencing UMI-ligated DNA

Initial findings indicate that establishing a threshold for UMI redundancy leads to a substantial reduction in sequencing errors attributable to the methodology. A single plasmid molecule was amplified via E. coli cloning and subsequently linearized by restriction enzyme digestion. An oligonucleotide containing a primer-binding site and a unique molecular identifier (UMI) sequence at its 5' end was then ligated to the linearized plasmid. The resulting UMI-tagged linear plasmid was amplified by PCR and subjected to sequencing using the Revio system. The resulting single-molecule sequences were then subjected to analysis using the umierrorcorrect function. During the analysis, we applied GroupSize thresholds of 1, 2, and 3 were applied, and only reads that met each threshold were used to generate consensus sequences for identical UMIs. These consensus sequences were then mapped to the reference sequence. Consequently, the error rates for GroupSize thresholds of 1, 2, and 3 were 3.59×10^{-4}, 1.35×10^{-4}, and 3.81×10^{-5},

respectively (Fig.2a). When the threshold was set to 4, no errors were observed among 37,307 base pairs. These findings demonstrate that filtering sequences based on UMI redundancy can significantly reduce method-derived errors, improving sequencing accuracy by nearly an order of magnitude.

It is widely acknowledged that PacBio sequencers exhibit elevated error rates in homopolymer regions, defined as stretches of identical nucleotides(28, 29). In this study, the term "homopolymer region" was defined as a sequence containing three or more identical consecutive bases, including one base on each side. Subsequently, an analysis was conducted to compare the error rates observed in homopolymer and non-homopolymer regions. In accordance with preceding reports, elevated error rates were detected in homopolymer regions. In the sequencing data of UMI-tagged linear plasmids employing a GroupSize threshold of 3, the error rate in homopolymer regions was 7.11×10^{-5}, whereas the error rate in non-homopolymer regions was 1.66×10^{-5} (Fig.2b). As the GroupSize threshold increased from 1 to 3, a decrease in the discrepancy between homopolymer and non-homopolymer regions was observed. This finding indicates that the averaging of sequences with identical UMIs effectively mitigates sequencing errors, particularly in homopolymer regions.

Quantification of Error Rate Caused by RT and Comparison with Virus-derived Mutation

In order to ascertain the mutation distribution of viral populations derived from a single virus particle, we sequenced viral genomic RNA was extracted from infected cells and subjected to sequencing. In order to differentiate between biological mutations and technical errors, a comparison was made with *in vitro* transcribed RNA derived from plasmids.

MDCK cells were seeded in a 6-well plate and infected with influenza virus at a multiplicity of infection (MOI) of 0.05 PFU/well. Following a 72-hour incubation period, wells exhibiting cytopathic effects (CPE) were harvested. Of the 96 wells, four exhibited clear CPE, suggesting successful viral amplification from single particles. The supernatants from these wells were collected as independent viral populations, and total viral genomic RNA was extracted. RT-PCR analysis of the NA gene confirmed that virus replication occurred only in samples where CPE was observed (Fig. S1).

For each viral RNA sample, reverse transcription (RT) was performed with the ligation of unique molecular identifiers (UMIs), followed by PCR amplification and sequencing. To control for errors introduced during RT-PCR and sequencing, we prepared plasmids containing sequences homologous to the PB2 and HA segments under a T7 promoter and synthesized RNA via *in vitro* transcription. These *in vitro* transcripts were subsequently processed using the same UMI-based RT-PCR and sequencing protocol. Distribution of the number of consensus sequences after filtering based on UMI group size for each gene and population is shown in Fig.S2

In order to differentiate between biological mutations and technical errors, a method was developed that involved mapping viral reads to the consensus sequence obtained from each viral population and the *in vitro* transcribed reads to the reference plasmid sequence that was utilized for transcription. The observed error rates in the *in vitro* transcripts are considered to reflect technical errors associated with RT and sequencing. Conversely, the higher error rates observed in the viral RNA are attributed to a combination of biological mutations and technical noise.

It is noteworthy that viral RNA exhibited a consistent pattern of elevated error rates (Fig. 3a). For the PB2 segment, the error rate for viral RNA was 1.19×10^{-4}, whereas that for in vitro transcribed PB2-like RNA was 2.22×10^{-6}. Accordingly, the probability that a detected mutation in

PB2 originates from viral replication rather than experimental error was calculated as 1 - (2.22 × 10^{-6})/(1.19 × 10^{-4}) ≈ 98.1%. Segment-specific error rates ranged from 1.17 × 10^{-4} to 2.19 × 10^{-4} (Table S1), and thus the PB2 value is within the genome-wide range. These error rates were significantly higher than those of in vitro transcribed PB2-like RNA, indicating that detected mutations are highly likely to originate from viral replication.

These results demonstrate that the majority of detected mutations are derived from biological processes during viral replication rather than technical artifacts. Consequently, this approach facilitates single-molecule genome sequencing and precise profiling of mutation heterogeneity within viral populations.

Mutation Profile of RNA Extracted from a Virus Population Derived from a Single Virus Clone

Table S1 presents a summary of the mutation frequency, transition/transversion ratio (Ti/Tv), and the ratio of nonsynonymous to synonymous substitutions (dN/dS) for each gene. The mutation detection rate across coding regions is illustrated in Fig.3b. The observed Ti/Tv ratios ranged from 2 to 5, which is consistent with previously reported values(30, 31). In the absence of any specific selection pressure was applied within the cell culture system, the dN/dS values were found to be generally below 1, thereby suggesting purifying selection as a mechanism for maintaining protein function. It is noteworthy that the NA gene exhibited dN/dS values approaching or exceeding 1, indicating a heightened degree of evolutionary plasticity.

An investigation was conducted into the presence of defective viral genomes that exhibited partial deletions. This investigation was motivated by the findings of prior studies, which reported the existence of internally deleted genomes in influenza viruses(32, 33). Among the 546,580 molecules that were analyzed, 837 (0.15%) contained continuous deletions longer than 10 nucleotides. However, no such deletions were identified in the *in vitro* transcribed RNAs, suggesting that these deletions likely emerged during the viral replication process (Fig.S3). Conversely, a mere 13 molecules (0.0024%) exhibited insertions that spanned more than 100 nucleotides. It should be noted that the RT-PCR primer design strategy did not take into account sequences with consecutive deletions at the termini, but only sequences with consecutive deletions in the center of the genome.

On average, over 10,000 consensus reads were obtained per gene after UMI-based error correction (group size ≥3). Independent digital particle counting estimated that each viral culture contained approximately ~10^{13} viral particles. Under these conditions, variants present at a frequency of at least ≥0.1% in the population can be detected with >95% probability, exceeding the background error rate shown in Fig. 3a.

On average, over 10,000 sequencing reads were obtained per gene. Independent digital particle counting estimated that each viral culture contained approximately ~10^{13} viral particles. Given the detection probability of rare variants (e.g., 0.1% frequency), the sequencing depth of 10,000 reads provides >95% probability of detecting such mutations, thereby exceeding the background error rate shown in Fig.3a. This finding suggests that mutations present at a frequency of at least ≥0.1% frequency in the population can be reliably detected.

In order to assess the applicability of this approach for comprehensive detection of drug-resistant mutations, a model was developed to demonstrate how mutant frequencies change with fitness using a competitive logistic growth model. This model is predicated on the premise that host cell resources are limited, and that there is a transition from exponential growth to saturation. It is

widely employed for the estimation of effective reproduction numbers(34–37). The minimum relative fitness (compared to wild type) necessary for mutants to surpass this threshold, depending on their initial abundance is shown in Fig.S4. Even late-arising mutants (1 in 10^{11}) with ≥5-fold fitness advantage could reach 0.1% of the population, indicating that such adaptive mutations can be detected by this method.

Characteristics of the Observed Mutation Distribution and Deviation from the Poisson Distribution

In order to investigate the characteristics of the mutation distribution in viral populations, a comparison was made between distribution of genome sequences derived from viral RNA and that of RNA synthesized by *in vitro* transcription. For each specific nucleotide position, we constructed the distribution P was constructed by defining P(k) as the number of positions at which mutations were observed k times. Assuming that mutations introduced by T7 RNA polymerase, reverse transcriptase, and PCR enzymes occur randomly, the sequencing errors detected in *in vitro* transcribed RNA are expected to approximate a Poisson distribution. Conversely, if mutations introduced during viral replication are entirely random with respect to genomic location, then the mutation distribution of viral RNA should also adhere to a Poisson distribution. Conversely, substantial deviations from a Poisson distribution may indicate the presence of selection pressures, such as differences in replication efficiency associated with viral protein function or genomic sequence features.

The quantification of the discrepancy between the experimentally observed mutation count distribution observed through experimentation and the corresponding Poisson distribution with a mean count (λ) was achieved through the implementation of Jensen–Shannon (JS) divergence. The JS divergence is a symmetric and normalized measure of relative entropy that is both symmetric and normalized. It takes on values approaching 1, indicating greater dissimilarity between distributions. The results indicated that, the entire gene set, the mutation distributions of viral RNA exhibited a statistically significant deviation from the Poisson distribution when compared with those of *in vitro* transcribed RNA (Fig.4a). These findings imply that the observed mutation patterns in viral populations are not simply the result of neutral, randomly introduced mutations but reflect constraints imposed by selection related to protein function or replication efficiency. However, it is noteworthy that many minor mutations are introduced in a manner that approximately follows a Poisson distribution. These results clearly demonstrate that most mutations are introduced randomly, contributing to the formation of quasi-species.

To further evaluate the phenotypic diversity within the viral population, we quantified the diversity of amino acid sequences encoded by the RNA was quantified Shannon entropy. Shannon entropy is a measure of uncertainty based on probability distributions, with higher values indicating greater variability. For each viral gene, the distribution of protein sequences derived from all sequencing fragments (with UMI count ≥ 3) was constructed, and the entropy values between viral RNA and *in vitro* transcribed RNA were compared. The results demonstrated that, across the entire gene pool, viral RNA exhibited a higher degree of amino acid sequence diversity in comparison to *in vitro* transcribed RNA (Fig.4b). This increased diversity is likely indicative of the genetic variation accumulated during viral propagation and can serve as an indicator of the potential evolutionary capacity of the virus under the culture conditions utilized in this study.

Comparison Between Mutations Derived from Viral Propagation and Actual Evolutionary Lineages

The majority of the mutations measured in this study within viral populations are likely to be neutral mutations that have been randomly introduced. However, it is important to note that each of these mutations has the potential to become fixed and dominate the population, representing possible origins of emerging variants. A comparison of the viral genome sequences between the strain used in this experiment, A/Puerto Rico/8/1934 (PR8), and its closely related strain, A/Alaska/1935, revealed three amino acid differences: I563R (PB1), D452E (HA), and M458V (HA)(38, 39). Of these, two mutations—I563R (PB1) and M458V (HA)—were detected in at least one of the four experimental populations (Fig.5a). These mutations manifested in non-homopolymeric regions, and their frequencies exceeded the mean mutation rate in these regions across all genes, as illustrated in Fig.3a. These findings suggest that these mutations may be preferentially introduced during replication or may be positively selected due to their capacity to increase their own copy numbers.

Subsequently, under the assumption that mutations introduced at each nucleotide position follow a Poisson distribution, we identified mutations that occurred at frequencies with a probability of ≤5% as candidates for errors that were positively selected and proliferated. A comparison of these mutations with those found between PR8 and the recent pandemic strain A/California/01/2009 (CA01) revealed that 14 amino acid substitutions were shared(40). Among the 4,450 amino acid residues, 402 differed between PR8 and CA01, and a total of 596 non-synonymous mutations exceeding the threshold were identified in this study. Assuming random selection of positions, the probability that a randomly introduced mutation matches both the position and the resulting amino acid of any of the 402 known substitutions is calculated follows: The resulting percentage, 1.6%, is then calculated by dividing 402 by 4450 and then multiplying by 0.177. In this context, 0.177 represents the mean probability that two independent random point mutations on the same codon will result in the same amino acid substitution. The z-score for observing 14 coincident mutations is calculated as follows:

$$\lambda = 0.016$$
$$n = 596$$
$$X = 14$$
$$z = \frac{X - n\lambda}{\sqrt{n\lambda}} = 1.43$$

The corresponding one-sided p-value is 0.0764. Consequently, no statistically significant difference was observed at the 5% level of significance, indicating that selective mutation introduction driving convergence toward recent pandemic strains cannot be concluded from this dataset.

Discussion

In this study, we established a methodology for single-molecule genome RNA sequencing within cultured influenza virus populations and demonstrated the potential for experimental mutation prediction through sequence distribution analysis. To validate this approach, the sequencing of UMI-tagged linearized plasmids was conducted. The analysis revealed that focusing solely on sequences with three or more reads that shared the same UMI led to an enhancement in sequence accuracy by more than tenfold. This resulted in an error rate that was on the order of 10^{-5} per bp per read. Assuming that PCR and sequencing introduce errors at a rate of approximately 10^{-3}, the probability of two or more errors occurring at the same base position among three reads theoretically decreases to the order of 10^{-6}. The discrepancy between this theoretical value and our measured data is likely attributable to mutation-prone regions, such as homopolymers and amplification bias, where error-containing sequences were preferentially amplified and detected as majorities.

It is widely acknowledged that PacBio sequencers frequently generate errors in homopolymer regions, and our study observed analogous trends. Specifically, when setting the UMI redundancy threshold was set to three, error rates in homopolymer regions found to be higher than those in non-homopolymer regions. However, the integration of UMIs has led to a significant reduction in this error gap, thereby validating the efficacy of UMI technology in addressing erroneous reads. This finding indicates that UMI-based methods have the potential to enhance the accuracy of PacBio, especially for genes and sequences that contain homopolymer regions. The implications of this study extend to other single-molecule sequencing platforms such as nanopore.

A subsequent analysis of RNA from virus populations derived from single virus particles exhibited elevated error rates when compared to *in vitro* transcribed RNA, a phenomenon that is presumably attributable to mutations introduced during viral replication. The observed mutation distribution comprised both mutations consistent with a neutral, Poisson-like accumulation and mutations that deviated substantially from a Poisson distribution. This pattern indicates the coexistence of neutral and non-neutral mutations within the viral population, forming a quasi-species structure. The deviation from Poisson expectations suggests that certain mutations were subject to selective pressures, likely influencing replication efficiency or protein function under the specific culture conditions. For instance, the mutation detection rate near the HA antigenic site (amino acids 180–200) was 1.62×10^{-4}, approximately 1.5 times higher than the genome-wide mutation rate, highlighting a potential hotspot under positive selection. This finding corroborates prior reports of higher sequence variability in antigenic regions(41, 42). On the other hand, the observed mutation rates among genes do not align with the findings from previous phylogenetic research[45] on 'highly conserved' and 'highly divergent' genes, suggesting a lack of correlation between the distribution size and the evolvability of each gene. Such discrepancies may reflect differences in observation timing. Traditional phylogenetic analyses capture fixed mutations shaped by long-term selection, while our study detects earlier-stage mutations that have yet to undergo full selective filtering. Thus, the weak correlation with phylogenetic conservation likely arises because many observed mutations are still under selection. A comparison between RNA extracted from viral populations and *in vitro* transcribed RNA revealed greater protein sequence diversity in the former, as quantified by Shannon entropy. This greater diversity reflects the accumulation of mutations during replication and the latent evolutionary potential of viral populations.

The reference sequences employed for mapping viral genomes in this study were derived from single particles that contributed to the formation of each virus population. Nevertheless, subtle differences were observed among the consensus sequences from four virus populations, suggesting that even within the same PR8 strain, various mutations had accumulated during laboratory passaging, resulting in genetically diverse populations at the outset. This finding suggests that the experimental strain already possessed a mutation pool, and the observed mutation distribution reflects this background diversity. A comprehensive understanding of the effects of long-term passaging on viral population structure and mutation origins is imperative to obtain significant insights.

As this study did not impose specific selective pressures, we did not observe a significant increase of particular mutations previously linked to drug resistance or host adaptation was not observed within the populations. However, resistance mutations such as I38M, which have been demonstrated to confer resistance to the endonuclease inhibitor baloxavir(43, 44), were detected (see Datasets S1 and S2 for a list of all mutations detected). Conversely, mutations fixed in PR8-related strains were already present in populations derived from single particles. These findings imply that the viral quasi-species may serve as a latent genetic reservoir, from which advantageous variants can be selected in response to environmental pressures. While genetic

variation was also detected in HA and NA, we did not impose drug or immune selection pressure in this study. Therefore, we did not expect to observe mutations that are already known to confer major antigenic changes in these proteins, and we consider it difficult to speculate on their functional implications in this context. Nevertheless, the detection of resistance-associated mutations indicates that the quasi-species pool may indeed harbor functionally relevant variation, even in the absence of explicit selective pressures. Thus, the real-time observation of mutation proliferation under diverse culture conditions will yield pivotal insights into the mechanisms underlying existing mutation expansion, thereby facilitating the prediction of novel mutations.

The predominant paradigm in evolutionary biology is the neutral evolution hypothesis, which posits that most evolutionary processes can be explained by random genetic drift. Consequently, elucidating the origins of these evolutionary processes is paramount for making accurate evolutionary predictions. A comprehensive analysis of neutral mutations necessitates the quantification of minor variants. The sUMI method was employed to detect mutations present at 0.1% frequency in populations by sequencing 10,000 molecules. Furthermore, the sequencing error rate was reduced to the order of 10^{-5}, comparable to reverse transcriptase error rates. This enabled theoretical detection of mutations a frequency of 0.05% by analyzing over 100,000 molecules with high accuracy. It is anticipated that this approach will yield comprehensive insights into mutation the occurrence rates and distributions of mutations in neutral evolution. Furthermore, we have demonstrated the applicability of this method for mutation forecasting by using logistic modeling based on mutation fitness and initial frequencies. Subsequent applications will encompass the comprehensive detection and quantitative estimation of adaptive mutations under diverse environmental conditions, including the presence of drugs and different host species.

In summary, experimental evidence has demonstrated the efficacy of UMI technology in reducing sequencing errors and accurately measuring mutation distributions within viral populations. Furthermore, evidence was presented demonstrating that sequence distributions within individual populations manifest non-random directional biases. With the continued development of methods to quantify mutation bias and latent evolutionary potential, it is anticipated that laboratory-scale prediction of drug-induced mutations and pandemic-capable strains will become a reality. The broad distribution of mutations indicates that viral populations possess diverse mutation pools, where selective pressures enhance robustness through adaptive mutation selection. This mechanism signifies the ability of viruses to adapt to environmental changes with flexibility, thereby providing critical insights for predicting long-term viral evolution and pandemic emergence. A more thorough examination of the roles of mutations in the context of adaptive viral evolution in response to drug treatment is merited.

Materials and Methods

1) Cells and Virus

Madin–Darby Canine Kidney (MDCK) cells, which are derived from canine renal tubular epithelial cells, were obtained from the American Type Culture Collection (ATCC). The cells were cultivated in Dulbecco's Modified Eagle Medium (DMEM; FUJIFILM), with the addition of 10% fetal bovine serum (FBS; BioWest) and 1% Penicillin–Streptomycin Solution (P/S; FUJIFILM). The cells were maintained at 37°C in a humidified atmosphere with 5% CO_2.

The A/Puerto Rico/8/1934 (H1N1) strain of influenza A virus was prepared as previously described(45), and stored at −80°C. The virus stocks were meticulously thawed and subsequently diluted to the appropriate concentrations prior to utilization.

379

380 **2) UMI Ligation to Control Plasmid**

381 The experiment utilized a control plasmid containing an HA-like sequence. The plasmid was
382 subjected to a series of digestions with the restriction enzymes NheI and EcoRI, of which were
383 obtained from Nippon Genetics. Following the digestion with NheI, dephosphorylation was carried
384 out. A double-stranded oligonucleotide containing a unique molecular identifier (UMI) and EcoRI-
385 compatible overhangs was ligated to the plasmid using Ligation Mix (Takara Bio).

386

387 **3) *In Vitro* Transcription of Control RNA**

388 Plasmids containing PB2- and HA-like sequences under the control of a T7 promoter were
389 constructed by RT-PCR and In-Fusion cloning from viral RNA extracts, as they were selected for
390 representing both highly and poorly conserved regions across strains, showing substantial
391 sequence divergence, and having relatively large genome sizes that facilitate read acquisition
392 with Revio. *In vitro* transcription was performed using the *In Vitro* Transcription T7 Kit (Takara
393 Bio) with T7 RNA polymerase.

394

395 **4) Single-Virus Particle Infection and RNA Extraction**

396 MDCK cells were cultivated to achieve confluency in 6-well plates. Following the removal of the
397 medium and a thorough washing step with Phosphate Buffered Saline (PBS), the cells were
398 infected with the influenza virus at 0.05 PFU/well. The infection was carried out at 35°C for 1 hour
399 with gentle rocking every 15 minutes. After removing the inoculum, the cells were incubated in 2
400 mL of 1× Minimum Essential Medium (MEM) at 37°C with 5% CO_2. Culture supernatants were
401 collected 3-4 days post-infection from wells where complete cytopathic effect was observed. The
402 extraction of viral RNA was conducted using the QIAamp Viral RNA Mini Kit (QIAGEN). Of the 96
403 wells examined, 4 (4.2%) exhibited signs of successful infection.

404

405 **5) RT-PCR**

406 Reverse transcription was performed using SuperScript IV (Thermo Fisher Scientific), and PCR
407 was carried out using KOD Plus NEO (TOYOBO). The reverse transcription primer consisted of
408 four functional elements arranged from the 5' end to the 3' end: PCR priming site
409 (AGGTTATGCGTTGATACCACTGCTTGC), unique molecular identifier (UMI) composed of 15
410 randomized nucleotides (NNGNNTNNGNNTNNGNNTNNN), population-specific barcode of six
411 nucleotides, and annealing site complementary to the viral genome (AGCRAAAGCAGG). PCR
412 was performed using an average of 4.15 × 10^7 input molecules as the template, as quantified
413 by qPCR targeting the HA gene. The cycling conditions were as follows: initial denaturation at
414 94 °C for 2 min, followed by 25 cycles of denaturation at 98 °C for 10 s and annealing/extension
415 at 68 °C for X s (X = 80 for PB2, PB1, and PA, 60 for HA, NP, and NA, 40 for MP and NS). All
416 primers were custom synthesized by FASMAC.

417

418 **6) Sequencing**

Sequencing using the PacBio Revio platform was outsourced to Takara Bio and Bioengineering Lab.

7) Analyzing sequencing results

1.Read processing and barcode demultiplexing

HiFi reads with an average quality score ≥60 were first processed using SeqKit (v0.16.1). Reverse-complement sequences were generated and combined with the original reads to count reads whose reverse-complement strand was sequenced.

Barcode sequences were extracted from using seqkit grep. Reads corresponding to the reference and sample barcodes were separated into individual FASTQ files.

2. Primer trimming

To remove PCR priming site, cutadapt(46) (v4.0) was applied. Reads matching at least 90% identity were retained, and sequences shorter than 20 nt were discarded:

3. UMI grouping and variant calling

Consensus-based error correction and variant calling were performed with umierrorcorrect (run_umierrorcorrect.py). Each dataset was aligned to the reference sequence, and consensus sequences for each UMI group were remapped. The pipeline produced BAM alignment files, which were opened with Genetyx-NGS, and variant summary tables were compiled in TSV format.

8) Virus particle synthesis and quantification

Virus particle quantification was performed using digital influenza assay(47).

9) Bioinformatic Analysis

Initial demultiplexing of fastq files by population barcode was performed using Seqkit(48). UMI consensus sequences were generated using UMIerrorcorrect(49), and reads were mapped to reference sequences using Genetyx-NGS/Mac. The mapping results were then further processed to extract the base counts at each nucleotide position. The following metrics were calculated:

・Jensen–Shannon Divergence (JSD):

The frequency distribution of sites with k mutations is denoted by P(k). The Poisson distribution with mean count λ is denoted by Q(k). The JSD between P(k) and Q(k) was calculated as follows:

$$D_{KL}(P,Q) = \sum_{x} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

$$M = \frac{1}{2}(P+Q)$$

$$D_{JS}(P,Q) = \frac{1}{2}(D_{KL}(P,M) + D_{KL}(Q,M))$$

・Shannon Entropy:

The frequency distribution of an amino acid sequence "x" translated from an RNA sequence obtained by sequencing is denoted by P(x). The Shannon entropy H is calculated as follows:

$$H = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

10) Growth Simulation Using a Logistic Model

Assumed that the time evolution of the particle numbers of wild-type and mutant species, denoted as N_w and N_m respectively, follows a system of differential equations. This assumption was then used to calculate the temporal changes in the number of each particle type.

$$\frac{dN_w}{dt} = r_w N_w \left(1 - \frac{N_w + N_m}{K}\right)$$
$$\frac{dN_m}{dt} = r_m N_m \left(1 - \frac{N_w + N_m}{K}\right)$$
$$K = 10^{13}, r_w = 1$$

Data availability

The raw sequencing data generated in this study have been deposited in the DDBJ Sequence Read Archive under the BioProject accession number PRJDB35851.

Acknowledgments

This research was supported by the Japan Science and Technology Agency for Core Research for Evolutional Science and Technology, JST CREST, Japan (JPMJCR22N2), and The 2024 - 2025 UTOPIA AI Research Discovery Program, AMED, Japan (JP223fa627001). K. T. is supported by the MERIT-WINGS Program at the University of Tokyo.

References

1. N. J. Cox, T. L. Brammer, H. L. Regnery, Influenza: Global surveillance for epidemic and pandemic variants. *Eur J Epidemiol* **10**, 467–470 (1994).

2. I. G. Barr, *et al.*, Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: Basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009–2010 Northern Hemisphere season. *Vaccine* **28**, 1156–1167 (2010).

488 3. W. K. Ampofo, *et al.*, Improving influenza vaccine virus selectionReport of a WHO
489 informal consultation held at WHO headquarters, Geneva, Switzerland, 14–16
490 June 2010. *Influenza Other Respir Viruses* **6**, 142–152 (2012).

491 4. J. K. Taubenberger, D. M. Morens, 1918 Influenza: the Mother of All Pandemics.
492 *Emerg Infect Dis* **12**, 15–22 (2006).

493 5. N. Pardi, M. J. Hogan, F. W. Porter, D. Weissman, mRNA vaccines — a new era in
494 vaccinology. *Nat Rev Drug Discov* **17**, 261–279 (2018).

495 6. E. A. Belongia, *et al.*, Variable influenza vaccine effectiveness by subtype: a
496 systematic review and meta-analysis of test-negative design studies. *Lancet Infect*
497 *Dis* **16**, 942–951 (2016).

498 7. Y. Zhu, A. Yongky, J. Yin, Growth of an RNA virus in single cells reveals a broad
499 fitness distribution. *Virology* **385**, 39–46 (2009).

500 8. J. D. Parvin, A. Moscona, W. T. Pan, J. M. Leider, P. Palese, Measurement of the
501 mutation rates of animal viruses: influenza A virus and poliovirus type 1. *J Virol* **59**,
502 377–383 (1986).

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

505 10. I. Skountzou, *et al.*, Incorporation of Glycosylphosphatidylinositol-Anchored
506 Granulocyte- MacrophageColony-Stimulating Factor or CD40 Ligand Enhances
507 Immunogenicity of Chimeric Simian Immunodeficiency Virus-Like Particles. *J Virol*
508 **81**, 1083–1094 (2007).

509 11. B. F. Koel, *et al.*, Substitutions Near the Receptor Binding Site Determine Major
510 Antigenic Change During Influenza Virus Evolution. *Science (1979)* **342**, 976–979
511 (2013).

512 12. I. Deznabi, B. Arabaci, M. Koyutürk, O. Tastan, DeepKinZero: zero-shot learning for
513 predicting kinase–phosphosite associations involving understudied kinases.
514 *Bioinformatics* **36**, 3652–3661 (2020).

515 13. N. N. Thadani, *et al.*, Learning from prepandemic data to forecast viral escape.
516 *Nature* **622**, 818–825 (2023).

517 14. Q. Yan, J. Zhao, UniBind: a novel artificial intelligence-based prediction model for
518 SARS-CoV-2 infectivity and variant evolution. *Signal Transduct Target Ther* **8**, 464
519 (2023).

520 15. S. A. W. Shah, *et al.*, Seasonal antigenic prediction of influenza A H3N2 using
521 machine learning. *Nat Commun* **15**, 3833 (2024).

522 16. S. Dellicour, *et al.*, Dispersal dynamics of SARS-CoV-2 lineages during the first
523 epidemic wave in New York City. *PLoS Pathog* **17**, e1009571 (2021).

524 17. E. Domingo, J. Sheldon, C. Perales, Viral Quasispecies Evolution. *Microbiology and*
525 *Molecular Biology Reviews* **76**, 159–216 (2012).

526 18. K. M. Peck, A. S. Lauring, Complexities of Viral Mutation Rates. *J Virol* **92** (2018).

527 19. R. Sanjuán, A. Moya, S. F. Elena, The distribution of fitness effects caused by
528 single-nucleotide substitutions in an RNA virus. *Proceedings of the National*
529 *Academy of Sciences* **101**, 8396–8401 (2004).

530 20. J. J. Bull, R. Sanjuán, C. O. Wilke, "Lethal Mutagenesis" in *Origin and Evolution of*
531 *Viruses*, (Elsevier, 2008), pp. 207–218.

532 21. M. Wang, *et al.*, Viral quasispecies quantitative analysis: a novel approach for
533 appraising the immune tolerant phase of chronic hepatitis B virus infection.
534 *Emerg Microbes Infect* **10**, 842–851 (2021).

535 22. J. Sardanyés, C. Perales, E. Domingo, S. F. Elena, Quasispecies theory and
536 emerging viruses: challenges and applications. *npj Viruses* **2**, 54 (2024).

537 23. D. Zucha, P. Androvic, M. Kubista, L. Valihrach, Performance Comparison of
538 Reverse Transcriptases for Single-Cell Studies. *Clin Chem* **66**, 217–228 (2020).

539 24. S. Martín-Alonso, E. Frutos-Beltrán, L. Menéndez-Arias, Reverse Transcriptase:
540 From Transcriptomics to Genome Editing. *Trends Biotechnol* **39**, 194–210 (2021).

541 25. D. H. Westfall, *et al.*, "Optimized SMRT-UMI protocol produces highly accurate
542 sequence datasets from diverse populations-Application to HIV-1 quasispecies."

543 26. M. L. Menéndez, J. A. Pardo, L. Pardo, M. C. Pardo, The Jensen-Shannon
544 divergence. *J Franklin Inst* **334**, 307–318 (1997).

545 27. C. E. Shannon, A Mathematical Theory of Communication. *Bell System Technical*
546 *Journal* **27**, 379–423 (1948).

547 28. A. C. Daly, K. J. Geras, R. Bonneau, A convolutional neural network for common
548 coordinate registration of high-resolution histology images. *Bioinformatics* **37**,
549 4216–4226 (2021).

550 29. R. Pourmohammadi, J. Abouei, A. Anpalagan, Error analysis of the PacBio
551 sequencing CCS reads. *Int J Biostat* **19**, 439–453 (2023).

552 30. M. D. Pauly, M. C. Procario, A. S. Lauring, A novel twelve class fluctuation test
553 reveals higher than expected mutation rates for influenza A viruses. *Elife* **6**, 1–18
554 (2017).

555 31. S. Duchêne, S. Y. Ho, E. C. Holmes, Declining transition/transversion ratios through
556 time reveal limitations to the accuracy of nucleotide substitution models. *BMC*
557 *Evol Biol* **15**, 36 (2015).

558 32. F. G. Alnaji, W. K. Reiser, J. Rivera-Cardona, A. J. W. te Velthuis, C. B. Brooke,
559 Influenza A Virus Defective Viral Genomes Are Inefficiently Packaged into Virions
560 Relative to Wild-Type Genomic RNAs. *mBio* **12** (2021).

561 33. R. Penn, *et al.*, Levels of Influenza A Virus Defective Viral Genomes Determine
562 Pathogenesis in the BALB/c Mouse Model. *J Virol* **96** (2022).

563 34. J. Hedge, S. J. Lycett, A. Rambaut, Real-time characterization of the molecular
564 epidemiology of an influenza pandemic. *Biol Lett* **9** (2013).

565 35. P. Skums, *et al.*, QUENTIN: reconstruction of disease transmissions from viral
566 quasispecies genomic data. *Bioinformatics* **34**, 163–170 (2018).

567 36. A. Bing, *et al.*, Comparison of empirical and dynamic models for HIV viral load
568 rebound after treatment interruption. *Stat Commun Infect Dis* **12** (2020).

569 37. A. Kartono, S. T. Wahyudi, A. A. Setiawan, I. Sofian, Predicting of the Coronavirus
570 Disease 2019 (COVID-19) Epidemic Using Estimation of Parameters in the Logistic
571 Growth Model. *Infect Dis Rep* **13**, 465–485 (2021).

572 38. Y. Qu, *et al.*, Evolutionary genomics of the pandemic 2009 H1N1 influenza viruses
573 (pH1N 1v). *Virol J* **8** (2011).

574 39. Q. Zhuang, *et al.*, Diversity and distribution of type A influenza viruses: an updated
575 panorama analysis based on protein sequences. *Virol J* **16**, 85 (2019).

576 40. G. J. D. Smith, *et al.*, Origins and evolutionary genomics of the 2009 swine-origin
577 H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).

578 41. B. Thyagarajan, J. D. Bloom, The inherent mutational tolerance and antigenic
579 evolvability of influenza hemagglutinin. *Elife* **3** (2014).

580 42. N. C. Wu, *et al.*, Major antigenic site B of human influenza H3N2 viruses has an
581 evolving local fitness landscape. *Nat Commun* **11**, 1233 (2020).

582 43. J. C. Jones, P. N. Q. Pascua, W. N. Harrington, R. J. Webby, E. A. Govorkova,
583 Multiple polymerase acidic (PA) I38X substitutions in influenza A(H1N1)pdm09
584 virus permit polymerase activity and cause reduced baloxavir inhibition. *Journal of*
585 *Antimicrobial Chemotherapy* **76**, 957–960 (2021).

586 44. K. Taniguchi, *et al.*, The impact of PA/I38 substitutions and PA polymorphisms on
587 the susceptibility of zoonotic influenza A viruses to baloxavir. *Arch Virol* **169**, 29
588 (2024).

589 45. A. J. Eisfeld, G. Neumann, Y. Kawaoka, Influenza A virus isolation, culture and
590 identification. *Nat Protoc* **9**, 2663–2681 (2014).

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

593 47. K. V. Tabata, *et al.*, Antibody-free digital influenza virus counting based on
594 neuraminidase activity. *Sci Rep* **9**, 1–13 (2019).

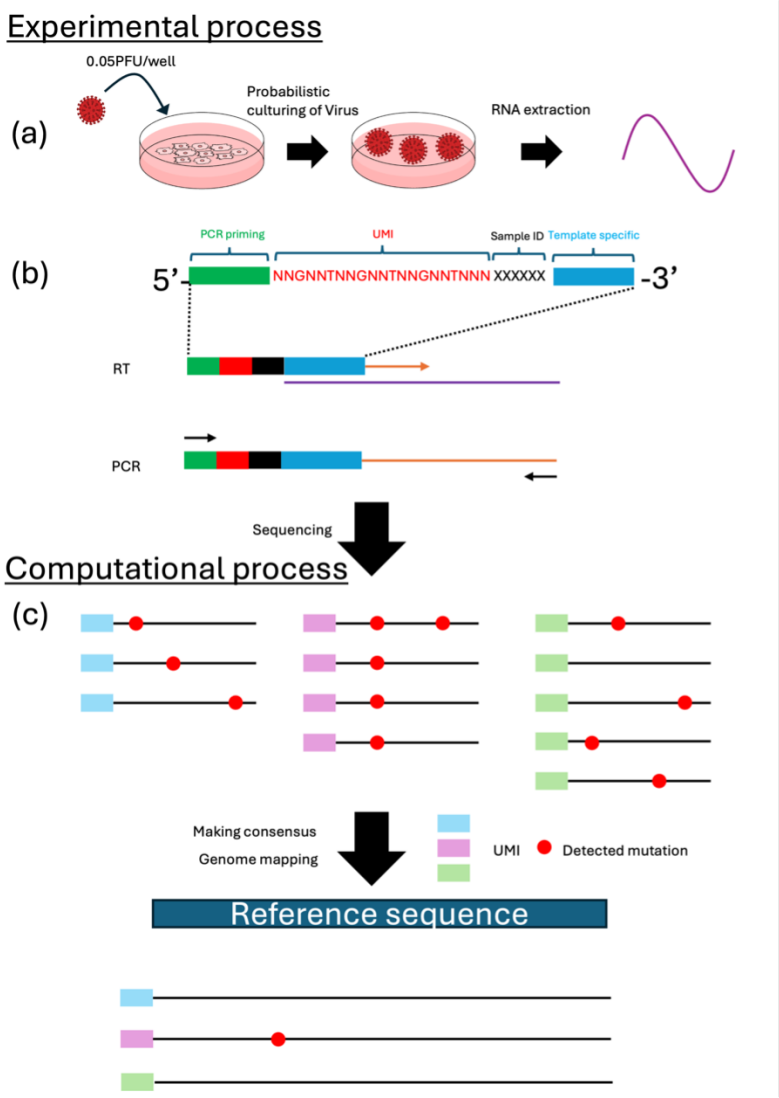
595 48. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A cross-platform and ultrafast toolkit for
596 FASTA/Q file manipulation. *PLoS One* **11**, 1–10 (2016).

597 49. T. Österlund, S. Filges, G. Johansson, A. Ståhlberg, UMIErrorCorrect and
598 UMIAnalyzer: Software for Consensus Read Generation, Error Correction, and
599 Visualization Using Unique Molecular Identifiers. *Clin Chem* **68**, 1425–1435 (2022).

600

601 **Figures and Tables**

602



603

Fig. 1 Overview of the process of obtaining genomic diversity of influenza virus in a population replicated from a single particle.

(a) MDCK cells were infected with influenza virus at a concentration of 0.05 PFU/well. After 3-4 days, supernatants were collected from wells where cellular tissue destruction had occurred, and viral RNA was extracted.

(b) The RT primer consisted of a 25-nucleotide PCR priming site at the 5' end, followed by a 6-nucleotide sample ID, a 21-nucleotide UMI containing a 15-nucleotide random nucleotide, and a 12-nucleotide sequence complementary to the template. cDNA was synthesized via reverse transcription, followed by PCR amplification using primers containing the PCR priming site sequence and gene-specific sequence.

(c) The sequences were grouped according to their UMI. For each unique UMI, the original RNA sequence was reconstructed by majority voting among identical sequences. The reconstructed RNA molecules were mapped to the reference genome, and the distribution of mutations was quantified.

604

605

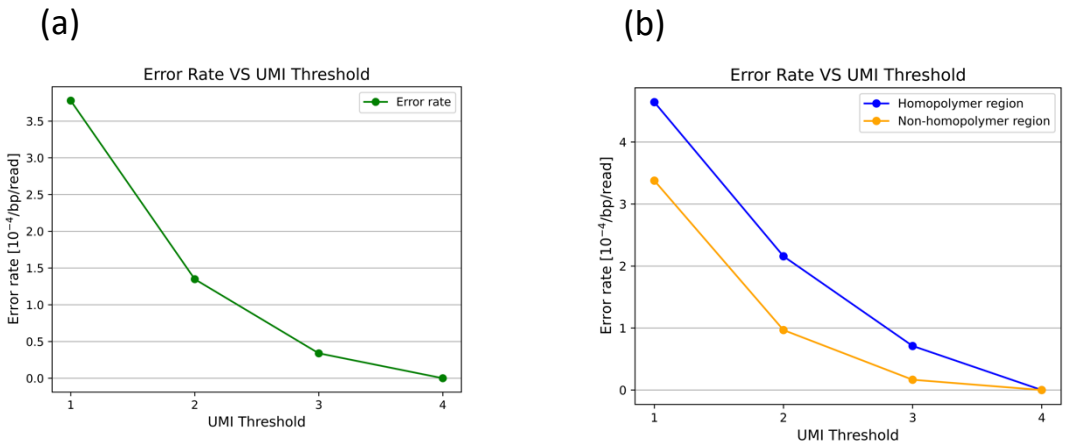


Fig. 2 Error rates of UMI-ligated plasmid filtered by different group size

The plasmid, replicated by cloning in E. coli. was linearized using restriction enzyme treatment. A unique molecular identifier (UMI) was then attached through ligation, followed by amplification via PCR and subsequent sequencing. After sequencing, consensus sequences were obtained from reads sharing the same UMI, allowing for the reconstruction of single-molecule sequences.

(a) Changes in the error rate of linearized plasmids when filtered based on group size (UMI duplication detection count).

(b) Changes in the error rate in homopolymer and non-homopolymer regions when filtered based on group size.

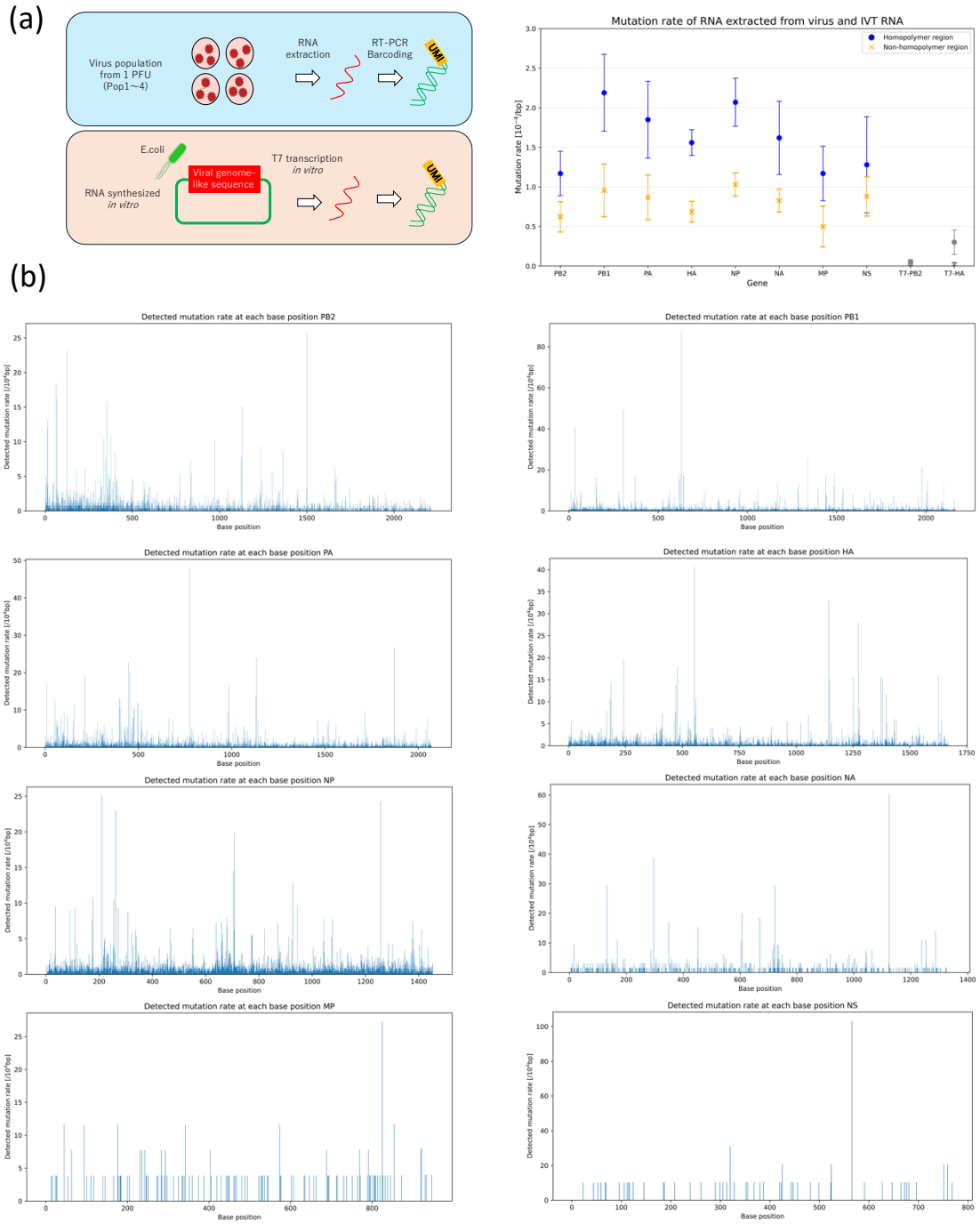


Fig. 3 Error rates obtained from virus extracted RNA and in vitro transcribed RNA

(a) Sequencing results of RNA extracted from a virus population propagated from a single molecule and RNA synthesized by *in vitro* transcription from a plasmid. The viral RNA represents the average mutation rate obtained from four virus populations, with bars indicating the standard deviation. The *in vitro* transcribed RNA represents the average of two experiments, with bars indicating the standard deviation.

(b) The bar graph of mutation rates obtained at each position within the coding region of the extracted viral RNA. The horizontal axis represents the nucleotide position within the coding region.

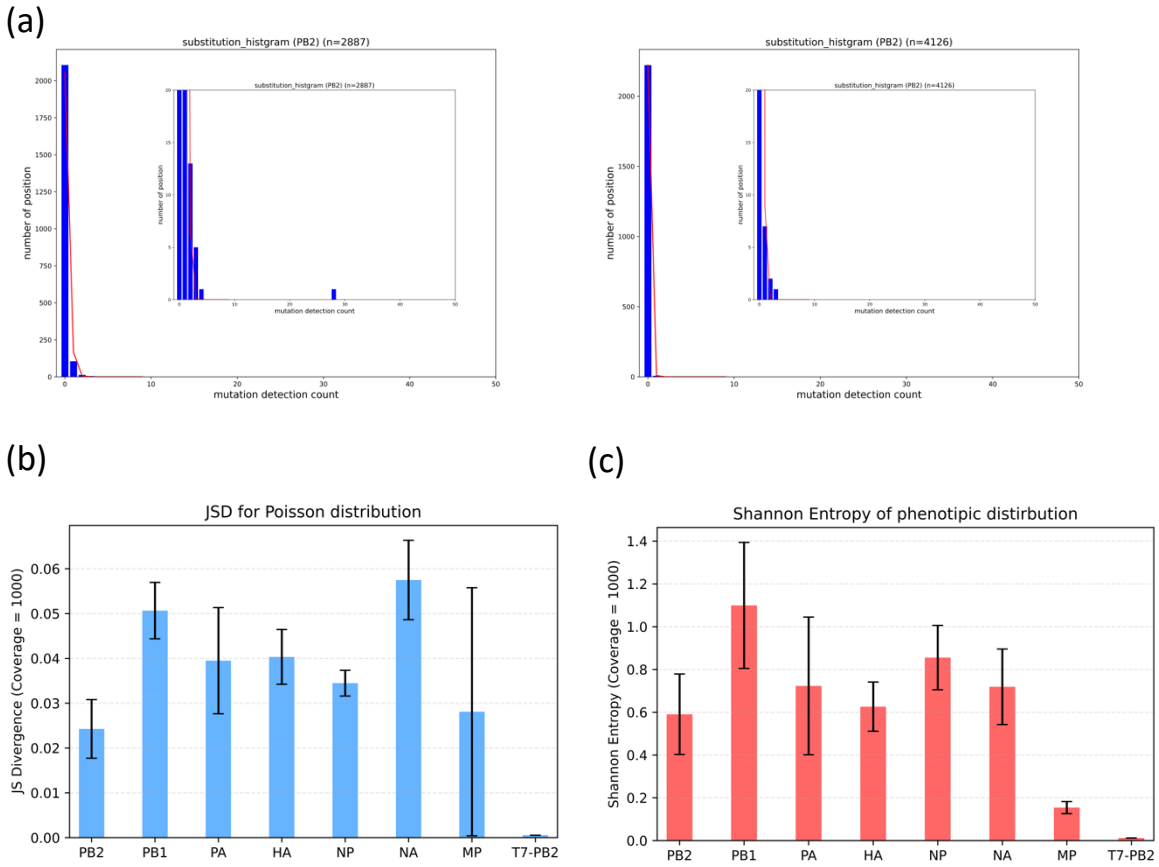


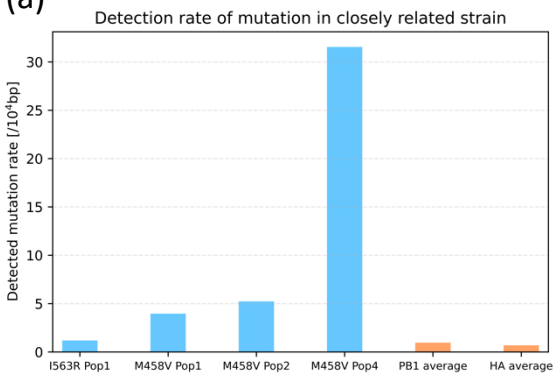
Fig. 4 The Distance from the Poisson Distribution and Sequence Diversity in Viral Extracted RNA and In Vitro Transcribed RNA

(a) Distribution of mutation detection counts by position in the PB2 gene extracted from Population1 virus (Left) and PB2-like sequence RNA synthesized via *in vitro* transcription (Right). The horizontal axis represents the number of times a mutation was detected, while the vertical axis represents the number of positions where mutations with the corresponding detection count were observed. The fitting line represents the Poisson distribution calculated based on the mean value for each dataset.

(b) Jensen-Shannon Diversity between the Poisson distribution and each gene extracted from the virus or the PB2-like RNA synthesized via *in vitro* transcription. To ensure equal conditions across genes, distributions were normalized to a sequencing coverage of 1000. The NS gene was not included in the graph because no population had sequencing coverage exceeding 1000. The bar height for the viral populations represents the mean value across four populations, and the error bars indicate the standard deviation.

(c) Shannon entropy of each gene extracted from the virus and the PB2-like gene synthesized via *in vitro* transcription. To ensure equal conditions across genes, sequencing results from 1000 molecules were randomly sampled, and the Shannon entropy was calculated for each sample. This procedure was repeated 10 times, and the average value was used as the Shannon entropy for a given gene in a given population. The bar height for the viral populations represents the mean value across four populations, and the error bars indicate the standard deviation.

(a)



(b)

Gene	Homology	Mutation detected in the experiment
PB2	94.6%	R121K
PB1	95.0%	R54K, M325I, S361R, M645V, V682I
PA	95.8%	Q57R, R204K
HA	86.2%	N4K, D52N, S102P
NP	90.8%	K236R, T423A, R452K
NA	80.4%	
MP	92.8%	
NS	84.3%	

Fig. 5 Detection Rate of Variants in Related Strains and Variants Matching the Latest Strain

(a) Among the three mutations observed between the strain used in this experiment (PR8) and its related strain (Alaska/1935), the detection rates of the two mutations observed in the cultured population of this experiment within each population are presented. The orange bars represent the average mutation rate in non-homopolymer regions across all populations for PB1 and HA.

(b) A list of mutations observed in the cultured population of this experiment among the multiple mutations found between PR8 and the latest circulating strain (CA01). "Homology" indicates the amino acid sequence similarity between PR8 and CA01.

Supporting Information for

Heterogeneity of Genetic Sequence within Quasi-species of Influenza Virus Revealed by Single-Molecule Sequencing

Kenji Tamao[1], Hiroyuki Noji[1,2], Kazuhito V. Tabata[1,2*]

1 Department of Applied Chemistry, Graduate School of Engineering, the University of Tokyo, Tokyo, Japan

2 Research Institute of Planetary Health, The University of Tokyo, Japan.

*Corresponding author: Kazuhito V.Tabata

Email: tabatak@g.ecc.u-tokyo.ac.jp

This PDF file includes:

Figures S1 to S4

Tables S1

Other supporting materials for this manuscript include the following:

Datasets S1 to S2

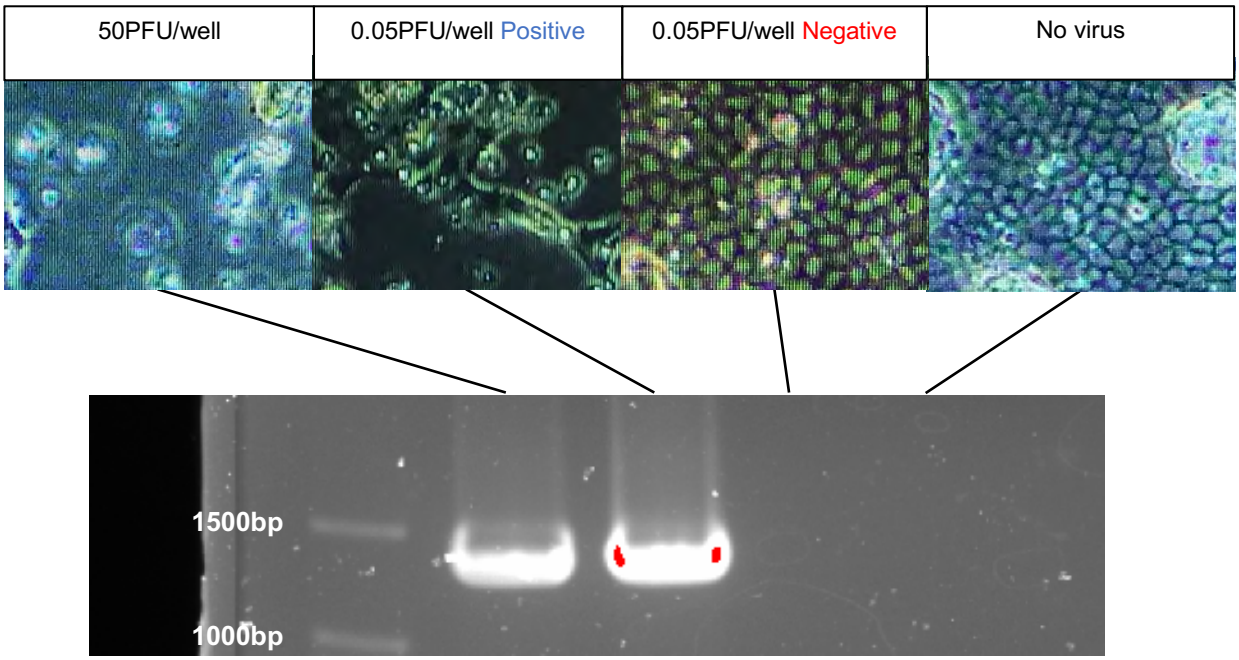


Fig. S1. Population formation from single PFU

Viral populations were generated from a single PFU of influenza virus by infecting cells with diluted virus. Each figure shows a microscopic image of MDCK cells 96 hours post-infection. The leftmost image shows infection at 50 PFU/well, in which the cell monolayer is completely destroyed. The rightmost figure shows virus-free conditions, in which the cell monolayer is maintained. The two middle images show infection at 0.05 PFU/well, where some wells show cell destruction while others retain an intact cell monolayer. The relationship between cell destruction and viral replication was confirmed by RT-PCR targeting the NA gene (1347 nt).

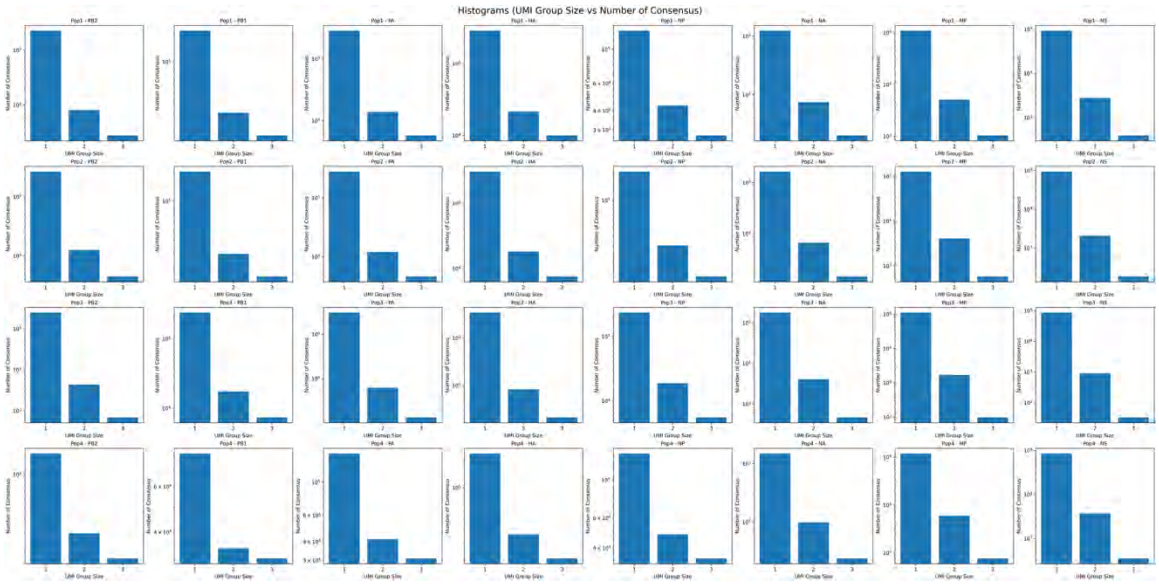


Fig. S2. Distribution of the number of consensus sequences after filtering based on UMI group size for each gene and population

The histogram shows how many consensus sequences remain after applying a filter on UMI group size, illustrating the impact of filtering on sequence counts.

(a) (b)

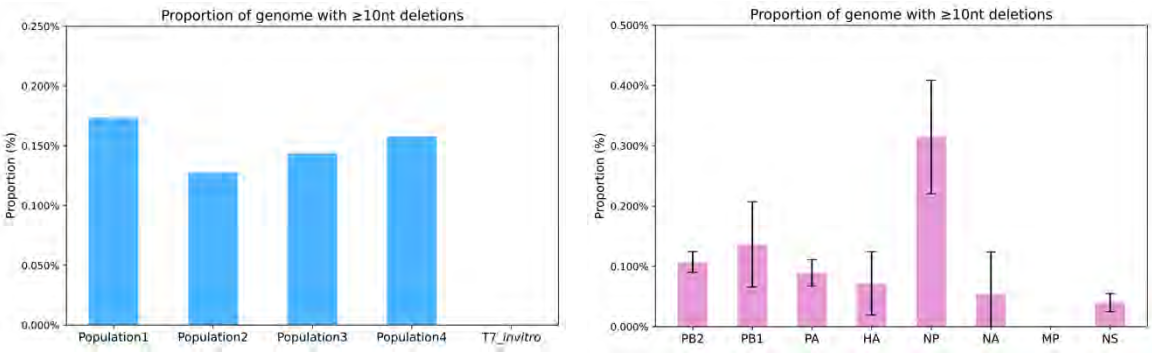


Fig. S3. The proportion of sequences with ≧ 10nt deletions

Proportion of sequences with consecutive deletions of 10 or more bases a) in RNA extracted from each viral population and in in vitro transcribed RNA, and b) within each gene. Error bars represent the weighted standard deviation (SD).

(a) (b)

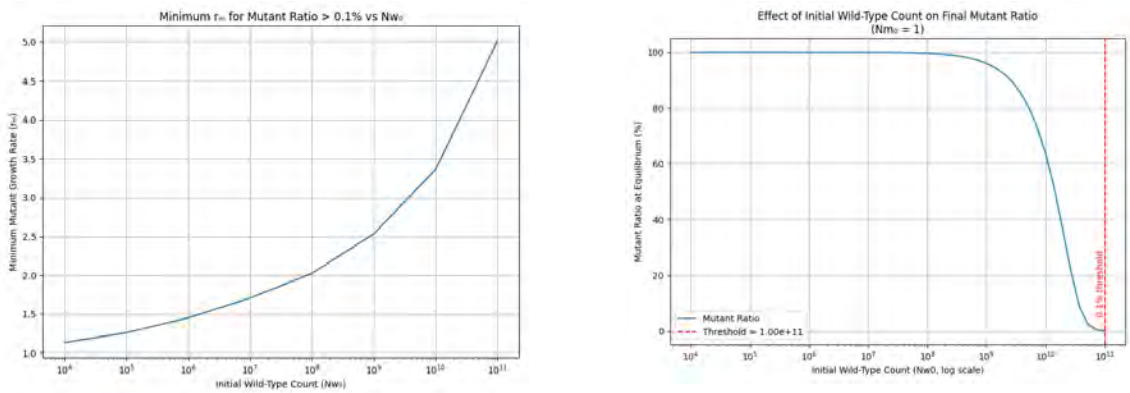


Fig. S4 Prediction of detection rates of mutant strains using a logistic growth model

(a) Minimum r_m values required for the final mutant fraction to exceed 0.1% in each initial condition.

(b) Final mutant fractions for each initial condition when r_m is fixed at 5 times than r_w

121 **Table S1. Overview of the sequencing results from virus extracted RNA**

Gene	mutation rate[/10^4nt]	Ti/Tv	dN/dS
PB2	1.17 ± 0.28	2.84 ± 0.32	0.74 ± 0.13
PB1	2.19 ± 0.49	3.58 ± 0.49	0.86 ± 0.10
PA	1.85 ± 0.49	2.98 ± 0.16	0.76 ± 0.05
HA	1.56 ± 0.16	2.46 ± 0.12	0.68 ± 0.04
NP	2.07 ± 0.30	3.62 ± 0.31	0.69 ± 0.07
NA	1.62 ± 0.46	1.91 ± 0.66	1.05 ± 0.15
MP	1.17 ± 0.34	4.58 ± 1.74	0.53 ± 0.07
NS	1.28 ± 0.61	3.13 ± 2.69	1.63 ± 1.01

122

123 **Dataset S1 (separate file).** A list of all mutation counts detected, written in amino acid form

124 **Dataset S2 (separate file).** A list of all mutation counts detected, written in nucleotide form

125