The Effects of Statistical Multiplicity of Infection on Virus Quantification and Infectivity Assays

B. Mistry, M. D'Orsogna, T. Chou

November 22, 2017

Abstract

Many biological assays are employed in virology to quantify parameters of interest. Two such classes of assays, virus quantification assays (VQA) and infectivity assays (IA), aim to estimate the number of viruses present in a solution, and the ability of a viral strain to successfully infect a host cell, respectively. VQAs operate at extremely dilute concentrations and results can be subject to stochastic variability in virus-cell interactions. At the other extreme, high viral particle concentrations are used in IAs, resulting in large numbers of viruses infecting each cell, enough for measurable changes in total transcription activity. Furthermore, host cells can be infected at any concentration regime by multiple particles, resulting in a statistical multiplicity of infection (SMOI) and yielding potentially significant variability in the assay signal and parameter estimates. We develop probabilistic models for SMOI at low and high viral particle concentration limits and apply them to the plaque (VQA), endpoint dilution (VQA), and luciferase reporter (IA) assays. We test our proposed new methods for inferring experimental parameters from data using numerical simulations and show improvement on existing procedures in all limits.

1 Introduction

Understanding viral dynamics is an important task in medicine, epidemiology, public health, and, in particular, for the development of antiviral therapies and vaccines. Drugs that hinder viral infection include blockers of viral entry into the host cell (1–6) and inhibitors of genetic activity and protein assembly inside the cytoplasm and nucleus (7–9). Mechanistic models of drug action have recently emerged as useful tools in helping design ad-hoc experiments to study drug efficacy and in interpreting results (10–13). Mathematical models typically assume prior knowledge of given physical quantities pertaining to the virus, host cell, or the biological assay being studied. Once these parameters are assigned, viral and cell population dynamics and their statistical properties can be predicted. Among the different experimental assays, one often seeks to evaluate the number of virus particles in a stock solution or the number of viruses that have successfully infected host cells (6, 14–19).

In the case of virus quantification assays (VQA), performing repeated controlled experiments on viral dynamics or comparing results across multiple studies requires knowing how many viruses are present in the initial stock solution of each experiment (4, 5). Similarly, antigens that induce immune responses against viral infections may be engineered from viral constituent parts such as capsid proteins, viral enzymes, and genetic vectors (20), and may be used in the development of vaccines. Being able to determine the exact number of virus-derived antigens helps control the efficacy of vaccines and optimize yield (21–23).

Given the central role of VQAs, several assays have been designed to estimate viral particle counts. These include plaque (24) and endpoint dilution (23, 25) assays, which will be discussed in more detail in the remainder of this work. For now, we note that these assays involve repeatedly diluting an initial solution of virus particles in the presence of a layer of plated cells, until viral concentrations are low enough that the dynamics of an individual virus can be extrapolated. At these low particle counts, however, the discrete nature of the infection process cannot be neglected and can cause substantial discrepancies when replicating experiments. Average quantities are not necessarily representative, and a more in-depth approach in quantifying virus-cell interactions is necessary.

Infectivity assays (IA) on the other hand aim to quantify the number of viruses that have successfully infected host cells under varying antiviral drug environments (14–16). IAs may measure the total transcription activity across all cells, such as the luciferase reporter assay (15, 26), or may count the number of host cells that were successfully infected, such as the enzyme-linked immunosorbent assay (ELISA) and the immunofluorence assay with fluorescence activated cell sorting (FACS) (4, 14, 15, 27). These assays are performed using undiluted solutions with large numbers of viral particles, reducing stochastic variability. The average number of viruses that infect a cell is estimated as the ratio of the number of viruses in solution to the number of plated cells, a quantity known as the multiplicity of infection (MOI) (19). However each cell may be infected by varying numbers of viruses distributed around the average given by the MOI. In these cases, one may be interested in the complete probability distribution for the number of virus infections in each plated cell and in the related statistical variance.

In this paper we will derive a probability model for the distribution of viral infections per host cell which we call the statistical multiplicity of infection (SMOI). The SMOI can be used as a starting point to help estimate the number of viral particles in solution in VQAs, and to determine a viral strain's ability to successfully infect host cells in IAs. In Section 2.1, we present the mathematical foundations for the SMOI in the two experimentally relevant parameter regimes of small and large viral particle counts and derive a probability model for the total number of infected cells under any dilution level. In Section 3.1, we apply our models to the plaque assay and formulate a new method of analyzing plaque count data. In Section 3.2 we employ a special case of the derived probability distribution to the endpoint dilution assays and compare our results to those arising from traditional titration techniques such as the Reed and Muench (28) and Spearman-Karber methods (29). In Section 3.3, we use the large particle limit of our model to describe the luciferase reporter assay. A discussion of our results and a side-by-side comparison with existing methods are provided in Section 4.

2 Methods

2.1 Probabilistic Models of Statistical Multiplicity of Infection (SMOI)

An infection assay is initiated by laying a monolayer of M cells to the bottom of a microtiter well, as illustrated in Fig. 1 (17, 24, 25). Though variability exists among experiments, M is typically in the range of 10^4–10^5 (14, 26) and is assumed to be a known experimental parameter. A supernatant containing N_0 virus particles in the range of 10^5–10^7 (24–26), is then added to the microtiter well. Not all N_0 particles are capable of successful infection. This may be due to lethal mutations in the viral genome or damage to the virus glycoprotein spikes required for receptor binding and entry into the cell (30, 31). Furthermore, infection of a host cell requires a complex sequence of biochemical processes that may include receptor binding, membrane fusion, reverse transcription, nuclear pore transport, and DNA integration (10, 19). Virus particles that fail at one or several of these steps lead to abortive infections. To differentiate, the particles that do succeed are called infectious units

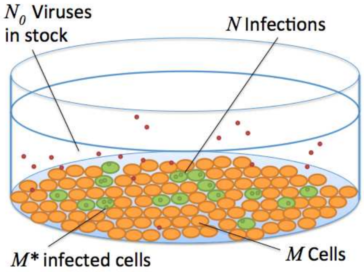


Table of Parameters	
N_0	Number of viruses
N	Number of infections
M	Number of host cells
M^*	Number of infected cells
M_r	Number of cells infected by exactly r viruses
Q	Particle to PFU ratio
D	Dilution factor
T	Number of assay trials
$P_{d,t}$	Plaque assay data
E_d	Endpoint dilution data
L_t^{data}	Luciferase reporter assay data

Figure 1: A typical assay includes a plate of M host cells inoculated with a solution of N_0 viruses. Each viral particle has some probability of infection and the total number N of infections are distributed to the M^* infected cells. The probability of infection is roughly estimated with the reciprocal of the *a priori* measured particle to PFU ratio Q.

(IU) or plaque forming units (PFU). We will denote the number of IUs as $N \leq N_0$. A statistical measure of the relationship between N_0 and N is given by the "particle to PFU ratio" Q, an experimentally determined parameter associated with each viral species that quantifies, on average, how many initial viral particles are required to successfully infect a single host cell (30, 32). Low values of Q, such as with poliovirus ($Q = 30$) (32), have a high likelihood of successful infection compared to viruses with large Q, such as HIV-1 ($Q = 10^7$) (33). Thus, the reciprocal Q^{-1} can be interpreted as the probability for a single virus to infect a host cell. Assuming an initial stock of N_0 particles, the discrete probability density function of N is

$$\Pr\left(N = n | N_0, Q\right) = \binom{N_0}{n} \left(Q^{-1}\right)^n \left(1 - Q^{-1}\right)^{N_0 - n}, \tag{1}$$

which defines a binomial distribution with parameters N_0 and Q^{-1}. Whereas N, N_0 may or may not be known, throughout this work we will assume Q to be an *a priori* measured quantity. In actuality, the probability of a virus successfully infecting a host is highly dependent on the methods used to harvest the virus stock, the experimental parameters of the assay, the host receptor concentrations and binding rates, and the dynamics of the physiological processes leading to infection (30, 34). A thorough investigation into these processes would be necessary to model this probability and is outside of the scope of this paper.

We assume each viral particle in solution acts independently of others and that host cell infection attempts are random events. At high ratios N_0/M of particles to cells, a quantity referred to as the "multiplicity of infection" (MOI), it becomes increasingly probable for more than one IU to infect the same host cell. We can define M_0 as the count of cells not infected by any IU, M_1 as the count of cells infected by exactly one IU, up to M_N, the number of cells infected by all N viable IUs. The statistical multiplicity of infection (SMOI) can now be defined as the ensemble of cell counts $\{M_0, M_1, \cdots, M_N\}$. Note that two constraints must hold: $\sum_{r=0}^{N} M_r = M$ to account for all infected and un-infected cells, and $\sum_{r=0}^{N} rM_r = N$ for conservation of the total number of IUs. Since the M cells are assumed to be of identical size and volume, they carry equal probability of being infected by a particular virus. Thus, evaluating the probability distribution of M_r reduces to the well-known problem of randomly placing balls into bins of equal size (35) and we derive

$$\Pr(M_r = m_r | M, N) = \sum_{j=m_r}^{M} \binom{j}{m_r} \binom{M}{j} \binom{N}{r, \cdots, r, (N - rj)} \frac{(-1)^{j - m_r} (M - j)^{N - rj}}{M^N}, \tag{2}$$

where we use the binomial and multinomial coefficients. The derivation of Eq. 2 is detailed in Appendix 4 and an investigation into the effects of inhomogeneous cell sizes is presented in Appendix 4. Furthermore, in Appendix 4, we derive the expected value and variance of M_r as

$$\mathrm{E}\left[M_r\right] = M \binom{N}{r} \left(\frac{1}{M}\right)^r \left(1 - \frac{1}{M}\right)^{N - r}, \tag{3}$$

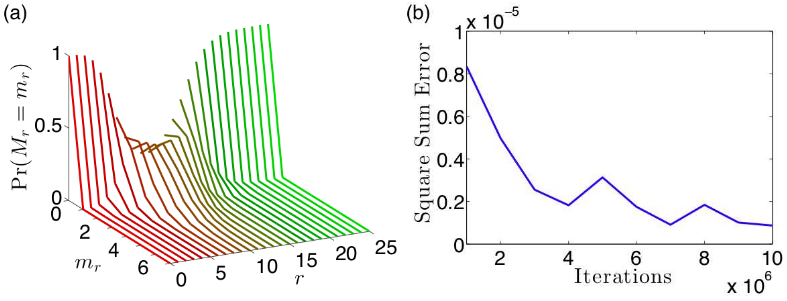


Figure 2: (a) A collection of plots of the probability of finding m_r cells that have been infected by exactly r IUs given a total number of IUs $N = 100$ and a total number of cells $M = 10$ using Eq. 2. With $N/M = 10$, we expect very few cells to be uninfected, resulting in the probability distribution concentrated close to 0 for low values of r. Similarly, we expect few cells to be infected by a very large number of IUs, accumulating the probability distribution close to 0 for large r. Only at intermediate values of $r \approx N/M = 10$ we observe a Poisson-like distribution. (b) A numerical comparison between our analytical result in Eq. 2 and the statistical frequency of virus-cell counts from a simulation of $N = 100$ IUs being randomly assigned to $M = 10$ cells. The square sum error between the simulated proportions and the analytical result was calculated with increasing number of iterations of the simulation. For iterations around 10^6, our square sum error is on the order of 10^{-6}, indicating strong confidence in our analytical result.

and

$$\mathrm{Var}\left[M_r\right] = M\binom{N}{r}\left(\frac{1}{M}\right)^r\left(1-\frac{1}{M}\right)^{N-r} + \frac{M(M-1)N!(M-2)^{N-2r}}{(r!)^2(N-2r)!M^N} - \frac{M^2(N!)^2(M-1)^{2N-2r}}{(r!)^2\left[(N-r)!\right]^2 M^{2N}}. \tag{4}$$

Note that the variance is equal to the expected value with an additional correction term that vanishes as N and M increase, indicating the probability distribution of M_r is Poisson-like for large N and M. A qualitative plot of the probability distribution and a numerical check of its accuracy is provided in Fig. 2.

We also derive the joint probability $\Pr(M_0 = m_0, \cdots, M_N = m_N|M, N)$ that the SMOI $\{M_0, M_1, \cdots, M_N\}$ takes on the set of values $\{m_0, m_1, \cdots, m_N\}$ as

$$\begin{aligned}\Pr(M_0 = m_0, \cdots, M_N = m_N|M, N) &= \frac{1}{M^N}\binom{M}{m_0, m_1, \cdots, m_N}\binom{N}{0, \cdots, 0, 1, \cdots, 1, \cdots, N, \cdots, N} \\ &= \frac{M!N!}{M^N}\prod_{r=0}^{N}\frac{1}{m_r!\,(r!)^{m_r}}. \end{aligned}\tag{5}$$

The first and second multinomial expressions enumerate the degeneracy of how the M identical cells are distributed across the configuration $\{m_0, \cdots, m_N\}$ and how the N identical IUs are chosen for those cells respectively. Although the second expression in Eq. 5 is more succinct, it must be explicitly conditioned on the constraints $\sum_{r=0}^{N} m_r = M$ and $\sum_{r=0}^{N} rm_r = N$.

The expressions in Eqs. 2 and 5 provide an exact discrete description of the stochasticity of the MOI, but are computationally expensive to evaluate for large values of N and M. In a typical virology experiment, the number of viral particles N_0 and host cells M are large enough for certain asymptotic methods to be applicable. Furthermore, for intermediate values of Q and based on Eq. 1, the expected number of IUs N would be similarly large. We can thus take the mathematical limit $N, M \to \infty$ while keeping the ratio $\mu = \frac{N}{M}$ fixed and approximate Eq. 2 as:

$$\Pr(M_r = m_r|M, N) \approx \frac{1}{m_r!}\left[\frac{M\mu^r e^{-\mu}}{r!}\right]^{m_r}\exp\left[-\frac{M\mu^r e^{-\mu}}{r!}\right]. \tag{6}$$

Eq. 6 implies that M_r is Poisson-distributed with mean and variance

$$\mathrm{E}[M_r] = \mathrm{Var}[M_r] \approx \frac{M\mu^r e^{-\mu}}{r!}. \tag{7}$$

A mathematical justification of Eq. 6 is given in Appendix 4 and a visual comparison of Eq. 6 and the analytical result in Eq. 2 with simulation is provided in Fig. 3.

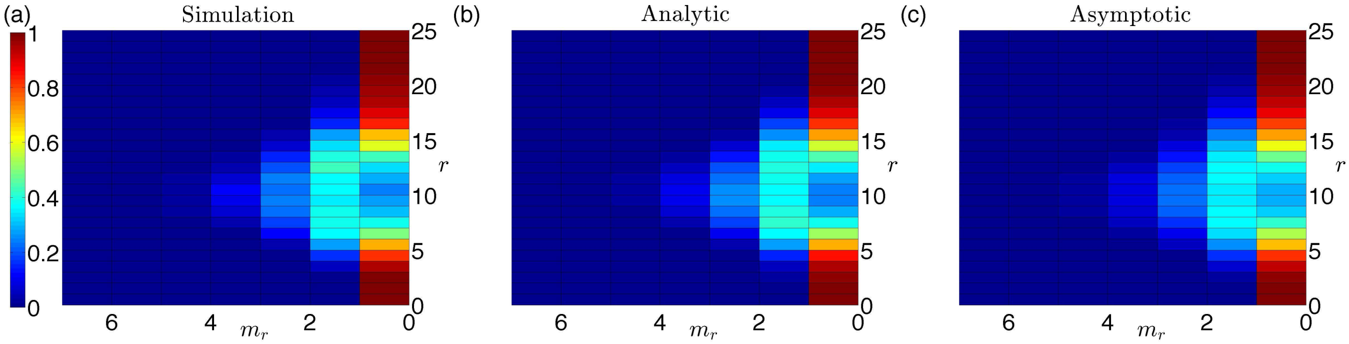


Figure 3: Heat maps of the probability distribution $\Pr(M_r = m_r|M, N)$ of finding m_r cells that have been infected by exactly r IUs given a total number of viruses $N = 100$ and $M = 10$ cells. (a) The statistical frequency of virus-cell counts after simulating IUs randomly distributing to the M cells, averaged over 1000 iterations. (b) The analytical result obtained from Eq. 2. (c) The asymptotic approximation with $M = 10$ and $\mu = \frac{N}{M} = 10$, using the expression in Eq. 6. There is close agreement between the simulated and analytical results. The relatively low values of M and N makes the asymptotic formula in Eq. 6 inappropriate for this parameter regime, explaining the discrepancy between the asymptotic result and the exact analytical result. However, it is noteworthy how qualitatively small that deviation is, which will continue to vanish as M and N increase in value.

Under the same large M, N limit and using Eq. 6, we show in Appendix 4

$$\Pr(M_0 = m_0, \cdots, M_N = m_N|M, N) \approx \prod_{r=0}^{N} \Pr(M_r = m_r|M, N), \tag{8}$$

which implies that as $M, N \to \infty$, the random variables $M_0, \cdots, M_N$ are independently distributed. In Section 2.2, we will apply our probability model of SMOI to the case of a repeatedly diluted solution of virus particles.

2.2 Serial Dilution

Low viral particle concentrations in assays are typically obtained via serial dilution processes in order to increase the sensitivity to individual viral infections (4, 24, 25). The initial viral stock containing N_0 particles is diluted by a fixed factor of D and the process is repeated $d_{\max}$ times. At each dilution number d, an assay can be performed to determine if the concentration of virus particles in the diluted solution is sufficient to generate a qualitative signal of infection, known as a "cytopathic effect" (CE). For example, the diluted stock can be administered *in vivo* to a model organism such as a mouse. The mouse's death would indicate that at least one lethal unit of the virus was present at that dilution level. Alternatively, an *in vitro* assay can be carried out to measure a signal that, for example, quantifies the exact number of plated cells that were successfully infected. To model these assays, we first define M^* as the number of host cells infected by at least one IU and that are capable of producing new viruses. In Appendix 4 we derive the discrete probability density function for finding $M^* = m$ infected cells at a given dilution number d and find

$$\Pr(M^* = m) = \binom{M}{m} \left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]^m \exp\left(-\frac{N_0}{QMD^d}\right)^{M-m}. \tag{9}$$

Eq. 9 shows that the number of infected cells M^* is binomially distributed with expected value

$$\mathrm{E}[M^*] = M \left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right], \tag{10}$$

and variance

$$\mathrm{Var}[M^*] = M \left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right] \exp\left(-\frac{N_0}{QMD^d}\right). \tag{11}$$

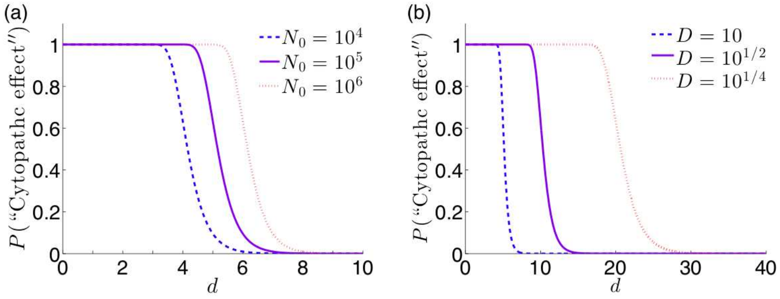


Figure 4: The probability of observing a cytopathic effect (CE) given in Eq. 12 as a function of the dilution number d and with $Q = 1$. (a) For $D = 10$, as the initial particle count N_0 increases, the critical dilution moves toward higher d. (b) Common dilution factors include logarithmic dilution ($D = 10$), half-logarithmic dilution ($D = 10^{1/2}$), and quarter-logarithmic dilution ($D = 10^{1/4}$). Logarithmic dilution requires a lower number of dilutions to cause the characteristic decrease in probability, requiring less individual assays to perform. Quarter-logarithmic dilution, though requiring more dilutions, has a slower transition from high to low probability across d, making the assay less sensitive to experimental error or noise. The plot above can be used to quantify the tradeoffs between the choices of D.

We can define the probability of observing a CE at dilution number d as the probability of finding one or more infected cells:

$$\begin{aligned}\text{Pr(“Cytopathic effect”)} &\equiv \sum_{m=1}^{M} \Pr(M^* = m) \\ &= 1 - \exp\left(-\frac{N_0}{QD^d}\right). \end{aligned} \tag{12}$$

The definition we use in Eq. 12 assumes an *in vitro* assay that can exhibit a cytopathic signal after a single cell infection or more. For *in vivo* assays, the probability that m infected cells are sufficient for a CE will depend on many complex physiological factors such as immune pressure, in-host viral evolution, and virion burst size (36). A plot of how the initial particle count N_0 and dilution factor D effect the characteristic functional form of Eq. 12 are shown in Fig. 4. Although both Eqs. 9 and 12 assume each IU contains all viral genes required for in-host replication, an extended probability model that factors in genetic mutation and degradation is provided in Appendix 4. Furthermore, for the case of retroviruses, infectious processes inside the host cytoplasm may be suppressed by previous infections, known as viral interference, and is explored in Appendix 4. In Section 3.1, we will use Eq. 9 to analyze the plaque assay. Eq. 12 will be used for "binary" assays that are only concerned with the presence or absence of a CE such as the endpoint dilution assay, which we will explore in Section 3.2.

3 Results and Discussion

3.1 Plaque Assay

The plaque assay is an example of a virus quantification assay (VQA) where the objective is to infer the total number of viruses N_0 present in a solution (24, 25, 37). After d serial dilutions the viral stock is added to a monolayer of M cells and a layer of agar gel is added to the well to inhibit the diffusion of virus particles in the plate. If a virus successfully infects a host cell, the agar will limit the range of new infections to the most adjacent cells. Viral infection thus spreads out radially from the initial nucleation infection and forms a visible discoloration in the plate called a "plaque." For high particle concentrations, the number of plaques formed may be large enough to cover the entire plate surface. After a sufficient critical dilution number d_c however, the number of plaques formed are low enough to be visibly distinct and countable. For each dilution number d, the assay can be performed for T number of trials. The 'signal' data arising from the plaque assay $P_{d,t}$ is the number of visible plaques counted, where $t = 1, \cdots, T$ is the trial number. The standard method of obtaining an estimate $\hat{N}_0$ of the particle count N_0 is to apply the sample mean of the data $P_{d_c,t}$ at the critical dilution level d_c to the formula

$$\hat{N}_0 = D^{d_c}\left(\frac{1}{T}\sum_{t=1}^{T} P_{d_c,t}\right), \tag{13}$$

which posits that the average number of plaques is directly proportional to the particle count N_0. Eq. 13 assumes that each infected cell corresponds to one IU, which is not necessarily true in the context of SMOI. Furthermore, although data corresponding to dilution numbers $d < d_c$ are unusable, data for $d > d_c$ corresponding to countable plaques are not used at all in Eq. 13.

In order to improve on Eq. 13 by using the entire set of plaque counts $P_{d,t}$ for our estimate of N_0, we propose a maximum likelihood estimation (MLE) scheme. Using the mathematical models derived above, we can construct an expression $\mathcal{L}(P_{d,t}|N_0)$ of the probability that the data observed $P_{d,t}$ can be generated assuming a particular value for N_0, known as a likelihood function. A value for N_0 that maximizes $\mathcal{L}(P_{d,t}|N_0)$, the MLE, corresponds to the most probable estimate $\hat{N}_0$ that could have generated the data. As each nucleation of a plaque corresponds to a distinct infected cell and under the assumption that overlapping lesions of dead cells are still discernible as distinct plaques, we can equate $P_{d,t}$ to the total number of successfully infected cells M^*. We will ignore the dynamics of coinfection and viral interference. Using Eq. 9, we propose the following likelihood function of the data given N_0:

$$\mathcal{L}(P_{d,t}|N_0) = \prod_{d=d_c}^{d_{\max}} \prod_{t=1}^{T} \binom{M}{P_{d,t}} \left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]^{P_{d,t}} \exp\left(-\frac{N_0}{QMD^d}\right)^{M-P_{d,t}}. \tag{14}$$

To obtain the MLE $\hat{N}_0$, we take the derivative of the natural log of Eq. 14 with respect to N_0 and set the result to zero to obtain

$$0 = \sum_{d=d_c}^{d_{\max}} \sum_{t=1}^{T} \frac{M \exp\left(\frac{-\hat{N}_0}{QMD^d}\right) - M + P_{d,t}}{QMD^d \left[1 - \exp\left(\frac{-\hat{N}_0}{QMD^d}\right)\right]}. \tag{15}$$

We can solve Eq. 15 using numerical methods such as Newton-Raphson (38), an iterative scheme that approaches the solution of an equation asymptotically starting from an initial guess $\hat{N}_0^{\text{init}}$. To increase the stability of convergence to the solution, we choose $\hat{N}_0^{\text{init}}$ by using the expectation in Eq. 10 to derive

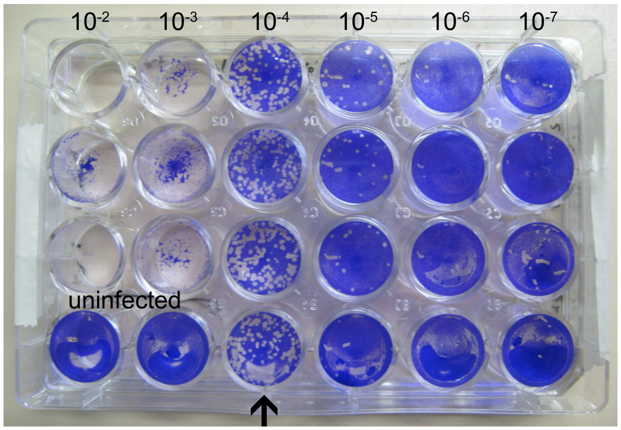
$$\hat{N}_0^{\text{init}} = -QMD^{d_c} \ln\left[1 - \frac{1}{M}\left(\frac{1}{T}\sum_{t=1}^{T} P_{d_c,t}\right)\right]. \tag{16}$$

An example of raw plaque count data and the resulting estimates for N_0 are given in Fig. 5. In order to quantify the relative improvement of the MLE of N_0 over the former method in Eq. 13, we simulate plaque assay data assuming a fixed, known N_0 value. In our simulation, we use the models established in Section 2.1 to sample the N_0 particles according to Eq. S9 to account for serial dilution and sample again the resulting particles according to Eq. 1 to obtain the number of IUs N. The IUs are distributed randomly to the M cells with equal probability and the resulting number of infected cells M^* are recorded. Since plates of cells with too many infections render the number of plaques uncountable, a “countable plaque threshold” renders the data unusable when the number of infected cells exceed the threshold. Thus, the resulting plaque data $P_{d,t}$ for a given dilution d and trial t is assigned the number of simulated infected cells if the latter is less than the given threshold. A scatter plot of the data $P_{d,t}$ of one such simulation is shown in Fig. 6a and the corresponding likelihood function from Eq. 14 is plotted in Fig. 6b. Because the MLE method utilizes a full probabilistic model of the plaque count distribution instead of relying only on the expected value at the single critical dilution d_c, it generates an estimate consistently closer to the *a priori* N_0.

3.2 Endpoint Dilution Assay

Another widely used assay for quantifying the initial viral particle count N_0 is the endpoint dilution or endpoint titration assay (23, 25, 39). It is often used in place of the plaque assay as it can be more rapidly performed and is useful for viral strains that are unable to form plaques. Here, serial dilutions at a factor of D are employed and at every dilution number d, an assay is performed T times to test for a successful CE. The number E_d of observed CEs at a given dilution number d is recorded as the signal. For low dilution, we expect many cells to be infected and the probability of observing a CE, as shown in Eq. 12, is close to 1. If every trial of the assay is likely to display a CE, then E_d is expected to be close to T. However, at high dilution, the probability in Eq. 12 rapidly decreases to 0, as shown in Fig. 4, and E_d will be similarly small. For a large initial stock of viral particles N_0, a larger dilution number d is needed to ensure the dramatic change in probability in Eq. 12. Thus, the critical dilution at which E_d most rapidly decreases from T can be used to estimate the particle count N_0.

One commonly used way to estimate N_0 is the Reed and Muench (RM) method (28). The RM method was developed to utilize the two dilution numbers that capture the greatest change in the data E_d. We first define a critical dilution number $d_{50\%}$



Plaque Count Data from Sloutskin et al (37).						
Trial	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
1	–	–	111	24	1	0
2	–	–	126	24	1	0
3	–	–	121	13	0	0

Figure 5: An example of raw plaque count data taken from Sloutskin et al (37). A viral solution was assayed in a plate of $M = 3 \times 10^5$ cells at dilution number $d = 2, 3, 4, 5, 6$, and 7 at a dilution factor of $D = 10$. The particle to PFU ratio is assumed to be $Q = 1$. For $T = 3$ separate trials, the number of plaques were counted at each dilution level. The bottom row of plates used as a control is ignored. For dilution numbers $d = 2$ and 3, the entire plate of cells show cytotoxicity so that the numbers of plaques were undiscernable and, thus, the countable data starts at $d_c = 4$. For the old method featured in Eq. 13, the estimate for N_0 was $\hat{N}_0 = 1.19 \times 10^6$ and for the MLE derived from Eq. 15, $\hat{N}_0 = 1.26 \times 10^6$. This results in a relative difference of 5.5%.

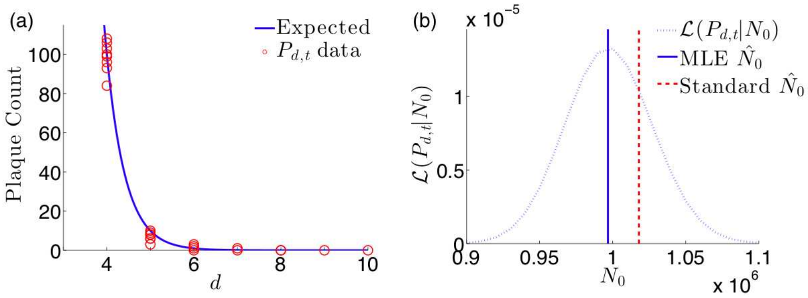


Figure 6: Results of plaque assay simulation for parameters $N_0 = 10^6$, $M = 10^5$, $Q = 1$, $D = 10$, $d_{\max} = 10$, and $T = 10$. (a) The scatter plot of simulated data $P_{d,t}$ (circles) and the expected value of plaque counts as given by Eq. 10 show close agreement. (b) The likelihood function $\mathcal{L}(P_{d,t}|N_0)$ with respect to N_0 using the same simulated data. The MLE obtained by iteratively solving Eq. 15 is $\hat{N}_0 = 9.97 \times 10^5$ and is relatively closer to the true value of N_0 than the estimate calculated from the standard method in Eq. 13 $\hat{N}_0 = 1.02 \times 10^6$.

to be the largest dilution such that at least 50% of the trials exhibit a CE. The estimate $\hat{N}_0$ for the particle count N_0 is given by

$$\log_{10}(\hat{N}_0) = d_{50\%} + \frac{E_{d_{50\%}} - 0.5T}{E_{d_{50\%}} - E_{d_{50\%}+1}}. \tag{17}$$

Another commonly used estimation scheme is the Spearman-Karber (SK) method (29, 39). This method similarly utilizes a heuristic understanding of how the dilution d effects the data E_d. We define the critical dilution number $d_{100\%}$ as the largest dilution such that 100% of trials exhibit a cytopathic effect. The estimate $\hat{N}_0$ is given by

$$\log_{10}(\hat{N}_0) = d_{100\%} - \frac{1}{2}\log_{10}(D) + \log_{10}(D)\sum_{d=d_{100\%}}^{d_{\max}} \frac{E_d}{T}. \tag{18}$$

Both methods are derived from the heuristic observation that E_d exhibits sigmoidal behavior as a function of the dilution number d, but an underlying probabilistic model is missing. Neither method uses the "particle to PFU ratio" Q, factors in the stochasticity of serial diluting viral samples, or considers the dynamics of SMOI. Furthermore, both methods fail to employ the entire set of data E_d.

We present an alternative way to infer N_0 using Eq. 12 to establish a maximum likelihood estimation scheme. We restrict ourselves to *in vitro* assays in which a single infected cell is sufficient to display a CE. Then each cytopathic count is binomially distributed with parameters T and the probability given in Eq. 12. Thus, for a set of data $\{E_1, E_2, \cdots, E_{d_{\max}}\}$, we propose the likelihood function

$$\mathcal{L}(E_d|N_0) = \prod_{d=1}^{d_{\max}} \binom{T}{E_d}\left[1-\exp\left(\frac{-N_0}{QD^d}\right)\right]^{E_d} \exp\left(\frac{-N_0}{QD^d}\right)^{T-E_d}. \tag{19}$$

Eq. 19 is an expression of the probability of the data $\{E_1, \cdots, E_{d_{\max}}\}$ given the current assumed value of N_0. To obtain the best estimate $\hat{N}_0$ of N_0, we maximize the likelihood function by taking the log and derivative of $\mathcal{L}(E_d|N_0)$ with respect to N_0 and set it equal to zero to obtain

$$0 = \sum_{d=1}^{d_{\max}} \frac{E_d - T + T\exp\left(\frac{-\hat{N}_0}{QD^d}\right)}{QD^d\left(1-\exp\left(\frac{-\hat{N}_0}{QD^d}\right)\right)}. \tag{20}$$

As with Eq. 15, solving Eq. 20 for $\hat{N}_0$ requires a numerical method such as Newton-Raphson. As an appropriate initial estimate for $\hat{N}_0$, the formula

$$\hat{N}_0^{\text{init}} = -0.5QD^{d_c}\left[\ln\left(1-\frac{E_{d_c}}{T}\right) + D\ln\left(1-\frac{E_{d_c+1}}{T}\right)\right], \tag{21}$$

can be used, where d_c is the largest dilution number such that at least half of the assays exhibit a cytopathic effect. Eq. 21 is the average of the N_0 estimates at dilutions d_c and d_{c+1} when setting the CE probability in Eq. 12 to $1/2$. For a comparison of our MLE method with the RM and SK methods, we simulate data similar to that described in Section 3.1. Here we take the number of trials such that the simulated count of infected cells is greater than zero as the values of E_d for a given dilution number d. We plot the likelihood from Eq. 19 and compare the MLE of N_0 with those derived by the RM and SK methods in Fig. 7a. As both RM and SK estimate very similar values of $\hat{N}_0$, they both consistently over-estimate the *a priori* set N_0 relative to the MLE method. This demonstrates the utility of a probabilistic model for parameter inference over heuristically determined formulas.

The expressions we derived in Eqs. 14 and 19 applied to simulated data can also help quantify tradeoffs in experimental design. As discussed above, there exist viruses that cannot form plaques, restricting the options of VQAs to endpoint dilution. However, for many cases, the choice between using one assay over the other can be one of convenience. More specifically, endpoint dilution assays can often be performed more rapidly than plaque assays. Using the same simulated data for both assays, we plot Eqs. 14 and 19 together in Fig. 7b. The plots clearly show the superiority of the plaque assay for estimating the viral stock number N_0 in respect to both how close the MLE infers the true N_0 value and the amount of variance in that estimate. While the amount of variability and error that is tolerable for an experiment may be context-dependent, the plots in Fig. 7b provide a quantitative way to differentiate between the two methods.

3.3 Luciferase Reporter Assay

The luciferase reporter assay, an example of an IA, is used to measure the infectivity of a viral strain. Here the ratio $\mu = N/M$ of total infections over the number of plated cells is estimated by measuring the transcription activity of viral proteins (14–16). The reporter employs an oxidative enzyme luciferase that facilitates a reaction when introduced to the substrate luciferin, resulting in bioluminescence. The protocol begins with attaching the luciferase encoding gene to the viral genome. The altered viral strain is cloned to a total particle count N_0 which, in this case, is assumed to be fixed and known. The solution of viruses is added to a plated monolayer of M host cells. An incubation time is allowed for transcription of viral proteins and, incidentally, the luciferase enzyme. Subsequently all cells are lysed to release all cytoplasm content into the solution upon which

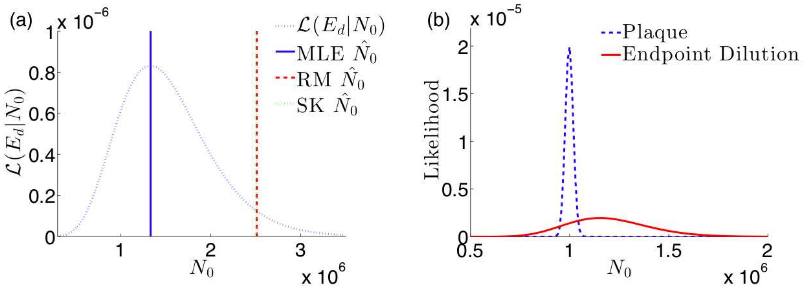


Figure 7: (a) The likelihood function $\mathcal{L}(E_d|N_0)$ in Eq. 19 for the endpoint dilution assay and the corresponding maximum likelihood, Reed and Muench, and Spearman-Karber estimates given simulated data generated with $N_0 = 10^6$, $Q = 1$, $D = 10$, and $d_{\max} = 10$. The estimates for maximum likelihood ($\hat{N}_0 = 1.33 \times 10^6$), RM ($\hat{N}_0 = 2.51 \times 10^6$), and SK ($\hat{N}_0 = 2.51 \times 10^6$) all overestimate N_0, but the smaller relative error of the MLE is an improvement on the errors of the existing two methods. (b) The likelihood functions $\mathcal{L}(P_{d,t}|N_0)$ and $\mathcal{L}(E_d|N_0)$ for the plaque and endpoint dilution assays respectively given simulated data. The data was generated with parameters $N_0 = 10^6$, $M = 10^5$, $Q = 1$, $D = 10^{1/4}$, $d_{\max} = 30$, and a "countable plaque threshold" of 150. The plaque assay likelihood is concentrated close to the true N_0 value while the endpoint dilution likelihood is far more spread out and overestimates N_0.

luciferin is added. The oxidation of luciferin is facilitated by the luciferase enzyme and the resulting bioluminescence yields a measurable signal (40). The light intensity is thus a measure of total transcription activity of the viral genome in the cell and can be used as a proxy for the total number of viruses N that successfully infected host cells. The experiment may be repeated for T number of trials.

Although there is stochasticity in transcription factor binding and, in the case of retroviruses, the number of integration sites on the host DNA, we will assume that each successful virus infection contributes one viral genome to be transcribed and each transcription occurs at a constant rate proportional to the total number of integrated viral genomes. Note that the limited number of transcription factors, ribosomes, and other cell machinery necessary to produce viral proteins and the luciferase reporter causes the production rate to saturate as the number of infecting viruses r per cell increases. Thus, transcription activity saturates with increasing number of infections r. We can model this effect by defining a monotonically increasing function $f(r)$ representing the number of transcribed viral proteins when a cell is infected by r viruses over the course of the assay. Thus, for a given SMOI $\{M_0, \cdots, M_N\}$, we will model the intensity signal L of the total luciferase reporter luminescence with

$$L = \sum_{r=0}^{N} L_0 f(r) M_r, \tag{22}$$

where L_0 is the fluorescence intensity arising from a single luciferase reporter present in the solution. Although $f(r)$ may take on many functional forms, a commonly used model for transcription factor kinetics is the Hill function (41) given by

$$f(r) = \frac{f_{\max} r^h}{K + r^h}, \tag{23}$$

where $f_{\max}$ is the maximum transcription activity of luciferase, h is the Hill coefficient that quantifies cooperative binding of multiple transcription factors at a promoter region, and K is an effective disassociation constant relating the binding and unbinding rates of transcription factor. The functional form of Eq. 23 accounts for the limited transcription machinery available for the multiple copies of viral genome present in the cell. In Fig. 8a we calculate the discrete probability distribution $\Pr(L = l)$ by considering the cumulative weight of every allowable configuration of N viruses infecting M cells through Eq. 22.

The luciferase reporter assay maintains large values of initial virus count N_0 and cell count M. We can thus use the asymptotic approximations in Eqs. 6 and 7 with the Central Limit Theorem (35) to assume L is normally distributed with expected value

$$\mathrm{E}[L] = L_0 f_{\max} M e^{-\mu} \sum_{r=0}^{N} \frac{r^h \mu^r}{(K + r^h) r!}, \tag{24}$$

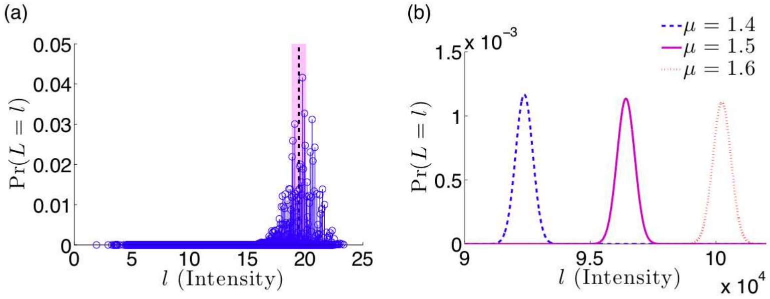


Figure 8: Probability distributions of the luciferase assay fluorescence intensity L from Eq. 22. (a) A toy example of a discrete probability distribution of each allowable configuration for $N = 30$ viruses infecting $M = 20$ using Eq. 5. The parameters used for the reporter kinetics are $f_{\max} = 2$, $h = 1$, $K = 1$ and $L_0 = 1$. The mean intensity of the signal is $\mathrm{E}[L] = 19.5$, represented by the vertical dotted line, and variance $\mathrm{Var}[L] = 1.49$, represented by the shaded region. (b) The normally distributed approximation for $M = 1 \times 10^5$, $f_{\max} = 2$, $h = 1$, $K = 1$ and $L_0 = 1$. The distributions are plotted for $\mu = 1.4$, 1.5, and 1.6 by computing the expected values $\mathrm{E}[L] = 9.23 \times 10^4$, 9.64×10^4 and 1×10^5 and the variances $\mathrm{Var}[L] = 1.7 \times 10^5$, 1.24×10^5 and 1.3×10^5 respectively.

and variance

$$\mathrm{Var}[L] = L_0^2 f_{\max}^2 M e^{-\mu} \sum_{r=0}^{N} \frac{r^{2h}\mu^r}{(K + r^h)^2 r!}. \tag{25}$$

A visualization of the normal approximation of the probability distribution of L is shown in Fig. 8b. Furthermore, with Eqs. 24 and 25, we can derive the likelihood function $\mathcal{L}(L_t^{\mathrm{data}}|\mu)$ of the data L_t^{data}, given μ

$$\mathcal{L}(L_t^{\mathrm{data}}|\mu) = \prod_{t=0}^{T} \frac{1}{\sqrt{2\pi \mathrm{Var}[L]}} \exp\left[-\frac{\left(L_t^{\mathrm{data}} - \mathrm{E}[L]\right)^2}{2\mathrm{Var}[L]}\right], \tag{26}$$

where t is the trial number. Due to the complicated functional form of the mean and variance of L, creating a maximum likelihood scheme to estimate μ from experimental data is intractable, so we use Eq. 24 by replacing the expected value with the experimental average of measurements L_t^{data}. If we assume no cooperative transcription binding ($h = 1$), we solve for the estimate $\hat{\mu}$ by applying the Newton-Raphson iterative method to the equation

$$0 = \frac{1}{T} \sum_{t=0}^{T} L_t^{\mathrm{data}} - L_0 f_{\max} M e^{-\hat{\mu}} \sum_{r=0}^{N_0} \frac{r\hat{\mu}^r}{(K + r)r!}, \tag{27}$$

The typical method, under the assumption that luminescent intensity is proportional to the number of IUs N, is to use the sample mean via the formula $\hat{\mu}^{\mathrm{init}} = \frac{1}{L_0 M T} \sum_{t=0}^{T} L_t^{\mathrm{data}}$. This fails to account for the effects of SMOI. However, we employ it as an initial guess for our iterative method to solve Eq. 27. In order to compare the two estimates, we simulate data similar the descriptions in the previous two sections. Here, we do not dilute the initial particle count and, after distributing the N IUs to the M cells with equal probability, we compile the SMOI configuration and calculate L_t^{data} using Eq. 22. The results are shown in Fig. 9. The iterative method produces an estimate $\hat{\mu}$ far closer to the true value of μ than the former method. A similar approach can be used to compare methods for alternative functional forms of the viral protein transcription dynamics described in Eq. 23.

4 Conclusion

In this work, we derived probability models that quantify the the viral infectivity of host cells in an *in vitro* environment. By factoring in the stochastic nature of virus-host engagement, defective and/or abortive events, and the possibility of multiple infections of a single host, we defined the statistical multiplicity of infection (SMOI) and determined related probabilistic models. We analyzed two limiting regimes: one of small, dilute numbers of infecting viruses N, and the other of large N. For the low N regime, Eqs. 2 and 5 model how the limited number of infectious units are distributed amongst the M host cells. Alternatively, for large N, we showed the cell counts of the SMOI become statistically independent, as displayed in Eq. 8,

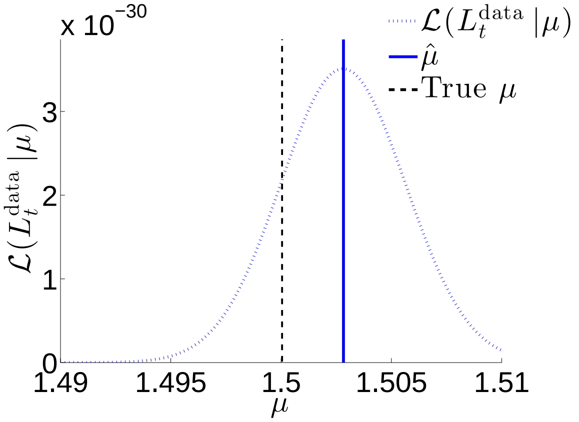


Figure 9: The likelihood function $\mathcal{L}(L_t^{\text{data}}|\mu)$ using Eq. 26 and simulated data. We set $\mu = 1.5$ and assign other parameters with $M = 1 \times 10^5$, $f_{\max} = 2$, $h = 1$, $K = 1$ and $L_0 = 1$. The estimate derived from solving Eq. 27 is $\hat{\mu} = 1.502$ while the standard method based on the sample mean yields $\hat{\mu} = 0.97$, far lower than what is displayed in the plot.

Comparison of Virological Assay Analyses		
Assay (Parameter)	New Method	Standard Method
Plaque (N_0)	$0 = \sum_{d=d_c}^{d_{\max}} \sum_{t=1}^{T} \frac{M \exp\left(\frac{-\hat{N}_0}{QMD^d}\right) - M + P_{d,t}}{QMD^d \left[1 - \exp\left(\frac{-\hat{N}_0}{QMD^d}\right)\right]}$ Initial guess: $\hat{N}_0^{\text{init}} = -QMD^d \ln\left(1 - \frac{1}{MT}\sum_{t=1}^{T} P_{d_c,t}\right)$	$\hat{N}_0 = D^{d_c}\left(\frac{1}{T}\sum_{t=1}^{T} P_{d_c,t}\right)$
Endpoint Dilution (N_0)	$0 = \sum_{d=1}^{d_{\max}} \frac{E_d - T + T\exp\left(\frac{-\hat{N}_0}{QD^d}\right)}{QD^d\left(1 - \exp\left(\frac{-\hat{N}_0}{QD^d}\right)\right)}$ Initial guess: $\hat{N}_0^{\text{init}} = \frac{-QD^{d_c}}{2} \ln\left[\left(1 - \frac{E_{d_c}}{T}\right)\left(1 - \frac{E_{d_c+1}}{T}\right)^D\right]$	Reed and Muench: $\log_{10}(\hat{N}_0) = d_{50\%} + \frac{E_{d_{50\%}} - 0.5T}{E_{d_{50\%}} - E_{d_{50\%}+1}}$ Spearman-Karber: $\log_{10}(\hat{N}_0) = d_{100\%} - \left[\frac{1}{2} - \sum_{d=d_{100\%}}^{d_{\max}} \frac{E_d}{T}\right] \log_{10} D$
Luciferase Reporter $\left(\mu = \frac{N}{M}\right)$	$0 = \frac{1}{T}\sum_{t=0}^{T} L_t^{\text{data}} - L_0 f_{\max} M e^{-\hat{\mu}} \sum_{r=0}^{N_0} \frac{r\hat{\mu}^r}{(K+r)r!}$ Initial guess: $\hat{\mu}^{\text{init}} = \frac{1}{L_0MT}\sum_{t=0}^{T} L_t^{\text{data}}$	$\hat{\mu} = \frac{1}{L_0MT}\sum_{t=0}^{T} L_t^{\text{data}}$

Table 1: A summary of the analytically derived expressions used to analyze experimental results. For virus quantification assays, the desired parameter to estimate is the number of initial viral particles N_0 for the plaque and endpoint dilution assays. For infectivity assays, the ratio $\mu = N/M$ is the objective of the luciferase reporter assay. Each expression is displayed next to the most common method used to estimate these parameters currently.

and that they display a Poisson distribution in Eq. 6. Lastly, we explored the effects of serial dilution on the total number of infected cells and the probability of observing an infectious signal in Eq. 9.

Using our probability models along with reasonable assumptions of applied combinatorics and nonlinear inference, we analytically derived expressions for several virological assays to improve on existing methods of experimental data analysis. For virus quantification assays, serial dilution of virus particles creates an environment akin to the low particle regime. Using the appropriate probability model, we created new methods of estimating the particle count N_0 in the initial viral stock for the plaque assay and the endpoint dilution assay. For measuring infectivity of a viral strain, the objective is to determine the effective multiplicity of infection $\mu = N/M$ as the ratio of successfully infecting viruses N and the total number of cells M included in the assay. As these assays operate under no dilution, we employed the large N limit probability model to analytically derive expressions for the luciferase reporter assay to estimate μ. A summary of each estimation method along with the most commonly used counterpart is displayed in Table 1.

Further work can be done to increase the accuracy of our probabilistic models and expressions derived for the virological assays. A thorough result of cell count distributions conditioned on the inhomogeneity in cell sizes is still needed to better quantify the variance in our results. A mechanistic model to quantify the "particle to PFU ratio" for a given virus and experimental protocol would greatly increase the accuracy and confidence in virus quantification as it primarily determines the relationship between the physical viral stock's particle count N_0 and the theoretical number of infectious units N used in our derivations. Finally, the aggregation of viral particles into clusters acting as a single infectious unit is a significant source of multiplicity of infection and coinfection. A model of aggregate nucleation and cluster size distributions can be included into our model to account for these confounding factors.

Author Contributions

BM derived mathematical formulae, developed statistical inference framework, performed simulations, generated plots, and wrote the initial draft. MRD and TC verified the mathematical results, contributed to their analyses, and edited the manuscript. TC conceptualized, designed, and supervised the research.

Acknowledgments

This work was supported in part by grants from the NSF (DMS-1516675) and the Army Research Office (W911NF-14-1-0472). We are especially grateful to Dr. Nicholas Webb and Prof. Benhur Lee for insightful discussions.

References

1. Pegu, A., Z.-Y. Yang, J. C. Boyington, L. Wu, S.-Y. Ko, S. D. Schmidt, K. McKee, W.-P. Kong, W. Shi, X. Chen, J.-P. Todd, N. L. Letvin, J. Huang, M. C. Nason, J. A. Hoxie, P. D. Kwong, M. Connors, S. S. Rao, J. R. Mascola, and G. J. Nabel, 2014. Neutralizing Antibodies to HIV-1 Envelope Protect More Effectively In Vivo Than Those to the CD4 Receptor. Sci. Transl. Med. 6:88.
2. Osbourn, J. K., J. C. Earnshaw, K. S. Johnson, M. Parmentier, V. Timmermans, and J. McCafferty, 1998. Directed Selection of MIP-1α Neutralizing CCR5 Antibodies from a Phage Display Human Antibody Library. Nat. Biotechnol. 16:778–781.
3. Qiu, S., H. Yi, J. Hu, Z. Cao, Y. Wu, and W. Li, 2012. The Binding Mode of Fusion Inhibitor T20 onto HIV-1 gp41 and Relevant T20-Resistant Mechanisms Explored by Computational Study. Curr. HIV Res. 182–194.
4. Platt, E. J., M. M. Gomes, and D. Kabat, 2014. Reversible and Efficient Activation of HIV-1 Cell Entry by a Tyrosine-Sulfated Peptide Dissects Endocytic Entry and Inhibitor Mechanisms. J. Virol. 88:4304–4318.
5. Platt, E. J., J. P. Durnin, and D. Kabat, 2005. Kinetic Factors Control Efficiencies of Cell Entry, Efficacies of Entry Inhibitors, and Mechanisms of Adaptation of Human Immunodeficiency Virus. J. Virol. 79:4347–4356.
6. Fatkenheuer, G., A. L. Pozniak, M. A. Johnson, A. Plettenberg, S. Staszewski, A. I. M. Hoepelman, M. S. Saag, F. D. Goebel, J. K. Rockstroh, B. J. Dezube, T. M. Jenkins, C. Medhurst, J. F. Sullivan, C. Ridgway, S. Abel, I. T. James, M. Youle, and E. van der Ryst, 2005. Efficacy of Short-Term Monotherapy with Maraviroc, a New CCR5 Antagonist, in Patients Infected with HIV-1. Nat. Med. 11:1170–1172.
7. Jonckheere, H., J. Anné, and E. De Clercq, 2000. The HIV-1 Reverse Transcription (RT) Process as Target for RT Inhibitors. Med. Res. Rev. 20:129–154.
8. Thierry, S., S. Munir, E. Thierry, F. Subra, H. Leh, A. Zamborlini, D. Saenz, D. N. Levy, P. Lesbats, and A. Saib, 2015. Integrase Inhibitor Reversal Dynamics Indicate Unintegrated HIV-1 DNA Initiate De Novo Integration. Retrovirology 12:24.
9. Paterson, D. L., S. Swindells, J. Mohr, M. Brester, E. N. Vergis, C. Squier, M. M. Wagener, and N. Singh, 2000. Adherence to Protease Inhibitor Therapy and Outcomes in Patients with HIV Infection. Ann. Intern. Med. 133:21–30.
10. Wilen, C. B., J. C. Tilton, and R. W. Doms, 2012. HIV: Cell Binding and Entry. Cold Spring Harb. Persp. Med. 2.
11. Qian, K., S. L. Morris-Natschke, and K.-H. Lee, 2009. HIV Entry Inhibitors and Their Potential in HIV Therapy. Med. Res. Rev. 29:369–393.
12. Boulant, S., M. Stanifer, and P.-Y. Lozach, 2015. Dynamics of Virus-Receptor Interactions in Virus Binding, Signaling, and Endocytosis. Viruses 7:2794–2815.
13. Chou, T., 2007. Stochastic Entry of Enveloped Viruses: Fusion versus Endocytosis. Biophys. J. 93:1116–1123.
14. Chikere, K., T. Chou, P. R. Gorry, and B. Lee, 2013. Affinofile Profiling: How Efficiency of CD4/CCR5 Usage Impacts the Biological and Pathogenic Phenotype of HIV. J. Virol. 435:81–91.

15. Johnston, S. H., M. A. Lobritz, S. Nguyen, K. Lassen, S. Delair, F. Posta, Y. J. Bryson, E. J. Arts, T. Chou, and B. Lee, 2009. A Quantitative Affinity-Profiling System That Reveals Distinct CD4/CCR5 Usage Patterns among Human Immunodeficiency Virus Type 1 and Simian Immunodeficiency Virus Strains. J. Virol. 83:11016–11026.
16. Webb, N. E., and B. Lee, 2016. Quantifying CD4/CCR5 Usage Efficiency of HIV-1 Env Using the Affinofile System. In HIV Protocols, Springer New York, volume 1354 of Methods in Molecular Biology, 3–20.
17. Killian, M. L., 2008. Hemagglutination Assay for the Avian Influenza Virus, Humana Press, Totowa, NJ, 47–52.
18. Mascola, J. R., M. K. Louder, C. Winter, R. Prabhakara, S. C. D. Rosa, D. C. Douek, B. J. Hill, D. Gabuzda, and M. Roederer, 2002. Human Immunodeficiency Virus Type 1 Neutralization Measured by Flow Cytometric Quantitation of Single-Round Infection of Primary Human T Cells. J. Virol. 76:4810–4821.
19. Brown, C. M., and K. D. Bidle, 2014. Attenuation of Virus Production at High Multiplicities of Infection in Aureococcus Anophagefferens. J. Virol. 466–467:71–81.
20. Sette, A., and J. Fikes, 2003. Epitope-Based Vaccines: An Update on Epitope Identification, Vaccine Design and Delivery. Curr. Opin. Immunol. 15:461–470.
21. Gerdil, C., 2003. The Annual Production Cycle for Influenza Vaccine. Vaccine 21:1776–1779.
22. Tree, J. A., C. Richardson, A. R. Fooks, J. C. Clegg, and D. Looby, 2001. Comparison of Large-Scale Mammalian Cell Culture Systems with Egg Culture for the Production of Influenza Virus A Vaccine Strains. Vaccine 19:25–26.
23. Neumann, G., K. Fujii, Y. Kino, and Y. Kawaoka, 2005. An Improved Reverse Genetics System for Influenza A Virus Generation and its Implications for Vaccine Production. Proc. Natl. Acad. Sci. U. S. A. 102:16825–16829.
24. Kropinski, A. M., A. Mazzocco, T. E. Waddell, E. Lingohr, and R. P. Johnson, 2009. Enumeration of Bacteriophages by Double Agar Overlay Plaque Assay, Humana Press, 69–76.
25. Johnson, V. A., R. E. Byington, and P. L. Nara, 1990. Quantitative Assays for Virus Infectivity, Palgrave Macmillan UK, 71–86.
26. Agrawal-Gamse, C., F.-H. Lee, B. Haggerty, A. P. O. Jordan, Y. Yi, B. Lee, R. G. Collman, J. A. Hoxie, R. W. Doms, and M. M. Laakso, 2009. Adaptive Mutations in Human Immunodeficiency Virus Type 1 Envelope Protein with a Truncated V3 Loop Restore Function by Improving Interactions with CD4. J. Virol. 83:11005–11015.
27. Mascola, J. R., M. K. Louder, C. Winter, R. Prabhakara, S. C. De Rosa, D. C. Douek, B. J. Hill, D. Gabuzda, and M. Roederer, 2002. Human Immunodeficiency Virus Type 1 Neutralization Measured by Flow Cytometric Quantitation of Single-Round Infection of Primary Human T Cells. J. Virol. 76:4810–4821.
28. Reed, L. J., and H. Muench, 1938. A Simple Method of Estimating Fifty Percent Endpoints. Am. J. Hygiene 27:493–497.
29. Hamilton, M. A., R. C. Russo, and R. V. Thurston, 1977. Trimmed Spearman-Karber Method for Estimating Median Lethal Concentrations in Toxicity Bioassays. Eviron. Sci. Technol. 11:714–719.
30. Schwerdt, C. E., and J. Fogh, 1957. The Ratio of Physical Particles per Infectious Unit Observed for Poliomyelitis Viruses. Virology 4:41–52.
31. Aguilera, E. R., A. K. Erickson, P. R. Jesudhasan, C. M. Robinson, and J. K. Pfeiffer, 2017. Plaques Formed by Mutagenized Viral Populations Have Elevated Coinfection Frequencies. Am. Soc. Microbiol. 8:1–12.
32. Klasse, P. J., 2015. Molecular Determinants of the Ratio of Inert to Infectious Virus Particles. Prog. Mol. Biol. Transl. Sci. 129:285–326.
33. Layne, S. P., M. J. Merges, M. Dembo, J. L. Spouge, S. R. Conley, J. P. Moore, J. L. Raina, H. Renz, H. R. Gelderblom, and P. L. Nara, 1992. Factors Underlying Spontaneous Inactivation and Susceptibility to Neutralization of Human Immunodeficiency Virus. Virology 189:695–714.
34. Turner, T. E., S. Schnell, and K. Burrage, 2004. Stochastic Approaches for Modeling In Vivo Reactions. Comp. Biol. and Chem. 28:165–178.
35. Lange, K., 2003. Applied Probability. Springer Sci. Bus. Med. Inc.
36. Gilchrist, M. A., D. Coombs, and A. S. Perelson, 2004. Optimizing Within-Host Viral Fitness: Infected Cell Lifespan and Virion Production Rate. J. Theor. Biol. 229:281–288.
37. Sloutskin, A., and R. S. Goldsten, 2014. Infectious Focus Assays and Multiplicity of Infection (MOI) Calculations for Alpha-Herpesviruses. Bio-Protocol 4:e1295.
38. Lange, K., 2013. Optimization. Springer Sci. Bus. Med. Inc.
39. Ramakrishnan, M. A., 2016. Determination of 50% endpoint titer using a simple formula. World J. Virol. 5:85–86.
40. Montefiori, D. C., 2009. Measuring HIV Neutralization in a Luciferase Reporter Gene Assay, Humana Press, 395–405.
41. Weiss, J. N., 1997. The Hill Equation Revisited: Uses and Misuses. FASEB J. 11:835–41.
42. Pineda, E., P. Bruna, and D. Crespo, 2004. Cell Size Distribution in Random Tessellations of Space. Phys. Rev. E 70:066119.
43. Stern, A., S. Bianco, M. T. Yeh, C. Wright, K. Butcher, C. Tang, R. Nielsen, and R. Andino, 2014. Costs and Benefits of Mutational Robustness in RNA Viruses. Cell Reports 8:1026–1036.
44. Nisole, S., and A. Saïb, 2004. Early Steps of Retrovirus Replicative Cycle. Retrovirol. 1:9.
45. Nethe, M., B. Berkhout, and A. C. van der Kuyl, 2005. Retroviral Superinfection Resistance. Retrovirol. 2:52.

Supplementary Information: Mathematical Appendices

SMOI Probability

To derive Eq. 2, we index all cells with $i \in \{1, \cdots, M\}$ and define A_i^r as the event that cell i is infected by exactly r IUs. Then, given N IUs across all M cells, the probability of A_i^r is given by

$$\Pr(A_i^r | M, N) = \binom{N}{r} \left(\frac{1}{M}\right)^r \left(1 - \frac{1}{M}\right)^{N-r}. \tag{S1}$$

Since cell sizes are assumed to be homogeneous, the probability in Eq. S1 is the same for all cells, but the events $\{A_1^r, \cdots, A_M^r\}$ are not independent as the number of IUs N shared among the M cells is finite. Thus, we use the inclusion-exclusion principle (35) to derive

$$\begin{aligned}
\Pr(M_r = m_r | M, N) &= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \sum_{\substack{I \subset \{1, \cdots, M\} \\ |I| = j}} \Pr\left(\bigcap_{i \in I} A_i^r\right) \\
&= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \binom{M}{j} \Pr\left(\bigcap_{i=1}^{j} A_i^r\right) \\
&= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \binom{M}{j} \binom{N}{r, \cdots, r, (N - rj)} \left[\prod_{i=1}^{j} \left(\frac{1}{M}\right)^r\right] \left(\frac{M-j}{M}\right)^{N-rj} \\
&= \sum_{j=m_r}^{M} \binom{j}{m_r} \binom{M}{j} \binom{N}{r, \cdots, r, (N - rj)} \frac{(-1)^{j-m_r} (M-j)^{N-rj}}{M^N}.
\end{aligned} \tag{S2}$$

Note that the inner summation in the first identity above is over every possible collection of cells of size j, but as each cell is identical, the sum can be reduced to a single joint probability with the binomial degeneracy $\binom{M}{j}$.

Expected Value and Variance

For the generalized c-th moment $\mathrm{E}\left[M_r^c\right]$ of the number of cells M_r infected by exactly r viruses, we start with Eq. 2 to obtain

$$\begin{aligned}
\mathrm{E}\left[M_r^c\right] &= \sum_{m_r=0}^{M} \sum_{j=m_r}^{M} m_r^c (-1)^{j-m_r} \binom{j}{m_r} \binom{M}{j} \left(\frac{N!}{(r!)^j (N - rj)!}\right) \frac{(M-j)^{N-rj}}{M^N} \\
&= \sum_{j=0}^{M} \left[\sum_{m_r=0}^{j} m_r^c (-1)^{j-m_r} \binom{j}{m_r}\right] \binom{M}{j} \left(\frac{N!}{(r!)^j (N - rj)!}\right) \frac{(M-j)^{N-rj}}{M^N}
\end{aligned} \tag{S3}$$

To aid our derivation, we define the function $u(j, c)$ as

$$\begin{aligned}
u(j, c) &= \sum_{m=0}^{j} m^c (-1)^{j-m} \binom{j}{m} \\
&= j \sum_{k=0}^{j-1} (k+1)^{c-1} (-1)^{j-1-k} \binom{j-1}{k} \\
&= j \sum_{i=0}^{c-1} \binom{c-1}{i} \sum_{k=0}^{j-1} k^i (-1)^{j-1-k} \binom{j-1}{k} \\
&= j \sum_{i=0}^{c-1} \binom{c-1}{i} u(j-1, i).
\end{aligned} \tag{S4}$$

This is a recursive relationship from which we can evaluate any $u(j, c)$ using all $u(j-1, i)$ such that $0 \leq i < c$. We evaluate the first three cases $u(j, 0) = \delta_{0,j}$, $u(j, 1) = \delta_{1,j}$, and $u(j, 2) = \delta_{1,j} + 2\delta_{2,j}$, where $\delta_{0,j}$ is the Kronecker delta operator that returns the value 1 when the two subscript arguments are equal and 0 otherwise. We use the result for $c = 1$ and Eq. S3 to calculate the expected value of M_r as

$$\begin{aligned}
\mathrm{E}\left[M_r\right] &= \sum_{j=0}^{M} \delta_{1,j} \binom{M}{j} \left(\frac{N!}{(r!)^j \, (N-rj)!} \right) \frac{(M-j)^{N-rj}}{M^N} \\
&= M \binom{N}{r} \left(\frac{1}{M} \right)^r \left(1 - \frac{1}{M} \right)^{N-r} . \qquad \text{(S5)}
\end{aligned}$$

We obtain the second moment $\mathrm{E}\left[M_r^2\right]$ using the same method in order to obtain the variance of M_r as

$$\begin{aligned}
\mathrm{Var}\left[M_r\right] &= \mathrm{E}\left[M_r^2\right] - \mathrm{E}\left[M_r\right]^2 \\
&= M \binom{N}{r} \left(\frac{1}{M} \right)^r \left(1 - \frac{1}{M} \right)^{N-r} + \frac{M(M-1)N!(M-2)^{N-2r}}{(r!)^2 (N-2r)! M^N} - \frac{M^2 (N!)^2 (M-1)^{2N-2r}}{(r!)^2 \left[(N-r)!\right]^2 M^{2N}} . \qquad \text{(S6)}
\end{aligned}$$

Asymptotic Approximation

For the derivation of Eq. 6, we take the mathematical limit $N, M \to \infty$ while keeping the ratio $\mu = \frac{N}{M}$ fixed and approximate Eq. 2 as follows:

$$\begin{aligned}
\Pr(M_r = m_r | M, N) &= \sum_{j=m_r}^{M} \frac{j! M! N! (-1)^{j-m_r} (M-j)^{N-rj}}{m_r! (j-m_r)! j! (M-j)! (N-rj)! (r!)^j M^{N-rj} M^{rj}} \\
&= \frac{1}{m_r!} \sum_{j=m_r}^{M} \frac{(-1)^{j-m_r}}{(j-m_r)!(r!)^j} \left[M \cdots (M-j+1) \right] \frac{\left[N \cdots (N-rj+1) \right]}{M^{rj}} \left(1 - \frac{j}{M} \right)^{N-rj} \\
&= \frac{1}{m_r!} \sum_{j=m_r}^{M} \frac{(-1)^{j-m_r}}{(j-m_r)!(r!)^j} \left[M \cdots (M-j+1) \right] \left[\mu \cdots \left(\mu - \frac{rj-1}{M} \right) \right] \left(1 - \frac{\mu j}{N} \right)^{N-rj} \\
&\approx \frac{1}{m_r!} \sum_{j=m_r}^{M} \frac{(-1)^{j-m_r}}{(j-m_r)!(r!)^j} M^j \mu^{jr} e^{-\mu j} \\
&= \frac{1}{m_r!} \left[\frac{M \mu^r e^{-\mu}}{r!} \right]^{m_r} \sum_{j=0}^{M-m_r} \frac{(-1)^j}{j!} \left[\frac{M \mu^r e^{-\mu}}{r!} \right]^j \\
&\approx \frac{1}{m_r!} \left[\frac{M \mu^r e^{-\mu}}{r!} \right]^{m_r} \exp\left[- \frac{M \mu^r e^{-\mu}}{r!} \right] . \qquad \text{(S7)}
\end{aligned}$$

Note that, although the first approximation in Eq. 6 requires j in the summation to be sufficiently smaller than M, any contribution from the summation for j close to M vanishes due to both the $(j - m_r)!$ term in the denominator and the $\left(1 - \frac{j}{M}\right)^{N-rj}$ term approaching 0. Under the same large M, N limit, we can derive an asymptotic approximation of the joint probability distribution by taking the natural log of both sides of Eq. 5:

$$\begin{aligned}
\ln \Pr(M_0 = m_0, \cdots, M_N = m_N) &= \ln\left(\frac{1}{M^N}\right) + \ln M! + \ln N! + \sum_{r=0}^{N} \ln\left(\frac{1}{m_r!(r!)^{m_r}}\right) \\
&\approx -N\ln M + M\ln(M) - M + N\ln(N) - N + \sum_{r=0}^{N} \ln\left(\frac{1}{m_r!(r!)^{m_r}}\right) \\
&= N\ln\left(\frac{N}{M}\right) + M\ln M - Me^{-\mu}e^{\mu} - \mu M + \sum_{r=0}^{N} \ln\left(\frac{1}{m_r!(r!)^{m_r}}\right) \\
&= \ln\mu\left(\sum_{r=0}^{N} r m_r\right) + (\ln M - \mu)\left(\sum_{r=0}^{N} m_r\right) - Me^{-\mu}\left(\sum_{r=0}^{\infty}\frac{\mu^r}{r!}\right) + \sum_{r=0}^{N} \ln\left(\frac{1}{m_r!(r!)^{m_r}}\right) \\
&= \sum_{r=0}^{N}\left[r m_r \ln\mu + m_r \ln M - m_r\mu - \frac{Me^{-\mu}\mu^r}{r!} + \ln\left(\frac{1}{m_r!(r!)^{m_r}}\right)\right] - \sum_{r=N+1}^{\infty}\frac{Me^{-\mu}\mu^r}{r!} \\
&= \sum_{r=0}^{N} \ln\left[\frac{\mu^{r m_r} M^{m_r} e^{-m_r\mu}}{m_r!(r!)^{m_r}} \exp\left(-\frac{Me^{-\mu}\mu^r}{r!}\right)\right] - \mathcal{O}\left(\frac{M\mu^N}{N!}\right) \\
&\approx \ln\left[\prod_{r=0}^{N}\frac{1}{m_r!}\left[\frac{M\mu^r e^{-\mu}}{r!}\right]^{m_r} \exp\left(-\frac{M\mu^r e^{-\mu}}{r!}\right)\right]. \qquad \text{(S8)}
\end{aligned}$$

Since the argument in the right-hand-side of the last approximation is the same as Eq. 6, we arrive at the result in Eq. 8.

Number of Infected Cells

To derive Eq. 9, we first define N_d as the number of virus particles present in the viral solution after dilution of a factor of D^d. Obtaining N_d is effectively analogous to taking a volume of the initial viral stock scaled by D^{-d} and counting the number of particles captured in the volume. Thus, we expect N_d to be Poisson-distributed with mean $N_0 D^{-d}$ and discrete probability density function given by

$$\Pr\left(N_d = n_d | N_0\right) = \frac{1}{n_d!}\left(\frac{N_0}{D^d}\right)^{n_d} \exp\left(-\frac{N_0}{D^d}\right). \qquad \text{(S9)}$$

Once N_d is chosen from the above distribution, for a given "particle to PFU ratio" Q, the number of IUs N follows a binomial distribution with a probability function similar to Eq. 1, but with N_0 replaced with N_d. Note that, given an SMOI $\{M_0, \cdots, M_N\}$, it is immediate that $M^* = M - M_0$. Using this modified density of N and Eqs. 2 and S9, we can derive the discrete probability density function of M^* at a given dilution number d as

$$\begin{aligned}
\Pr\left(M^* = m\right) &= \sum_{n_d=0}^{N_0}\sum_{n=0}^{n_d} \Pr(N = n | N_d = n_d)\Pr(M_0 = M - m | N = n)\Pr(N_d = n_d) \\
&= \sum_{j=M-m}^{M} (-1)^{j-M+m}\binom{j}{M-m}\binom{M}{j} e^{-\frac{N_0}{D^d}} \sum_{n_d=0}^{N_0} \frac{\left(\frac{N_0}{D^d}\right)^{n_d}}{n_d!}\left[1 - Q^{-1} + Q^{-1}\left(1 - \frac{j}{M}\right)\right]^{n_d} \\
&\approx \sum_{j=M-m}^{M} (-1)^{j-M+m}\binom{j}{M-m}\binom{M}{j} \exp\left[\frac{N_0}{D^d}\left(1 - \frac{j}{QM}\right) - \frac{N_0}{D^d}\right] \\
&= \binom{M}{m}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]^{m} \exp\left(-\frac{N_0}{QMD^d}\right)^{M-m}. \qquad \text{(S10)}
\end{aligned}$$

Note that the approximation that closes the exponential term in the final result employs the assumption that N_0 is sufficiently large.

Inhomogeneous Cell Size

We derived the probability distribution in Eq. 2 assuming the plated host cells are of identical size and volume. This may not necessarily be the case as each cell exists at different stages of the mitotic cycle, will attach to the plate bottom at random locations, and contain deformities in shape and size. Assuming cells cover the entire surface of the well bottom, Pineda et al. (42) showed that the cell size proportion p_i for cell i is gamma distributed with probability density

$$f(p_i) = \frac{M^\nu \nu^\nu p_i^{\nu-1} \exp(-\nu M p_i)}{\Gamma(\nu)}, \tag{S11}$$

where ν is a parameter that can be estimated, for example, by fitting imaging data of cells. Under a specific realization of cell size distributions $\{p_1, \cdots, p_M\}$, we define A_i^r as the event that cell i is infected by exactly r viruses with probability

$$\Pr(A_i^r) = \binom{N}{r} p_i^r (1-p_i)^{N-r}. \tag{S12}$$

Using the inclusion-exclusion principle as above, we derive the conditional probability distribution of the number of cells M_r that were infected by exactly r viruses as

$$\begin{aligned}
\Pr(M_r = m_r | p_1, \cdots, p_M) &= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \sum_{|\{i_w\}|=j} \Pr\left(\bigcap_{w=1}^{j} A_{i_w}^r \right) \\
&= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \sum_{|\{i_w\}|=j} \binom{N}{r, \cdots, r, (N-rj)} p_{i_1}^r \cdots p_{i_j}^r \left(1 - \sum_{w=1}^{j} p_{i_w}\right)^{N-rj} \\
&= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \sum_{|\{i_w\}|=j} \frac{N!}{(r!)^j (N-rj)!} \left(\prod_{w=1}^{j} p_{i_w} \right)^r \left(1 - \sum_{w=1}^{j} p_{i_w}\right)^{N-rj}.
\end{aligned} \tag{S13}$$

In order to obtain the full probability, we first take note that each cell size proportion p_i is dependent on each other as they are constrained by $\sum_i^M p_i = 1$. We avoid this dependency by noticing the expression in Eq. S11 approaches zero very rapidly as p_i moves away from the expected value $1/M$. If we define a sufficiently large proportion $\hat{p}$ such that the interval $[0, \hat{p}]$ contains the majority of the area under the probability density in Eq. S11, we can make the approximation

$$\begin{aligned}
\Pr(M_r = m_r) &= \int_0^1 \cdots \int_0^1 \Pr(M_r = m_r | p_1, \cdots, p_M) f(p_1, \cdots, p_M) \mathrm{d}p_1 \cdots \mathrm{d}p_M \\
&\approx \int_0^{\hat{p}} \cdots \int_0^{\hat{p}} \Pr(M_r = m_r | p_1, \cdots, p_M) f(p_1) \cdots f(p_M) \mathrm{d}p_1 \cdots \mathrm{d}p_M \\
&= \left[\frac{M^\nu \nu^\nu e^{-\nu}}{\Gamma(\nu)} \right]^M \int_0^{\hat{p}} \cdots \int_0^{\hat{p}} \Pr(M_r = m_r | p_1, \cdots, p_M) \left(\prod_{w=1}^{M} p_w \right)^{\nu-1} \mathrm{d}p_1 \cdots \mathrm{d}p_M.
\end{aligned} \tag{S14}$$

It is clear that introducing cell size inhomogeneity dramatically increases the complexity of our probabilistic SMOI model. For relatively small numbers of cells M, image processing can be used to determine an estimation of a particular realization of cell size distribution $\{p_1, \cdots, p_M\}$ for a given experiment and factored into Eq. S13. Note that once the probability distribution of cell counts $\{M_0, \cdots, M_N\}$ is determined for a given realization of cell sizes $\{p_1, \cdots, p_M\}$, all subsequent analysis and derivations follow the same way as in the homogeneous cell size assumption.

Coinfection

As a vector for infection, the primary function of a single virus particle is to deliver its genetic contents into the host cell cytoplasm or nucleus (10–12). The typical model for viral infection assumes each virus contains all the genetic material required to replicate within a host cell (14, 15). Certain plant and fungi viruses, however, require two or more particles to successfully replicate within a host cell since each particle contains only part of the complete genome (31). Similarly, RNA viruses that target animal cells undergo error prone replication, resulting in partially complete genome sequences. These damaged viral genes may encode proteins needed for the host cell to successfully replicate new viruses. In this case, regardless of a successful viral infection, new viruses capable of infecting further host cells will not be produced. Additional viral infections that contain the missing sequence fragments, though, can "rescue" the cell's ability to replicate the virus, a phenomenon known as coinfection. In the context of our definition of SMOI, we now make the distinction between M_r, the number of cells that have been infected by viral genomes from exactly r distinct virus particles, and M_r^*, the number of cells that are fully capable of replicating new functioning viruses upon undergoing r distinct viral infections. It is immediate that each $M_r^* \leq M_r$ and their sum $M^* \equiv \sum_{r=1}^{N} \leq M - M_0$, so the results in Eqs. 9 and 12 are not sufficient to quantify the total number of virus-producing cells.

In order to model coinfection, we need to consider the genome of the virus species of interest. Specifically, we assume the genome is made up of G distinct genes. For example, many variants of HIV-1 carry a gene sequence containing $G = 9$ genes (10). In our model, we assume each gene encodes a protein that is essential for replication. Though individual nucleotide changes due to random mutations may result in an amino acid chain that is no longer functioning, some genes may be robust to these changes due to codon degeneracy or the gene's shear length (43). Thus, we assume each gene $g = 1, \cdots, G$ contained within a viral particle has a probability q_g of losing function. If a cell is infected by exactly r viral genomes, we define B_g^r as the event that gene g is still no longer functional, so that $\Pr(B_g^r) = q_g^r$. To quantify the probability that k genes are no longer functional in a host cell that has been infected by exactly r viral genomes, we use the inclusion-exclusion principle (35) to derive

$$\begin{aligned}
\Pr\left(\text{“}k\text{ failed genes given }r\text{ infections”}\right) &= \sum_{j=k}^{G} (-1)^{j-k} \binom{j}{k} \sum_{\substack{I \subset \{1,\cdots,G\} \\ |I| = j}} \Pr\left(\bigcap_{g \in I} B_g^r \right) \\
&= \sum_{j=k}^{G} (-1)^{j-k} \binom{j}{k} \sum_{\sigma_1 = 0}^{1} \cdots \sum_{\sigma_G = 0}^{1} \mathbb{1}_{\sum_{g=1}^{G} \sigma_g = j} \prod_{g=1}^{G} q_g^{\sigma_g r}, \qquad \text{(S15)}
\end{aligned}$$

where $\mathbb{1}_{\sum_{g=1}^{G} \sigma_g = j}$ is an indicator function that returns zero when the number of nonzero σ_g is not exactly j. The infected cell is only capable of producing viable viruses if none of the genes have failed and is equivalent to setting $k = 0$ in Eq. S15. Then we define the probability H_r that a cell infected by exactly r viral genomes will successfully produce new viruses as

$$H_r = \sum_{j=0}^{G} (-1)^j \sum_{\sigma_1 = 0}^{1} \cdots \sum_{\sigma_G = 0}^{1} \mathbb{1}_{\sum_{g=1}^{G} \sigma_g = j} \prod_{g=1}^{G} q_g^{\sigma_g r}. \qquad \text{(S16)}$$

Note that the probability that a cell not infected by any viral genome will produce viruses is $H_0 = 0$. Then, given an SMOI $\{M_0, \cdots, M_N\}$, the number of cells M_r^* capable of virus replication after being infected by exactly r viral genomes is binomially distributed with parameters M_r and H_r. The probability of M^* cells producing viruses is given by

$$\Pr\left(M^* = m | M_0, \cdots, M_N, M, N\right) = \sum_{M_1^*, \cdots, M_N^*} \binom{m}{M_1^*, \cdots, M_N^*} \prod_{r=1}^{N} \binom{M_r}{M_r^*} H_r^{M_r^*} (1 - H_r)^{M_r - M_r^*}. \qquad \text{(S17)}$$

If we let $m = 0$ and sum over the density in Eq. 5 for all possible SMOI, given an IU count N, we can derive the probability of observing a cytopathic effect as

$$\begin{aligned}
\Pr(\text{“Cytopathic effect”}|N) &= 1 - \Pr\left(M^* = 0 | N\right) \\
&= 1 - \sum_{M_0, \cdots, M_N} \frac{1}{M^N} \binom{M}{M_0, \cdots, M_N} \binom{N}{0, \cdots, 0, 1, \cdots, 1, \cdots, N, \cdots, N} \prod_{r=1}^{N} (1 - H_r)^{M_r} \\
&= 1 - \frac{M!N!}{M^N} \prod_{r=0}^{N} \sum_{M_r = 0}^{M} \frac{(1 - H_r)^{M_r}}{M_r! (r!)^{M_r}} \\
&\approx 1 - \frac{M!N!}{M^N} \prod_{r=0}^{N} \exp\left[\frac{1 - H_r}{r!} \right] \\
&= 1 - \frac{M!N!}{M^N} \exp\left[\sum_{r=0}^{N} \frac{1 - H_r}{r!} \right], \qquad \text{(S18)}
\end{aligned}$$

where the approximation is due to the assumption that the number of cells M is large. For intermediate values of N, computing the summation in the exponential is numerically viable, assuming the probabilities of gene failure $q_1, \cdots, q_G$ are known.

Though this expression may be used in place of Eq. 12 to analyze some virus quantification assays, for large values of N, numerically evaluating H_r becomes computationally expensive.

Viral Interference

To infect healthy cells, all species of viruses must undergo a series of events including cell attachment, entry via membrane fusion or endocytosis, and intracellular transport. Retroviruses, such as HIV-1, must also undergo reverse transcription, nuclear pore transport, and DNA integration in order to use the host cell's transcription machinery to produce viral protein. In the models developed in this paper, the probabilities of success for each of these processes was assumed to be subsumed into the *a priori* estimated particle to PFU ratio Q. However, for certain retroviruses, it has been observed that after an initial infection, subsequent infections from the same virus species become less likely (44, 45). This phenomenon, known as viral interference, is often due to the host producing new viral proteins after a refractory period that can inhibit one or more of the intracellular processes leading to integration of subsequent viral infections. To include this dynamic into our models, we first decouple the probabilities of integration from Q and define N as the number of viruses that have successfully completed viral entry into the host cytoplasm, but before all intracellular processes that lead to integration. Note that all of our results concerning the statistical multiplicity of infection (SMOI) still hold and we make the distinction between the number M_r of cells infected by r of the N infectious units and the number M_s^* of cells with exactly s integrations. Furthermore, some species of virus can contain multiple copies of their genome, such as HIV-1 which contains two copies per particle (10). Let a be the number of genomes contained in a single virus particle to be integrated into the host cell. Then the maximum number of possible integrations for a cell from M_r is ra. Let p_s be the probability of a viral genome integrating into the host DNA given that $s-1$ integrations have already occurred. Define $H_{r,s}$ as the probability a cell contains s successful integrations given that it was infected by exactly r distinct virus particles and is given by

$$H_{r,s} = \begin{cases} p_1 p_2 \cdots p_s \left(1 - p_{s+1}\right)^{ra-s} & 0 \le s \le ra \\ 0 & s > ra. \end{cases} \tag{S19}$$

If we define $M_{r,s}^*$ as the number of cells with s integrations after infection by exactly r virus particles, then given an SMOI $\{M_0, \cdots, M_N\}$ and N, we can derive the probability function

$$\Pr(M_{r,s}^* = m | M_0, \cdots, M_N, N) = \binom{M_r}{m} H_{r,s}^m \left(1 - H_{r,s}\right)^{M_r - m}. \tag{S20}$$

Noting that $M_s^* = \sum_{r=0}^{N} M_{r,s}^*$ is the number of cells with exactly s integrations, we can use Eqs. 6 and S20 to derive the expected value as

$$\begin{aligned} \mathrm{E}\left[M_s^* | N\right] &= \sum_{r=0}^{N} \mathrm{E}\left[M_{r,s}^* | N\right] \\ &= \sum_{r=0}^{N} H_{r,s} \mathrm{E}\left[M_r | N\right] \\ &= M e^{-\mu} \sum_{r=0}^{N} \frac{H_{r,s} \mu^r}{r!}, \end{aligned} \tag{S21}$$

where $\mu = \frac{N}{M}$. Note that if we are concerned with the total number $M^* = M - M_0^*$ of cells with at least one integration, as is the case for the probability distributions derived for assays employing serial dilution, issue of viral interference is negligible, allowing us to subsume the probability of the first integration into the particle to PFU ratio Q as before and leave all subsequent virus quantification analysis unchanged from the results in Section 3.1 and 3.2. However, for assays that attempts to quantify the total number of integrations, such as the luciferase reporter assay, the expectation in Eq. S21 can be used, assuming the probabilities $p_1, \cdots, p_N$ have *a priori* been estimated.