RESEARCH ARTICLE

Estimating the basic reproduction number of a pathogen in a single host when only a single founder successfully infects

**Vruj Patel, John L. Spouge*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

* spouge@nih.gov

OPEN ACCESS

Citation: Patel V, Spouge JL (2020) Estimating the basic reproduction number of a pathogen in a single host when only a single founder successfully infects. PLoS ONE 15(1): e0227127. https://doi.org/10.1371/journal.pone.0227127

Editor: Max O. Souza, Universidade Federal Fluminense, BRAZIL

Received: June 10, 2019

Accepted: December 12, 2019

Published: January 10, 2020

Data Availability Statement: A JAVA program simulating the Delta and Gamma models is available without restriction on GitHub at https://github.com/mrvruj/hivSim. All relevant data are within the manuscript and its Supporting Information files.

Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Competing interests: The authors have declared that no competing interests exist.

Abstract

If viruses or other pathogens infect a single host, the outcome of infection may depend on the initial basic reproduction number R_0, the expected number of host cells infected by a single infected cell. This article shows that sometimes, phylogenetic models can estimate the initial R_0, using only sequences sampled from the pathogenic population during its exponential growth or shortly thereafter. When evaluated by simulations mimicking the bursting viral reproduction of HIV and simultaneous sampling of HIV gp120 sequences during early viremia, the estimated R_0 displayed useful accuracies in achievable experimental designs. Estimates of R_0 have several potential applications to investigators interested in the progress of infection in single hosts, including: (1) timing a pathogen's movement through different microenvironments; (2) timing the change points in a pathogen's mode of spread (e.g., timing the change from cell-free spread to cell-to-cell spread, or vice versa, in an HIV infection); (3) quantifying the impact different initial microenvironments have on pathogens (e.g., in mucosal challenge with HIV, quantifying the impact that the presence or absence of mucosal infection has on R_0); (4) quantifying subtle changes in infectability in therapeutic trials (either human or animal), even when therapies do not produce total sterilizing immunity; and (5) providing a variable predictive of the clinical efficacy of prophylactic therapies.

Introduction

When viruses or other pathogens infect a single host, the basic reproduction number R_0 is the expected number of cells infected by a single infected cell [1]. The initial R_0 is a fundamental determinant of whether an infecting viral population will establish itself in the host. On one hand, if $R_0 < 1$, the viral invaders reproduce below replacement and will go extinct. On the other hand, if R_0 is slightly greater than 1, an initial virus has a small positive probability of amplifying into a systemic infection, and if R_0 is large, infection is all but inevitable.

The initial basic reproduction number R_0 is therefore a continuous variable with direct biological pertinence to infection. As such, it may have many underappreciated applications. As a general example, consider therapeutic trials. More specifically, consider as a motivating application pre-clinical tests of HIV therapies like vaccines in animal models. In this context, the macaque model is popular, because simian immunodeficiency virus can cause a disease progression resembling AIDS in humans [2]. Historically, macaque trials often used a single high-dose intravenous or mucosal inoculation to ensure almost certain infection of unprotected animals [3]. The consequent assessment of therapeutic efficacy depends primarily on binary categorical data, i.e., whether or not infection occurred in treated animals [4,5]. New experimental designs, notably repeated low-dose challenges, have improved statistical power in animal trials [3,6], but do not remove the intrinsic statistical limitations of binary data. The ability to estimate a continuous variable like the initial R_0 in HIV infection would in principle permit a statistically more powerful analysis of therapeutic efficacies.

The initial R_0 is likely a primary determinant of whether systemic infection occurs, and initial microenvironments with relatively few target cells probably impede systemic infection by HIV. Consider, e.g., that the number of per-act transmissions per 10,000 exposures varies considerably by route of infection [7]. For sexual exposure, the number ranges from less than 4 to 138; for needle-sharing, it is about 63. In contrast, for vertical transmission between mother and child, the number is 2260; for blood transfusion, it is 9250. Thus, the initial microenvironment may starkly limit the reproduction of HIV until the virus escapes into systemic circulation, where target cells are plentiful.

Unfortunately, practical thresholds for HIV detection make the initial R_0 inaccessible to direct measurement, because on average, viremia is delayed until about 10 days after exposure to HIV [8,9]. Modern techniques for measuring the abundance of HIV [10] have yielded estimates, e.g., $R_0 \approx 6$ [11] or $R_0 \approx 8$ [12]. These estimates of R_0 pertain to viremia, however, when there are at least 20 viruses/ml (the current lower limit of detectability) and thus about 10^5 viruses in the total blood volume of 5L [13]. By the time HIV is detectable in blood samples and R_0 can be measured directly, infection has long since established itself.

The viral dosage may vary considerably between the modes of transmission above, and in sexual or mother-to-child transmission of HIV, e.g., the genetic diversity at early stages of infection usually reflects the number of viruses founding the infection [14–16]. In fact, however, about 80% of all HIV infections arise from a single founding viral sequence [17–19]. Moreover, the design of repeated low-dose challenge animal trials is likely to cause infections with a single founding virus. Thus, because of its practical importance, the case of a single founder virus is a convenient starting point for mathematical analysis, and the rest of this article assumes a single founder.

Although most of this article is self-contained, it continues a scientific program started in [20]. Direct measurement of R_0 early in infection may be impractical, but Fig 1 in [20] showed that the initial R_0 displays its footprint in HIV sequences sampled in early viremia, during exponential expansion of the viral population. To describe the essence of Fig 1 in [20], if only two daughters of the founder successfully contribute descendants to the viremia, and one daughter has a novel mutation away from the founder, about half the sequences sampled in viremia have the mutation. In contrast, if the founder has many daughters successfully contributing descendants to the viremia, far fewer than half the sequences sampled in viremia are likely to have any given mutation.

The structure of this article is as follows: the theory section applies standard statistical procedures to yield a robust method for estimating R_0 from sequence data. The Methods section then describes the simulation of a continuous-time branching process whose parameters are pertinent to sampling sequences of the HIV gp120 gene. The process is the "Gamma model" of [20], a special Bellman-Harris process [21] that idealizes HIV reproduction. The Results section displays the accuracy of our estimator in recovering R_0 in the idealized simulation and the Discussion section examines some consequences of the theory. Finally, the Supporting

Information compares our estimator to other (inferior) statistical techniques that we applied to recover R_0 from the same idealized simulation.

All approximations in this article are uncontrolled, i.e., we cannot provide bounds on the error that the approximations cause. Usually without comment, therefore, we rely on simulations of the Gamma model to assess their accuracy. Parameter regimes not pertinent to HIV reproduction and the sampling of gp120 sequences require separate assessment and are beyond the immediate practical purview of this article.

Finally, as motivation, in the context of the Gamma model, our estimator compares favorably with state-of-the-art methods. It has an exceptionally simple analytic form, e.g., yielding a negligible computation time in comparison to exact Bayesian calculations. Moreover, under methods analogous to ours, the coalescent process yields the same estimator as continuous-time branching process models of population growth. Under the Gamma model for HIV reproduction, however, the estimator has a singularity, making it useless for quantitation if $R_0 \geq e \approx 2.7$.

Theory

The following set-up assumes that a single founder virus has infected a single host (e.g., a single HIV infects a human). The set-up is mostly self-contained, drawing only on a few critical approximations presented elsewhere with some mathematical foundations [20]. First, before modeling viral ancestry, we examine the sampling of the viral sequences.

With "≔" denoting a definition, define $(m) \coloneqq \{1, 2, \ldots, m\}$. Fix an alphabet Λ, e.g., the unambiguous nucleotide alphabet $\Lambda = \{a, c, g, t\}$. In practice, sequence analysis must invoke a strategy for handling anomalous characters in an alignment (e.g., ambiguous nucleotides or gap characters). Sometimes, anomalous characters are infrequent, so that as an acceptable approximation, the analysis can treat them as ordinary characters by enlarging its alphabet. Sometimes, the analysis simply omits columns containing them. Without further comment, the following assumes that the practical analysis has adopted an unspecified strategy for handling anomalous characters.

Consider a set $S_\bullet \coloneqq (S_m\colon m \in (M))$ consisting of M sequences. The sequences are sampled simultaneously from the descendants of a single founder sequence Φ. Align the sequences $S_\bullet$, so that the sequences $S_\bullet$ form an alignment matrix $S_{\bullet,\bullet} \coloneqq \{S_{m,n}\colon m \in (M), n \in (N)\}$ of N columns. Given an unspecified strategy for processing sequences $S_\bullet$ into $S_{\bullet,\bullet}$, the analysis here simply starts with the alignment matrix $S_{\bullet,\bullet}$. Implicitly, Φ aligns with $S_{\bullet,\bullet}$, so the letter Φ_n in Φ is ancestral to each letter $S_{m,n}$ in the matrix column $n \in (N)$ (where $S_{m,n}$ is from the sequence S_m). In practice, Φ is often unknown, a complication we handle shortly.

The Iverson bracket for indicator random variates is a standard notation [22]: let $[A] = 1$ if the statement A is true, and $[A] = 0$ otherwise. Let $M_n(L) \coloneqq \sum_{m=1}^{M} [S_{m,n} = L]$ $(n \in (N))$ count the instances of letter $L \in \Lambda$ in column n of $S_{\bullet,\bullet}$. Given the strategy for handling anomalous characters, $M = \Sigma_{(L \in \Lambda)} M_n(L)$.

The difference $D_n \coloneqq D_n(\Phi) \coloneqq M - M_n(\Phi_n)$ counts letters in column n that have mutated away from the founder Φ; let $\mathbf{D} \coloneqq \mathbf{D}(\Phi) \coloneqq (D_n(\Phi)\colon n \in (N))$. Given $\mathbf{D}$, let $\eta_m \coloneqq \eta_m(\Phi) \coloneqq \sum_{n=1}^{N} [D_n(\Phi) = m]$ count the alignment columns $n \in (N)$ where m letters differ from the founder letter Φ_n, and define the site frequency spectrum (SFS) as $\boldsymbol{\eta} \coloneqq \boldsymbol{\eta}(\Phi) \coloneqq (\eta_m(\Phi)\colon m \in (M))$. Typically, Φ is unknown, so $\boldsymbol{\eta}$ is not observable.

To develop a statistical model for $\boldsymbol{\eta}$, let ε_n be the probability of a mutation per base per generation in column n of the alignment. As in the infinite-sites model [23], we neglect the extremely rare possibility that two or more mutations occur in the ancestry of a single letter $S_{m,n}$. Let $\mu = \sum_{n=1}^{N} \varepsilon_n$ be the expected number of novel mutations per generation in the sequences. Despite its biological importance, the effects of preferential selection on sequence data are practically imperceptible during the first six months of HIV infection (see the first paragraph in the Materials and Methods of [19] and Fig 1 in [24]). Assume therefore that (ε_n: $n \in (N)$) are all small, and that novel mutations are all independent. To a good approximation, in every daughter the counts of novel mutations are independent Poisson variates with fixed mean μ.

For linguistic convenience, let every virus be both her own ancestor and her own descendant. Let A_m count the non-founder viral ancestors with m descendants in the sample, and define as in [20] the ancestral sample frequency spectrum (AFS), $\mathbf{A} \coloneqq (A_m: m \in (M))$. (Each sampled sequence contributes to A_1, e.g., because by convention, each sampled sequence is its only descendant in the sample). Under an infinite-sites model, every novel mutation occurs in a different column of the alignment. Every alignment column with m mutations therefore corresponds to a novel mutation in an ancestor with m descendants in the sample [25]. Given $\mathbf{A}$, the coordinates of $\boldsymbol{\eta}$ are independent Poisson variates, with η_m having mean μA_m (see, e.g., Theorem 1 in [20]). Accordingly, the relationship is written as $\boldsymbol{\eta} =_d$ Poission($\mu\mathbf{A}$), where "$=_d$" indicates equality of distributions.

Eq (17) in [20] used the law of total variance to write

$$\sigma^2(\eta_m) = \mathbb{E}[\sigma^2(\eta_m|\mathbf{A})] + \sigma^2(\mathbb{E}[\eta_m|\mathbf{A}]) = \mathbb{E}[\mu A_m] + \sigma^2(\mu A_m) = \mu\mathbb{E}A_m + \mu^2\sigma^2(A_m). \qquad (1)$$

Now, we restrict the discourse to the Gamma model for HIV gp120 (as detailed in the Methods section, which need not be read yet). Simulations of the Gamma model showed [20] that for $\mu = 0.0551$ (the value for HIV gp120), the typical magnitude of the ratio $\mu^2\sigma^2(A_m)/(\mu\mathbb{E}A_m)$ from Eq (1) was at most about 18%. For $\mu = 0.0551$ in the Gamma model, therefore, the mutational variance $\mu\mathbb{E}A_m$ makes the dominant contribution to $\sigma^2(\eta_m)$. The form of Eq (1) shows that the dominance remains robust to varying μ (particularly decreasing it), as long as the ratio $\mu^2\sigma^2(A_m)/(\mu\mathbb{E}A_m)$ remains small (say, less than 50%, occurring about $\mu \approx 0.0551 \times (0.50/0.18) \approx 0.153$).

Let $a_m \coloneqq \mathbb{E}A_m$ and $\mathbf{a} \coloneqq (a_m: m \in (M))$. Given the distributional equality $\boldsymbol{\eta} =_d$ Poission($\mu\mathbf{A}$), our observations on Eq (1) therefore suggest that the distributional approximation $\boldsymbol{\eta} \approx_d$ Poission($\mu\mathbf{a}$) pertains, as follows. In the present context (the distribution of $\boldsymbol{\eta}$ under $\mu = 0.0551$ in the Gamma model), the variation of $\mathbf{A}$ contributes little to the variance of η_m: effectively, as noted by other authors [1,19,24], $\mathbf{A}$ behaves as though it were the constant $a_m = \mathbb{E}A_m$ when contributing to random fluctuations in η_m. In effect, the approximation $A_m \approx a_m$ treats Eq (1) as an expansion in μ around $\mu = 0$ and drops terms quadratic in μ to retain the approximation $\sigma^2(\eta_m) \approx \mu\mathbb{E}A_m$. As a linear approximation, it should improve as μ decreases and worsen as μ increases.

To avoid distracting subscripts in the following equations, let $r \coloneqq R_0$ denote the basic reproduction number R_0 from the Introduction. For some practical purposes, HIV reproduces almost in lockstep, with synchronous generations (see the Delta model of [20], a Galton-Watson branching process [26]). Let G count the generations of HIV after host infection. To summarize the previous paragraph,

$$p(\boldsymbol{\eta}|G, r, M) \approx \prod_{m=1}^{M} e^{-\mu a_m} \frac{(\mu a_m)^{\eta_m}}{\eta_m!}. \qquad (2)$$

In any ancestry with G synchronous generations, $MG = \sum_{m=1}^{M} mA_m$. (Proof: count the ancestors in each of the generations $g = 1, 2, \ldots, G$, accounting for the multiplicity m of their sampled descendants. The total count is equivalent to counting each of the M samples G

times). Take expectations to derive

$$MG = \sum_{m=1}^{M} m a_m. \tag{3}$$

Reference [20] showed that for the Gamma model,

$$a_m^{(\Delta)} := a_m^{(\Delta)}(r) := \binom{M}{m} \sum_{g=1}^{\infty} (r^{-g})^{m-1} (1 - r^{-g})^{M-m} = \sum_{g=1}^{\infty} r^g \binom{M}{m} (r^{-g})^{m} (1 - r^{-g})^{M-m} \tag{4}$$

accurately approximated $a_m = \mathbb{E}A_m$ (m = 2, 3 . . ., M). The heuristic behind the approximation follows. To a good approximation, HIV has synchronous generations g = 1, 2, . . ., G. On average, generation g contains r^g individuals. Each viral sequence in the sample therefore has an approximate probability r^{-g} of descending from any particular individual $\mathscr{g}$ in generation g. Thus, the probability that $\mathscr{g}$ has m descendants in a sample of size M is approximately the binomial probability on the right of Eq (4). Sum the binomial probability over the individuals $\mathscr{g}$ in generation g (on average, r^g in number) and then over all generations g = 1, 2, . . ., G. Let G tend to infinity to derive Eq (4).

Eqs (3) and (4) therefore show that if in the Gamma model $a_1 \approx a_1^{(\Delta)}$, then $\lim_{G\to\infty}(MG - a_1)$ is approximately the (finite) quantity

$$a_{\bullet}^{(\Delta)} := a_{\bullet}^{(\Delta)}(r) := \sum_{m=2}^{M} m a_m^{(\Delta)}. \tag{5}$$

In statistical notations, "•" often suggests a sum, as in $a_{\bullet}^{(\Delta)}$.

The variable $\boldsymbol{\eta}$ in Eq (2) depends on the unknown founder sequence Φ. To relate $\boldsymbol{\eta}$ to an observable, define $\tilde{M}_n := \max_{L\in\Lambda} M_n(L)$, the maximum count of any single letter in column n. Loosely, $\tilde{D}_n := M - \tilde{M}_n$ then counts minority letters in column n. Unlike D_n, the observable $\tilde{D}_n$ has no dependency on the founder sequence Φ. Let $\lfloor x \rfloor := \max\{i \in \mathbb{Z} : i \leq x\}$ denote the floor function, with $\tilde{M} := \lfloor M/2 \rfloor$. Define the folded SFS $\tilde{\boldsymbol{\eta}} := (\tilde{\eta}_m : m \in (\tilde{M}))$ [27], with

$$\tilde{\eta}_m := \sum_{n=1}^{N} [\tilde{D}_n = m] = \sum_{n=1}^{N} ([D_n = m] + [D_n = M - m]) = \eta_m + \eta_{M-m} \tag{6}$$

for m = 1, 2, . . ., $\lfloor(M - 1)/2\rfloor$), where the second equality holds for an infinite-sites model (which we have assumed). If M is odd, $\lfloor (M-1)/2 \rfloor = \lfloor M/2 \rfloor = \tilde{M}$, so the definition of $\tilde{\boldsymbol{\eta}}$ is complete. If M is even, the pattern of pairs $\eta_m + \eta_{M-m}$ displayed in Eq (6) fails for $m = \tilde{M}$, because $\eta_{\tilde{M}}$ cannot be paired with a distinct $\eta_{M-\tilde{M}}$. If M is even, therefore, define $\tilde{\eta}_{\tilde{M}} := \eta_{\tilde{M}}$ for $m = \tilde{M}$. Loosely, in all cases, $\tilde{\eta}_m$ counts columns n where the number of minority letters equals $m(m \in (\tilde{M}))$.

The folded AFS $\tilde{\mathbf{A}} := (\tilde{A}_m : m \in (\tilde{M}))$ inherits the pattern for $\tilde{\boldsymbol{\eta}}$ established in Eq (6): $\tilde{A}_m := A_m + A_{M-m}$ for m = 1, 2, . . ., $\lfloor(M - 1)/2\rfloor$; if M is even, $\tilde{A}_{\tilde{M}} := A_{\tilde{M}}$. Henceforth and without comment, the same pattern generates folded quantities (denoted by over-tildes) from unfolded quantities, e.g., $\tilde{\mathbf{a}} := \mathbb{E}\tilde{\mathbf{A}}$. The folded SFS $\tilde{\boldsymbol{\eta}}$ inherits an approximate Poisson distribution from the SFS $\boldsymbol{\eta}$:

$$p(\tilde{\boldsymbol{\eta}}|G,r,M) \approx \prod_{m=1}^{\tilde{M}} e^{-\mu\tilde{a}_m} \frac{(\mu\tilde{a}_m)^{\tilde{\eta}_m}}{\tilde{\eta}_m!}. \tag{7}$$

Because

$$\lim_{G\to\infty}(MG - \tilde{a}_1) = \lim_{G\to\infty}(MG - a_1 - a_{M-1}) \approx a_{\bullet}^{(\Delta)} - a_{M-1}^{(\Delta)}, \tag{8}$$

the pattern suggests imposing the definition

$$\tilde{a}_{\bullet}^{(\Delta)} \coloneqq a_{\bullet}^{(\Delta)} - a_{M-1}^{(\Delta)}. \tag{9}$$

Eq (7) applied to the observable $\tilde{\boldsymbol{\eta}}$ yields a maximum likelihood estimate (MLE) $(\hat{G}, \hat{r})$. The asymptotic properties of an MLE make $\hat{r}$ a reasonable benchmark for other statistical estimates of r. Recall Eqs (8) and (9) relating $\tilde{a}_1$ and MG, and take natural logarithms in Eq (7) to derive the (approximate) log-likelihood

$$\ln L(G,r) \coloneqq \ln p(\tilde{\boldsymbol{\eta}}|G,r,M) \equiv -\mu(MG - \tilde{a}_{\bullet}) + \tilde{\eta}_1 \ln (MG - \tilde{a}_{\bullet}) + \sum_{m=2}^{\tilde{M}} (-\mu\tilde{a}_m + \tilde{\eta}_m \ln \tilde{a}_m), \tag{10}$$

where "≡" indicates an equality of functions, possibly ignoring an irrelevant additive term depending only on data (e.g., a term equaling a function of $\tilde{\boldsymbol{\eta}}$).

As a useful approximation, the following treats G as a continuous variate. Now, for any value of r (and not just $r = \hat{r}$),

$$\begin{aligned} \max_G \ln L(G,r) &= \max_G[-\mu(MG - \tilde{a}_{\bullet}) + \tilde{\eta}_1 \ln (MG - \tilde{a}_{\bullet})] + \sum_{m=2}^{\tilde{M}} (-\mu\tilde{a}_m + \tilde{\eta}_m \ln \tilde{a}_m) \\ &= -\tilde{\eta}_1 + \tilde{\eta}_1 \ln (\tilde{\eta}_1/\mu) + \sum_{m=2}^{\tilde{M}} (-\mu\tilde{a}_m + \tilde{\eta}_m \ln \tilde{a}_m) \end{aligned}, \tag{11}$$

where the second equality holds if the maximum is an internal maximum, so the argument $\hat{G}$ is determined by setting the derivative with respect to G equal to 0 (i.e., if $\hat{G}$ satisfies $\mu(M\hat{G} - \tilde{a}_{\bullet}) = \tilde{\eta}_1$). An MLE $\hat{r}$ then maximizes the profile log-likelihood, defined as

$$\ln L(r) \coloneqq \max_G \ln L(G,r) \equiv \sum_{m=2}^{\tilde{M}} [-\mu\tilde{a}_m(r) + \tilde{\eta}_m \ln \tilde{a}_m(r)]. \tag{12}$$

Now, $\tilde{a}_m(1) = \tilde{a}_m(\infty) = 0$ for $m = 2, 3, \ldots, \tilde{M}$. Because $\ln L(r) \downarrow -\infty$ at the boundaries of the interval $(1, \infty)$, a MLE $\hat{r} > 1$ therefore exists, such that

$$\frac{d}{dr} \ln L(r) = \sum_{m=2}^{\tilde{M}} \left[-\mu\tilde{a}'_m(r) + \tilde{\eta}_m \frac{\tilde{a}'_m(r)}{\tilde{a}_m(r)} \right] = 0 \tag{13}$$

has a root at $r = \hat{r}$. In the following, if an MLE lacked an explicit analytic expression, a golden-section search for maxima determined it numerically.

Unfortunately, the direct method of determining $\hat{r}$ by substituting $a_m(r) \approx a_m^{(\Delta)}(r)$ and then maximizing Eq (12) or solving Eq (13) entails many undependable numerical computations, because Eq (4) for $a_m^{(\Delta)}(r)$ requires multiplying unreasonably large and small numbers, followed by adding the resulting products with great precision. For completeness, the

Supporting Information develops an approximate MLE $\hat{r}^{(\Delta)}$ by approximating $a_m(r)$ with $a_m^{(\Delta)}(r)$ and compares it with our best estimator, derived as follows. Approximate the sum in $a_m^{(\Delta)}$ by an integral (the first term of an Euler-Maclaurin series [28]): because $(1 - r^0)^{M-m} = 0$ for $m = 2, 3, \ldots, M - 2$,

$$\begin{aligned} \binom{M}{m} \sum_{g=1}^{\infty} (r^{-g})^{m-1} (1 - r^{-g})^{M-m} &\approx \binom{M}{m} \int_0^{\infty} (r^{-g})^{m-1} (1 - r^{-g})^{M-m} dg \\ &= \binom{M}{m} \int_0^1 y^{m-1} (1 - y)^{M-m} \frac{dy}{y \ln r} \qquad (14) \\ &= \frac{M}{m(m-1)} \frac{1}{\ln r} \end{aligned}$$

where the final equality derives from the evaluation of a beta integral. Comparison of Eq (4) and the two sides of Eq (14) shows that

$$a_m^{(I)} \coloneqq \frac{M}{m(m-1)} \frac{1}{\ln r} \qquad (15)$$

approximates a_m for $m = 2, 3, \ldots, M - 2$.

After substituting $a_m^{(I)}(r)$ for $a_m(r)$ in Eq (13) and unfolding all folded quantities,

$$\sum_{m=2}^{M-2} \left[\frac{\mu M}{m(m-1)} \frac{1}{r \ln^2 r} - \eta_m \frac{1}{r \ln r} \right] = 0. \qquad (16)$$

Telescoping cancellation yields

$$\sum_{m=2}^{M-2} \frac{1}{m(m-1)} = \sum_{m=2}^{M-2} \left(\frac{1}{m-1} - \frac{1}{m} \right) = 1 - \frac{1}{M-2}. \qquad (17)$$

To estimate r, define $\tilde{\eta}_{-1} \coloneqq \sum_{m=2}^{M-2} \eta_m = \sum_{m=2}^{\tilde{M}} \tilde{\eta}_m$ (an observable, the sum of all $\tilde{\eta}_m$ except $\tilde{\eta}_1$), so the root $\hat{r}^{(I)}$ of Eq (16) satisfies

$$\ln \hat{r}^{(I)} = \frac{\mu M}{\sum_{m=2}^{M-2} \eta_m} \left(1 - \frac{1}{M-2} \right) = \frac{\mu M}{\tilde{\eta}_{-1}} \left(1 - \frac{1}{M-2} \right). \qquad (18)$$

The sum of independent Poisson variates is Poisson distributed, so the variate $\tilde{\eta}_{-1}$ is approximately Poisson distributed, with $\sigma^2(\tilde{\eta}_{-1}) \approx \mathbb{E}\tilde{\eta}_{-1}$. Define $\ln^2 x \coloneqq (\ln x)^2$. From a linear Taylor series approximation, the approximation $\operatorname{var} f(\tilde{\eta}_{-1}) \approx [f'(\mathbb{E}\tilde{\eta}_{-1})]^2 \operatorname{var} \tilde{\eta}_{-1}$ with $f(\eta) = \eta^{-1}$ yields

$$\hat{\sigma}^2(\ln \hat{r}^{(I)}) \approx \frac{\left[\mu M \left(1 - \frac{1}{M-2}\right)\right]^2}{(\mathbb{E}\tilde{\eta}_{-1})^4} \mathbb{E}\tilde{\eta}_{-1} \approx \frac{\ln^2 \hat{r}^{(I)}}{\tilde{\eta}_{-1}}, \qquad (19)$$

where the final approximation probably has a small relative error if $\mathbb{E}\tilde{\eta}_{-1}$ is large (i.e., $\tilde{\eta}_{-1} \approx \mathbb{E}\tilde{\eta}_{-1}$).

For comparison of our results with state-of-the-art methods, both the coalescent process and continuous-time branching process models of population growth produce the same approximation, that

$$a_m^{(C)} := \frac{M}{m(m-1)} \frac{1}{1-r^{-1}} \tag{20}$$

approximates a_m for $m = 2, 3, \ldots, M - 2$. To relate Eqs (15) and (20), recall the Taylor expansion $\ln r = -\ln[1-(1-r^{-1})] \approx 1 - r^{-1}$ near $1 - r^{-1} = 0$, i.e., near $r = 1$. The text near Eq (13) of [20] elaborates on the context of Eq (20). For comparison with Eq (18), which gives the estimator $\hat{r}^{(I)}$ in derived from Eq (15), consider the estimator $\hat{r}^{(C)}$ derived analogously from Eq (20),

$$1 - \left(\hat{r}^{(C)}\right)^{-1} = \frac{\mu M}{\sum_{m=2}^{M-2} \eta_m} \left(1 - \frac{1}{M-2}\right) = \frac{\mu M}{\tilde{\eta}_{-1}} \left(1 - \frac{1}{M-2}\right). \tag{21}$$

Routine algebra shows that the estimators are related by the equation

$$\hat{r}^{(C)} = \frac{1}{1 - \ln \hat{r}^{(I)}}. \tag{22}$$

Methods

Although the biological rationale given below in support of the Gamma model of HIV gp120 is brief, the interested reader can find a more detailed discussion elsewhere [20]. The Gamma model was simulated for each $r = R_0$, as follows. Each realization started with a single successfully infecting founder virus, which (for bookkeeping purposes) died at time $t = 0$, giving birth to a random number Z_1 of successfully infecting daughters. Biologically, each infected cell produces thousands of daughter virions, each with a small independent probability of infecting. The simulation therefore chose the number Z_1 from a Poisson distribution with mean $r = R_0$. Each daughter lived an independent random time after her birth. To approximate life-cycle times relevant to HIV [19,29], the random times had a gamma distribution with mean 2 days and standard deviation 0.24 days. The shape and rate parameters of the gamma distribution were therefore $(n, \lambda) = (69.4, 34.7)$ (to make the mean $n\lambda^{-1} = 2$ and variance $n\lambda^{-2} \approx 0.24^2$). The life-cycle of all the founder's descendants were similar, making the Gamma model a Bellman-Harris process [21] with parameters specific to HIV.

Each realization generated daughters until there were 6000 live viruses, to mimic a threshold of viral detection in blood. We also examined thresholds larger than 6000, but the exact threshold did not substantially alter scientific conclusions. If the viral population went extinct first, the realization restarted with a new founder. The 6000 live viruses were then sampled to produce six samples, of sizes $M = 2^k$ $(5 \le k \le 10)$. For each sample size M, tracing back the ancestry of the sample determined the ancestral sample frequency $\mathbf{A}$, which yielded the folded ancestral sample frequency $\tilde{\mathbf{A}}$.

In HIV, the gp120 gene is about 2550 nt long, and (with crossovers neglected) each HIV replication averages $\varepsilon \approx 2.16 \times 10^{-5}$ point mutations/base/replication [1]. On average, therefore, each RNA replication entails $\mu \approx 0.0551$ mutations in gp120. Simulating from the Poisson distribution in Eq (7) with $\mu \approx 0.0551$ yields a folded site-frequency spectrum $\tilde{\boldsymbol{\eta}}$ for the realization.

If $\tilde{\eta}_2 = \tilde{\eta}_3 = \ldots = \tilde{\eta}_{\tilde{M}} = 0$, the realization fails to estimate $\ln r$. For each r and each M, the simulation recorded the number F of failed realizations encountered before performing 1000 successful realizations.

For each of the 1000 successful realizations, $\tilde{\boldsymbol{\eta}}$ yielded $\ln \hat{r}$ for the estimate $\hat{r}^{(I)}$ in Eq (18). The simulation also estimated the corresponding standard deviations $\hat{\sigma}(\ln \hat{r})$ given in Eqs (19). For each r and each M, the simulation calculated a sample mean $\mathbb{E}(\ln \hat{r})$ and sample standard deviation $\sigma(\ln \hat{r})$ from the 1000 successfully sampled values of $\ln \hat{r}$. It also calculated the sample mean of the estimated standard deviation $\mathbb{E}\hat{\sigma}(\ln \hat{r})$ for comparison with the sample standard deviation.

We simulated ancestries as described above for a discrete grid of basic reproduction numbers $r = (1 + \varepsilon)^k$ $(k = 1, 2, \ldots, \lfloor \log_{1+\varepsilon} R \rfloor)$, so $1 < r \leq R$, with $R = 10$ and $\varepsilon = 0.1$ (as in [20]).

Results

Eq (18) suggests that $\sum_{m=2}^{M-2} \eta_m$ tends to decrease as r increases (the Introduction refers the reader to Fig 1 in [20] for an intuitive explanation). If $\eta_{-1} = \sum_{m=2}^{M-2} \eta_m = 0$ in Eq (18), $\hat{r}^{(I)} = \infty$ so one can only infer qualitatively that r is large. Thus, if every minority letter in an alignment column is a singleton, making $\eta_{-1} = 0$, a realization fails to estimate r quantitatively. Fig 1 displays the fraction of failed realizations (ones where $\tilde{\eta}_2 = \tilde{\eta}_3 = \ldots = \tilde{\eta}_{\tilde{M}} = 0$).

In Fig 1, the x-axis displays $x = \log_{10} r$ (where $r = R_0$); the y-axis, the fraction of failed realizations. The x-axis runs from $\log_{10} r = 0.0$ to $\log_{10} r = 1.0$, i.e., $r = 1$ to $r = 10$; the probabilities on the y-axis, from 0.0 to 1.0. The solid line corresponds to $M = 32$; the dashed line, to $M = 64$; the dashed-dot line, to $M = 128$; the dotted line, to $M = 256$. The fraction of failed realizations was identically 0.0 in all simulations with $M = 512$ and $M = 1024$.

In each subfigure of Fig 2 for $\hat{r} = \hat{r}^{(I)}$ (and also S1 and S2 Figs in the Supporting Information), the solid curves indicate the sample mean $y = \mathbb{E} \log_{10} \hat{r}$; the two dashed curves above and below each solid curve indicate $y = \mathbb{E} \log_{10} \hat{r} \pm \sigma(\log_{10} \hat{r})$, i.e., they indicate bands above and below $y = \mathbb{E} \log_{10} \hat{r}$ of height equal to the sample standard deviation $\sigma(\log_{10} \hat{r})$; and the

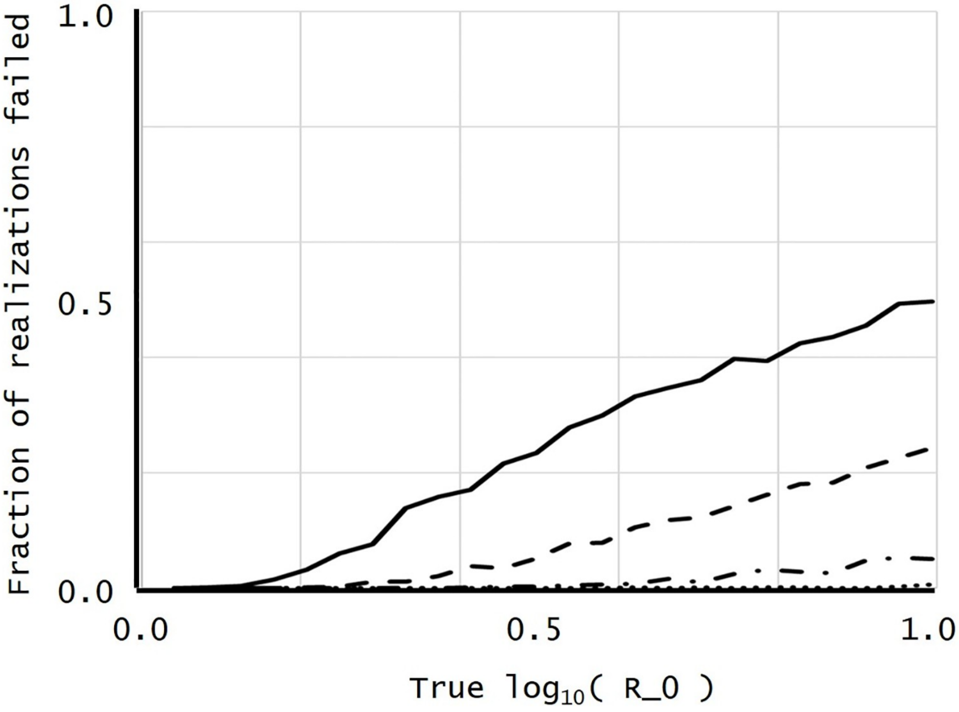

Fig 1. Plot of the fraction of realizations failed against the true $\log_{10}$ *r*.

https://doi.org/10.1371/journal.pone.0227127.g001

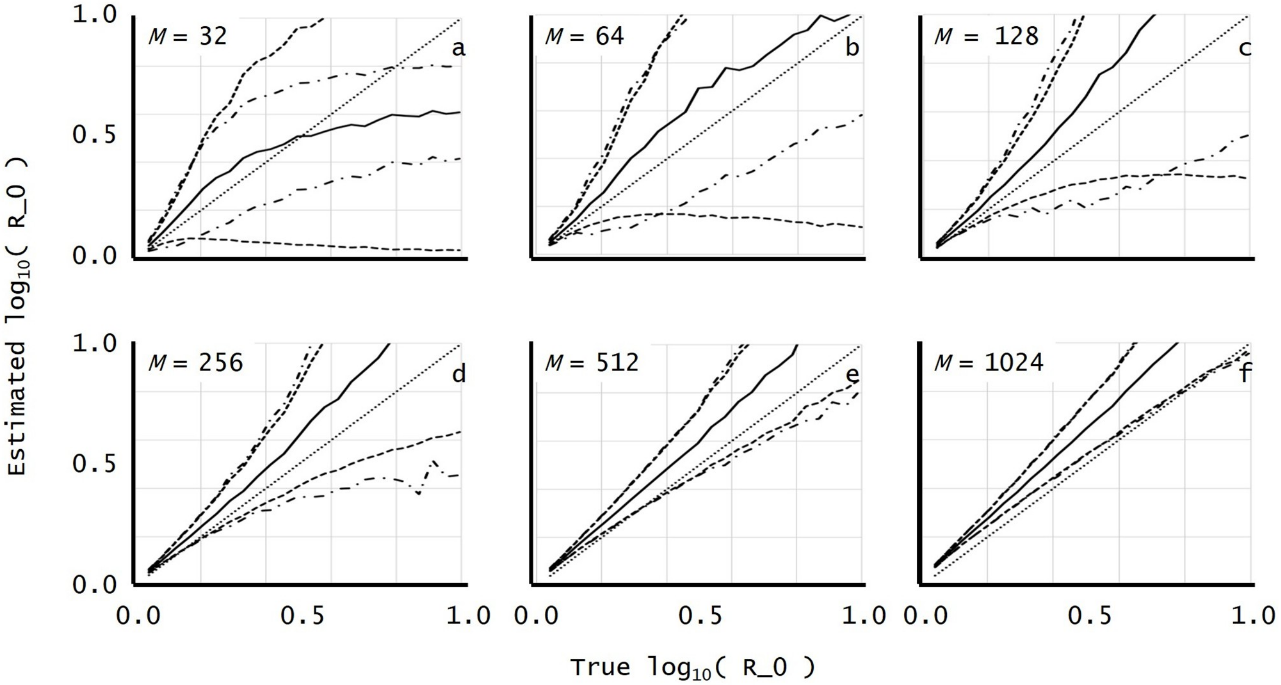

Fig 2. Plots for the maximum likelihood estimate (integral approximation $\hat{r}^{(I)}$).

https://doi.org/10.1371/journal.pone.0227127.g002

two dot-dashed curves above and below each solid curve indicate $y = \mathbb{E}\log_{10}\ \hat{r} \pm \mathbb{E}\hat{\sigma}(\log_{10}\hat{r})$, i.e., they indicate bands above and below $y = \mathbb{E}\log_{10}\hat{r}$ of height equal to the sample mean of the estimated standard deviation.

In Fig 2, both x- and y-axes display $\log_{10} r$: the horizontal, the true value $x = \log_{10} r$; the vertical, the estimated value $y = \log_{10}\hat{r}$. The x-and y-axes run from $\log_{10} r = 0.0$ to $\log_{10} r = 1.0$, i.e., $r = 1$ to $r = 10$. The dotted diagonal line indicates perfect estimation, $\hat{r} = r$. In their upper left, each of the subfigures (a)-(f) indicates the corresponding sample size $M = 2^k$ $(5 \leq k \leq 10)$.

Fig 2 for $\hat{r}^{(I)}$ displays progressively better recovery of r as M increases. For $M \geq 128$, the accompanying error estimate $\hat{\sigma}^2(\ln\hat{r}^{(I)})$ also has practical accuracy. Near $\log_{10} r = 0$, Fig 2e and 2f show some systematic overestimation away from the perfect estimate $\hat{r} = r$ (see also S1 and S2 Figs in the Supporting Information).

Fig 3 plots $y = \log_{10}\hat{r}^{(C)}$ from Eq (22) against $x = \log_{10}\hat{r}^{(I)}$. The dotted diagonal line indicates perfect agreement, $\log_{10}\hat{r}^{(C)} = \log_{10}\hat{r}^{(I)}$. In Fig 2, $y = \log_{10}\hat{r}^{(I)}$ slightly overestimated the true $x = \log_{10} r$. In Fig 3, $y = \log_{10}\hat{r}^{(C)}$ is consistently larger than $x = \log_{10}\hat{r}^{(I)}$. The two estimators $\hat{r}^{(C)}$ and $\hat{r}^{(I)}$ agree well as $\hat{r}^{(I)}$ decreases to 1, in accord with the Taylor expansion near $r = 1$ following Eq (20). (The Appendix of [20] also shows that in the present context, the Delta model of [20], the coalescent model, and the continuous-time branching-process model produce the same limiting SFS as $\boldsymbol{r}$ decreases to 1). Fig 3 also shows, however, that for larger values of $\hat{r}^{(I)}$, the coalescent estimate $\hat{r}^{(C)}$ becomes a gross overestimate, and it even blows up to infinity at $\log_{10}\hat{r}^{(I)} = \log_{10} e \approx 0.434$.

Discussion

In infection of a single host, the basic reproduction number R_0 is the expected number of cells infected by a single infected cell. This article shows how mutational variations observed in a

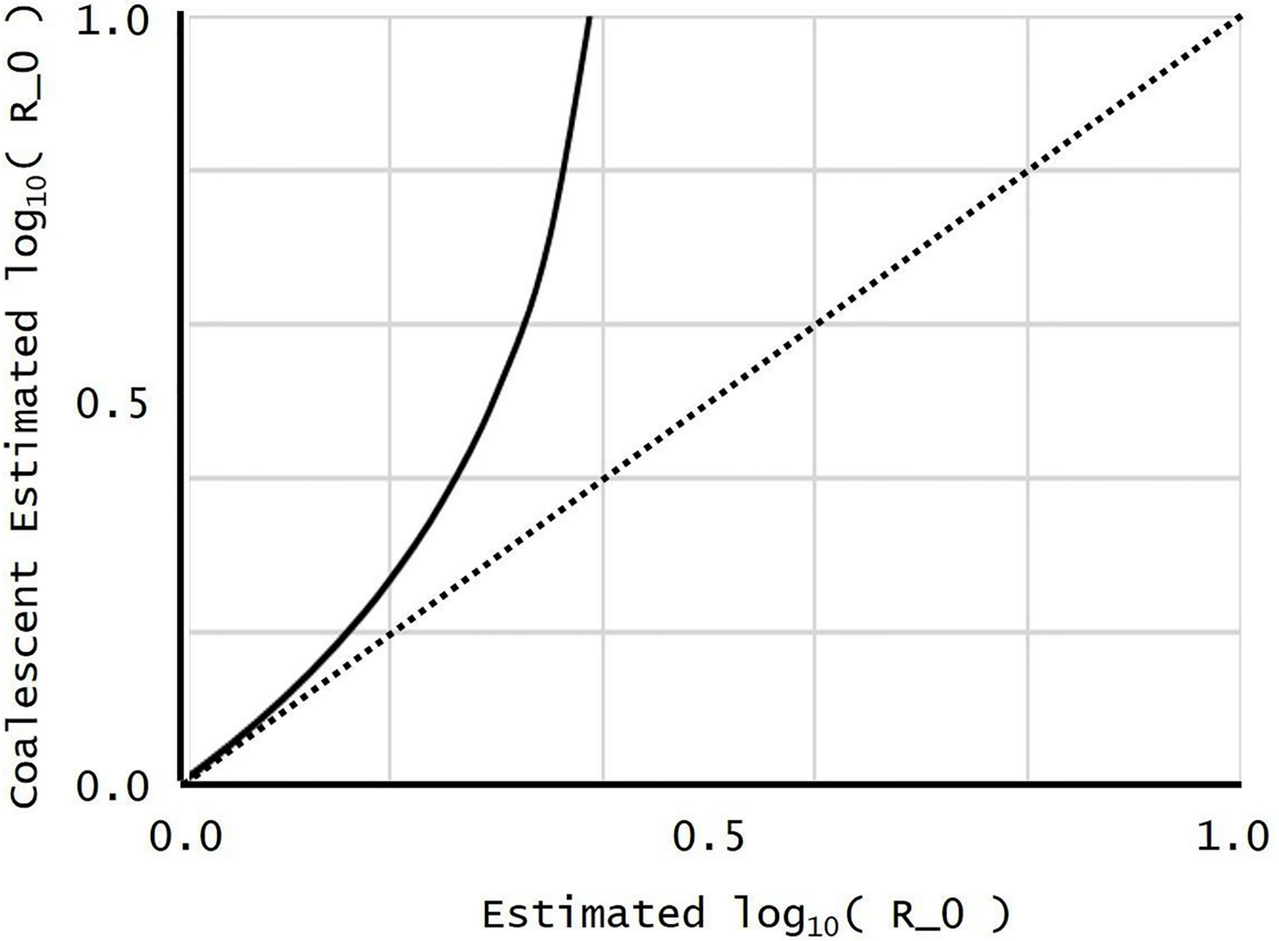

Fig 3. A plot of $\hat{r}^{(C)}$ against $\hat{r}^{(I)}$.

https://doi.org/10.1371/journal.pone.0227127.g003

sample from a population descended from a single founder yield an estimator $\hat{r}^{(I)}$ of R_0. It verifies the accuracy of its uncontrolled approximations with simulations that show that $\hat{r}^{(I)}$ reproduces R_0 within accuracies practicable for some purposes.

Our plots and statistics took log R_0 as the natural scale for the basic reproductive number R_0. The population size at generation g is $N_g = (R_0)^g$, so the population effects of changes in R_0 are linear on a log scale: log $N_g = g$ log R_0. On one hand, increasing R_0 by 1 has a greater effect on N_g when $R_0 = 1$ than when $R_0 = 10$. On the other, hand, doubling R_0 has the same additive effect on log N_g independent of the value of R_0.

Before considering the biological implications, we make a few technical statistical observations on the estimate itself. In simulations of HIV gp120 sequences, the absence of mutations common to sequence samples can cause estimation of R_0 to fail. For sample sizes $M \geq 128$, the probability of failing to estimate R_0 was less than 0.051 for all $1 < R_0 < 10$ (see Fig 1). To the authors' knowledge, studies sampling HIV sequences from patients typically sample between $M = 16$ and $M = 30$ [30] sequences per patient, an insufficient depth to test the present theory.

The estimator $\hat{r}^{(I)}$ in by Eq (18) has a simple analytic form, a negligible computation, that should be compared to the complexity of competing Bayesian calculations. Moreover, it does not contain derivatives that may lose accuracy because of numerical differencing. As noted after Eq (18), the coalescent and continuous-time birth-and-death branching process models of population growth yield estimators analogous to $\hat{r}^{(I)}$. In fact, the estimators are the same single estimator $\hat{r}^{(C)}$. The models can be manipulated to yield other estimators, so the comparison of models is by no means exhaustive, but in the present context $\hat{r}^{(C)}$ was clearly inferior to $\hat{r}^{(I)}$. It even displayed a singularity with our simulated HIV data (see Fig 3), suggesting that gradual reproduction in continuous time may have its limits when modeling the lytic viral bursts of HIV.

Like $\hat{r}^{(C)}$, but to a lesser extent, the estimator $\hat{r}^{(I)}$ overestimated R_0 throughout the full range tested, $1 < R_0 < 10$ (see Fig 2). Despite the overestimation, which became more severe as R_0 increased, estimates were accurate enough to be practicable for some purposes. As Fig 2 indicates, because of monotonicity of $\hat{r}^{(I)}$, experimental results with a large enough dataset can demonstrate a decrease in R_0. Moreover, near $R_0 = 1$, both the overestimation and the estimated error appeared relatively small, a useful property for detecting subtle therapeutic progress in reducing R_0. In summary, the integral estimator $\hat{r}^{(I)}$ is easy to compute throughout the full range $1 < R_0 < 10$ tested; its bias is noticeable not excessive for all sample sizes $M \geq 32$; and its error estimates are generally reliable for all sample sizes $M \geq 128$ (see Fig 2 for details).

The estimator $\ln \hat{r}^{(I)}$ in Eq (18) is linear in μ, suggesting that Eq (18) provides the first term of an expansion that our approximations have linearized around $\mu = 0$. Other articles [1,19,24] introduced and justified the linearizing approximation of an unvarying phylogeny, given here after Eq (1). The theory following Eq (1) indicates that the approximation steadily loses accuracy as μ increases, but simulations show that it retains enough accuracy to remain practicable in parameters ranges pertinent to HIV gp120.

We now turn to biological considerations. The chief limitation of our study derives from its attempt to use a simple mathematical model capture the complex biology of HIV infection. In fact, R_0 is likely to change as infection reaches new microenvironments in the host. Naïve use of the estimates here can only produce a single effective R_0 for early infection. In future (but beyond the purview of this article), we plan to incorporate time-dependencies into the simulation of R_0, and to develop estimates that can recover some of the time-dependency.

Presently, our mathematical model assumes a single founder. The extension of mathematical modeling from a single founder to multiple founders is an important relaxation of assumptions [19]. Regardless, the single-founder assumption is often satisfied in HIV infection, because most HIV infections have a single founder. For simplicity, it also assumes a constant R_0. After initial infection, HIV traverses different host microenvironments, potentially undergoing genetic bottlenecks. On one hand, a bottleneck can bias estimates based on sequence samples, because they obscure whether a most recent common ancestor dominates early in infection (a founder) or only after the bottleneck. In the present context, therefore, bottlenecks may bias estimates away from an initial R_0 and towards R_0 for a later microenvironment. On the other hand, even multiple escape lineages do not seriously bias genetic estimates of time since infection [24], so estimates of the initial R_0 may share a similar robustness.

Estimates of R_0 may also clarify HIV biology as an infection progresses. If target cells are scarce in a microenvironment, HIV may proliferate predominantly by cell-free spread, budding from an infected target cell, entering the extracellular fluid, and infecting another target cell by chance encounter [31]. Conversely, if target cell are plentiful, e.g., in the microenvironment of a lymph node [32,33], direct cell-to-cell spread may be more efficient than cell-free spread. Cell-to-cell spread has two distinct mechanisms (and therefore can occur in qualitatively distinct environments): (1) transmission of HIV by virological synapses between adjacent target cells or (2) transmission by capture and transfer of virions between proximal macrophages and dendritic cells [31]. Viral replication may not occur during cell-to-cell transmission, so regardless of the exact mechanism, shifts between cell-free spread and cell-to-cell spread may manifest themselves as concomitant changes in R_0. However fundamental R_0 may be to describing the reproduction of pathogens, cell-to-cell spread exemplifies the difficulties in interpreting an estimate of R_0 biologically.

The route of infection determines the initial microenvironment of HIV. Most routes transmit HIV much less effectively than hematological routes [7], suggesting that their initial R_0 is typically low. Mucosal infection can promote transmission of HIV [34], however, because it can increase the local concentration of activated T cells, promoting cell-to-cell spread, and probably increasing the initial R_0. Thus, the initial R_0 may provide valuable information about initial infection.

The dominance of cell-to-cell spread over cell-free spread may vary during infection. After infecting in a cell-rich mucosal microenvironment, HIV may move through the mucosal lamina, before being transported through the lymphatic system to lymph nodes, where the target cell density in its microenvironment increases dramatically [32,33]. The values of R_0 probably vary accordingly.

The present theory may also have an important application in animal trials of viral prophylaxis, when progress towards a therapy is subtle. Indeed, the design of animal trials using high-dose challenges may have unintentionally impeded practical assessment of candidate HIV therapies, because some vaccines and prophylactics may mitigate low- but not high-dose challenges [6]. Repeated low-dose challenge studies represent an important step forward in the pre-clinical assessment, because they mimic typical HIV challenges in humans [3]. Repeated low-dose challenges probably yield infections with a single founder virus, satisfying the primary assumption of the present theory. An estimate of R_0 in this context provides a new variable for statistical analysis, beyond the binary infection status of an animal, one with direct biological relevance to the establishment of infection. Even if a vaccine or prophylactic fails to produce total sterilizing immunity, a reduction in the initial R_0 encourages further investigation of an intervention, where previously the entire line of research might have been discarded.

Trials using repeated low-dose challenges also pose some unanswered experimental questions, the most pressing being the possibility that unsuccessful challenges potentially perturb the challenged animals. Do unsuccessful challenges foster partial immunity to further challenge? Do they increase the probability of future infection? Subtle perturbations in R_0 may be much more sensitive than a binary infection status in providing the answers to such questions.

Next-generation deep sequencing can produce around 10^5 reads of size comparable to gp120 [35,36]. Given the expectation that future experiments will likely be able to generate datasets with potentially even greater than 10^5 sequences, any robust estimator of R_0 must be able to handle extremely large sample sizes efficiently. The consistent tightening of the error estimates and accuracy as M is increased in Eq (19) suggests that the estimator $\hat{r}^{(I)}$ is particularly well-suited to application in such experiments. In addition, although a complete likelihood for the entire phylogeny and mutations might be desirable in some circumstances, a maximum likelihood method for the SFS permits inference about R_0 even if the reads are short, if the reads can be placed against a reference HIV genome.

In clinical trials of HIV therapies, guidelines previously suggested starting treatment only when CD4$^+$ T cell density declines below 350 cells/μl [37]. Recent studies (notably the SPARTAC trial) of temporally earlier interventions have shown increased efficacy compared to standard anti-retroviral protocols [38,39]. In some circumstances, the initial R_0 might measure the clinical efficacy of some such interventions, provide a variable predictive of their efficacy, or even help predict the clinical intervention with the greatest chance of success.

In conclusion, in the case of HIV and possibly other infectious agents, the integral approximation $\hat{r}^{(I)}$ provides a simple, easily computed estimate of the early basic reproduction number R_0 in a single host. The quantitative variable R_0 makes a well-characterized biological contribution to early HIV infection and should be useful assessing the efficacy of therapies in both human and animal trials.

Supporting information

S1 Fig. Plots for the maximum likelihood estimate (delta approximation) $\hat{r}^{(\Delta)}$.
(TIF)

S2 Fig. Plots for the estimate $\hat{r}^{(M)}$ from the method of moments.
(TIF)

S1 Data.
(DOCX)

Acknowledgments

We thank three anonymous referees for useful suggestions and Drs. Junyong Park, DoHwan Park, Anthony DeVico, and George Lewis for useful conversations.

Author Contributions

Conceptualization: John L. Spouge.

Formal analysis: Vruj Patel, John L. Spouge.

Methodology: John L. Spouge.

Software: Vruj Patel, John L. Spouge.

Supervision: John L. Spouge.

Validation: Vruj Patel, John L. Spouge.

Visualization: Vruj Patel.

Writing – original draft: Vruj Patel, John L. Spouge.

Writing – review & editing: Vruj Patel, John L. Spouge.

References

1. Giorgi EE, Funkhouser B, Athreya G, Perelson AS, Korber BT, et al. (2010) Estimating time since infection in early homogeneous HIV-1 samples using a poisson model. BMC Bioinformatics 11.
2. Koff WC, Johnson PR, Watkins DI, Burton DR, Lifson JD, et al. (2006) HIV vaccine design: insights from live attenuated SIV vaccines. Nature Immunology 7: 19–23. https://doi.org/10.1038/ni1296 PMID: 16357854
3. Nolen TL, Hudgens MG, Senb PK, Koch GG (2015) Analysis of repeated low-dose challenge studies. Statistics in Medicine 34: 1981–1992. https://doi.org/10.1002/sim.6462 PMID: 25752266
4. Gordon SN, Liyanage NPM, Doster MN, Vaccari M, Vargas-Inchaustegui DA, et al. (2016) Boosting of ALVAC-SIV Vaccine-Primed Macaques with the CD4-SIVgp120 Fusion Protein Elicits Antibodies to V2 Associated with a Decreased Risk of SIVmac251 Acquisition. Journal of Immunology 197: 2726–2737.
5. Strbo N, Vaccari M, Pahwa S, Kolber MA, Doster MN, et al. (2013) Cutting Edge: Novel Vaccination Modality Provides Significant Protection against Mucosal Infection by Highly Pathogenic Simian Immunodeficiency Virus. Journal of Immunology 190: 2495–2499.

6. Regoes RR, Longini IM, Feinberg MB, Staprans SI (2005) Preclinical assessment of HIV vaccines and microbicides by repeated low-dose virus challenges. Plos Medicine 2: 798–807.

7. Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A, et al. (2014) Estimating per-act HIV transmission risk: a systematic review. Aids 28: 1509–1519. https://doi.org/10.1097/QAD.0000000000000298 PMID: 24809629

8. Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, et al. (2003) Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. Aids 17: 1871–1879. https://doi.org/10.1097/00002030-200309050-00005 PMID: 12960819

9. Kahn JO, Walker BD (1998) Acute human immunodeficiency virus type 1 infection. New England Journal of Medicine 339: 33–39. https://doi.org/10.1056/NEJM199807023390107 PMID: 9647878

10. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, et al. (2008) Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. Journal of Virology 82: 3952–3970. https://doi.org/10.1128/JVI.02660-07 PMID: 18256145

11. Stafford MA, Corey L, Cao YZ, Daar ES, Ho DD, et al. (2000) Modeling plasma virus concentration during primary HIV infection. Journal of Theoretical Biology 203: 285–301. https://doi.org/10.1006/jtbi.2000.1076 PMID: 10716909

12. Ribeiro RM, Qin L, Chavez LL, Li DF, Self SG, et al. (2010) Estimation of the Initial Viral Growth Rate and Basic Reproductive Number during Acute HIV-1 Infection. Journal of Virology 84: 6096–6102. https://doi.org/10.1128/JVI.00127-10 PMID: 20357090

13. Kosaka PM, Pini V, Calleja M, Tamayo J (2017) Ultrasensitive detection of HIV-1 p24 antigen by a hybrid nanomechanical-optoplasmonic platform with potential for detecting HIV-1 at first week after infection. Plos One 12.

14. Wolinsky SM, Wike CM, Korber BTM, Hutto C, Parks WP, et al. (1992) Selective Transmission of Human-Immunodeficiency-Virus Type-1 Variants from Mothers to Infants. Science 255: 1134–1137. https://doi.org/10.1126/science.1546316 PMID: 1546316

15. Delwart E, Magierowska M, Royz M, Foley B, Peddada L, et al. (2002) Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection. Aids 16: 189–195. https://doi.org/10.1097/00002030-200201250-00007 PMID: 11807302

16. Derdeyn CA, Decker JM, Bibollet-Ruche F, Mokili JL, Muldoon M, et al. (2004) Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. Science 303: 2019–2022. https://doi.org/10.1126/science.1093137 PMID: 15044802

17. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, et al. (2008) Identification and characterisation of transmitted and early founder virus envelopes in primary HIV-1 infection. Proceedings of the National Academy of Sciences of the United States of America 105: 7552–7557. https://doi.org/10.1073/pnas.0802203105 PMID: 18490657

18. Haaland RE, Hawkins PA, Salazar-Gonzalez J, Johnson A, Tichacek A, et al. (2009) Inflammatory Genital Infections Mitigate a Severe Genetic Bottleneck in Heterosexual Transmission of Subtype A and C HIV-1. Plos Pathogens 5.

19. Love TMT, Park SY, Giorgi EE, Mack WJ, Perelson AS, et al. (2016) SPMM: estimating infection duration of multivariant HIV-1 infections. Bioinformatics 32: 1308–1315. https://doi.org/10.1093/bioinformatics/btv749 PMID: 26722117

20. Spouge JL (2019) An accurate approximation for the expected site frequency spectrum in a Galton-Watson process under an infinite sites mutation model. Theor Popul Biol 12: 30151–30155.

21. Bellman R, Harris TE (1948) On the Theory of Age-Dependent Stochastic Branching Processes. Proceedings of the National Academy of Sciences of the United States of America 34: 601–604. https://doi.org/10.1073/pnas.34.12.601 PMID: 16588841

22. Knuth DE (1992) 2 Notes on Notation. American Mathematical Monthly 99: 403–422.

23. Kimura M (1969) Number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903. PMID: 5364968

24. Park SY, Love TMT, Perelson AS, Mack WJ, Lee HY (2016) Molecular clock of HIV-1 envelope genes under early immune selection. Retrovirology 13.

25. Fu YX (1995) Statistical properties of segregating sites. Theoretical Population Biology 48: 172–197. https://doi.org/10.1006/tpbi.1995.1025 PMID: 7482370

26. Athreya KB, Ney PE (2004) Branching Processes. Mineola, New York: Dover.

27. Wakeley J (2009) Coalescent Theory. Greenwood Village CO: Roberts and Company.

28. Graham RL, Knuth DE, Ptashnik O (1994) Concrete mathematics: a foundation for computer science. New York: Addison-Wesley.

29. Markowitz M, Louie M, Hurley A, Sun E, Di Mascio M (2003) A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-Cell decay in vivo. Journal of Virology 77: 5037–5038. https://doi.org/10.1128/JVI.77.8.5037-5038.2003 PMID: 12663814

30. Lee HY, Giorgi EE, Keele BF, Gaschen B, Athreya GS, et al. (2009) Modeling sequence evolution in acute HIV-1 infection. Journal of Theoretical Biology 261: 341–360. https://doi.org/10.1016/j.jtbi.2009.07.038 PMID: 19660475

31. Zhang CW, Zhou S, Groppelli E, Pellegrino P, Williams I, et al. (2015) Hybrid Spreading Mechanisms and T Cell Activation Shape the Dynamics of HIV-1 Infection. Plos Computational Biology 11.

32. Layne SP, Merges MJ, Dembo M, Spouge JL, Nara PL (1990) {HIV} requires multiple gp120 molecules for {CD4}-mediated infection. Nature 346: 277–279. https://doi.org/10.1038/346277a0 PMID: 2374593

33. Spouge JL (1994) Viral multiplicity of attachment and its implications for human-immunodeficiency-virus therapies. Journal of Virology 68: 1782–1789. PMID: 8107240

34. Ward H, Rönn M (2010) The contribution of STIs to the sexual transmission of HIV. Current opinion in HIV and AIDS 5: 305–310. https://doi.org/10.1097/COH.0b013e32833a8844 PMID: 20543605

35. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, et al. (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 13: 375. https://doi.org/10.1186/1471-2164-13-375 PMID: 22863213

36. McElroy K, Thomas T, Luciani F (2014) Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microbial Informatics and Experimentation 4: 1. https://doi.org/10.1186/2042-5783-4-1 PMID: 24428920

37. Moore RD, Keruly JC (2007) CD4(+) cell count 6 years after commencement of highly active antiretroviral therapy in persons with sustained virologic suppression. Clinical Infectious Diseases 44: 441–446. https://doi.org/10.1086/510746 PMID: 17205456

38. Fidler (2013) Short-Course Antiretroviral Therapy in Primary HIV Infection. The New England journal of medicine 368: 207–217. https://doi.org/10.1056/NEJMoa1110039 PMID: 23323897

39. Sáez-Cirión A, Bacchus C, Hocqueloux L, Avettand-Fenoel V, Girault I, et al. (2013) Post-Treatment HIV-1 Controllers with a Long-Term Virological Remission after the Interruption of Early Initiated Antiretroviral Therapy ANRS VISCONTI Study. PLOS Pathogens 9: e1003211. https://doi.org/10.1371/journal.ppat.1003211 PMID: 23516360