Estimation of the Basic Reproduction Number in a Galton-Watson Process under an Infinite Sites Mutation Model

John L. Spouge1\*

1National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

\*Corresponding author

E-mail: [spouge@ncbi.nlm.nih.gov](mailto:manzouro@ncbi.nlm.nih.gov)

Keywords: approximate likelihood; Galton-Watson process; viral reproduction number

# Abstract

If viruses or other pathogens infect a single host, the outcome of infection often hinges on the fate of the initial invaders. The initial basic reproduction number , the expected number of cells infected by a single infected cell, helps determine whether the initial viruses can establish a successful beachhead. To determine , the Kingman coalescent or continuous-time birth-and-death process can be used to infer the rate of exponential growth in an historical population. Given  sequences sampled in the present, the two models can make the inference from the site frequency spectrum (SFS), the count of mutations that appear in exactly  sequences (). In the case of viruses, however, if  is large and an infected cell bursts while propagating virus, the two models are suspect, because they are Markovian with only binary branching. Accordingly, this article develops an approximation for the SFS of a discrete-time branching process with synchronous generations (i.e., a Galton-Watson process). When evaluated in simulations with an asynchronous, non-Markovian model (a Bellman-Harris process) mimicking the bursting viral reproduction of HIV, the approximation proved superior to approximations derived from the Kingman coalescent or continuous-time birth-and-death process. This article demonstrates that in analogy to methods in human genetics, the SFS of viral sequences sampled well after latent infection can remain informative about the initial . Thus, it suggests the utility of analyzing the SFS of sequences derived from patient and animal trials of viral therapies, because in some cases, the initial  may be able to indicate subtle therapeutic progress, even in the absence of statistically significant differences in the infection of treatment and control groups.

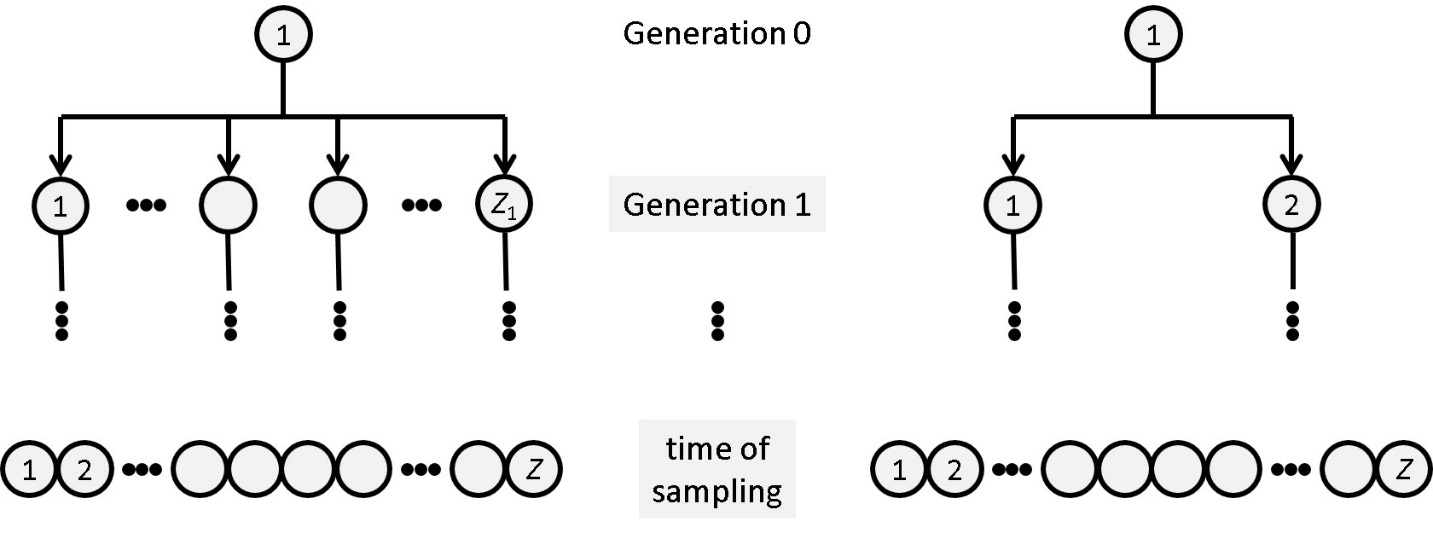
# Introduction

This article implements the second step in a scientific program initiated in a previous article ([Spouge 2018](#_ENREF_36)), referred to here as “I”. Before giving a self-contained description of the second step (including definitions for any technical terms and standard notations used in this paragraph), an outline of the three steps follows. Consider a discrete-time (Galton-Watson) branching process evolving under an infinite-sites model of mutation, with attention to its basic reproduction number , the mean number of daughters per mother. As a first step, I gave a probabilistic approximation describing sequence patterns generated by the process. As a second step, the present article inverts the description to estimate . The third step is to estimate  from sequence patterns if several similar Galton-Watson processes start simultaneously.

Our practical motivating application requires the completion of all three steps, as follows. Consider a pathogen infecting a single host. Our paradigmatic application is human immunodeficiency virus (HIV) infecting a human host, although other applications might include animal trials using simian immunodeficiency virus (SIV) or potentially, any combination of host and pathogen. Successful infection often hinges on the fate of the first arrivals among the invaders. In the viral infection of a single host, e.g., the invasive population often descends from a small set of founder viruses. It may even descend from a single founder, e.g., about 80% of HIV infections have a single founder ([Keele et al. 2008](#_ENREF_22), [Haaland et al. 2009](#_ENREF_18), [Love et al. 2016](#_ENREF_27)). Conversely, 20% of HIV infections have multiple founders, so a complete practical theory must include the case of multiple founders.

I explains our motivation further. “The initial basic reproduction number , in viral infection the expected number of cells that a single infected cell infects in the next generation ([Giorgi et al. 2010](#_ENREF_15)), contributes fundamentally to the chances of an invasive population establishing a successful beachhead. On one hand, if , the invasive population falls below replacement and dies. On the other hand, if  is slightly greater than 1, the invaders have a small but positive chance of survival, and if  is large, the invaders are likely to flourish. Thus, whether preventing infection or mitigating its impact by reducing an initial viral load, the initial , the basic reproduction number at the start of infection, could in principle provide a measure for setting therapeutic goals and benchmarking therapeutic progress.” I also gives a closely reasoned argument that because current methods for estimating  in HIV infection (e.g.,  ([Stafford et al. 2000](#_ENREF_38)) or  ([Ribeiro et al. 2010](#_ENREF_34))) require contemporaneous detection of virus, they pertain many generations after the viral founders, but not during the initial infection or to the founders themselves .

Although detection thresholds prevent direct measurement of the initial , genetic researchers have shown that the nucleic acids of present-day humans retain footprints from the population history (e.g., ([Durrett and Limic 2001](#_ENREF_11), [Durrett 2008](#_ENREF_9))). Similarly, viral sequences sampled during early viremia may be informative about the initial . Figure 1 illustrates the concept, and the remainder of this article refines the qualitative insight there. In particular, Figure 1 suggests that mutations appearing in two or more sampled sequences contain information about the initial  of an expanding population. Although the viral applications motivate the analysis, the theory presented here has broad applicability to inferring the early demographics of populations with very few founders.



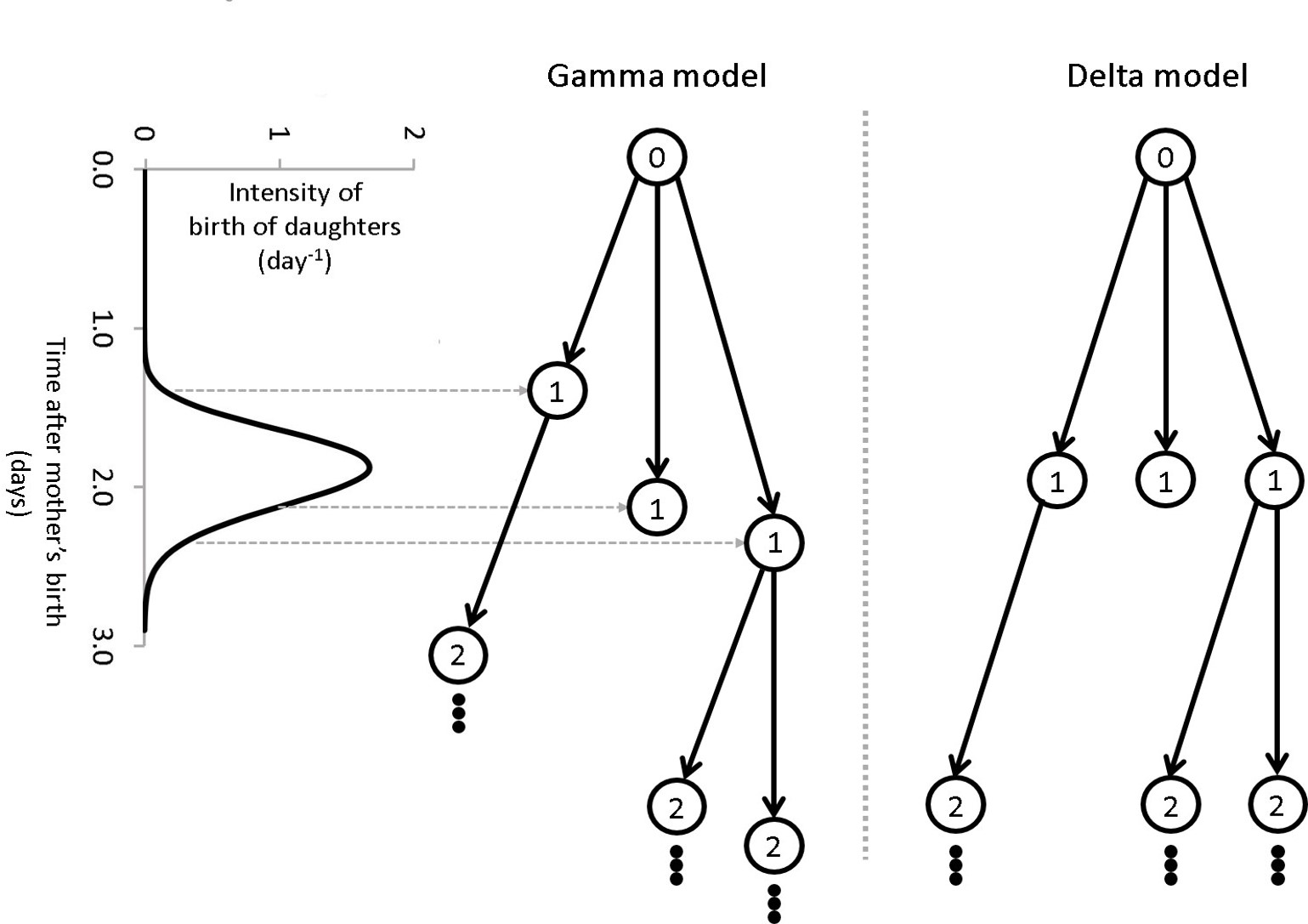
**Figure 1: Two viral ancestries, illustrating the effect of the initial**  **on samples.**

Figure 1 illustrates two hypothetical viral ancestries. The circles represent viruses. In the two ancestries, time runs downward, and for simplicity each viral population at the bottom has a single founder at the top (multiple founders introduce technical complications, but do not change the concept illustrated). The present population at the bottom contains some large number  of individuals, from which  sequences are sampled. The ancestry on the left has a large initial , so the founder has  daughters, where for illustration, we assume  is much larger than . Thus, the  sampled sequences are likely to descend from different daughters (many standard references on the “Birthday problem” implicitly prove this statement). If so, the sampled sequences cannot share any mutations away from the founder sequence. In contrast, the ancestry on the right has a small initial , so the founder has only  daughters. If either daughter’s sequence has a mutation away from the founder sequence, about half the  sampled sequences share the mutation. Thus, mutations appearing in more than one of the  sampled sequences are informative about the initial . Sampled mutations are conveniently partitioned into a “site frequency spectrum”, i.e., the numbers  of the sites where a mutation appears in exactly  of the  sampled sequences ().

Thus, given  aligned viral sequences sampled simultaneously from a single infected host, the aim of this article is to reconstruct as a function of  the site frequency spectrum (SFS) described in Figure 1. Before proceeding to the mathematical abstractions of Section 2, we first establish the parameter ranges relevant to an important application. In studies sampling HIV gp120 sequences from patients,  typically ranged from 16 to 30 per patient ([Lee et al. 2009](#_ENREF_26)). The gp120 gene is about 2550 nt long, and (with crossovers neglected) each HIV replication averages  point mutations/base/replication ([Giorgi et al. 2010](#_ENREF_15)). On average, therefore, each RNA replication entails  mutations in gp120. (The three significant figures represent an unrealistic precision but retain consistency with the literature.)

Studies in human genetics suggest some methods for estimating  from the SFS in gp120 sequence data, but the methods are not directly suited to viral infection in a single host. To elaborate, consider the following idealized model of an initial HIV infection, focusing first on a typical initial infected cell in a host. HIV lyses the cell, releasing about  viral particles ([Chen et al. 2007](#_ENREF_7), [De Boer et al. 2010](#_ENREF_8)). As noted above, the initial  is likely smaller than , so a typical viral particle has a miniscule chance of infecting. Given an initial environment, if viral particles infect independently, the count of infected daughter cells approximately follows an Poisson distribution whose mean equals the unknown initial  (see, e.g., ([Arratia et al. 1989a](#_ENREF_1), [Arratia et al. 1989b](#_ENREF_2))). Thus, the viruses from the initial infected cell infect  daughter cells, where  is a Poisson random variate with mean . Estimates of the replication time for HIV range from 1.76 days to 4.2 days ([Love et al. 2016](#_ENREF_27)), 2 days being a reasonable approximation ([Markowitz et al. 2003](#_ENREF_28)). We therefore model the time-intervals between the lysis of a mother cell and her infected daughter cells as independent random variates with a gamma distribution, approximating the mean by 2 days and the standard deviation by 0.24 = 2 – 1.76 days (1.76 days is the minimum estimate of the HIV replication time, so the resulting gamma distribution is likely less tightly concentrated around its mean than the true random HIV replication time). The gamma distribution approximating the random HIV replication time therefore has shape and rate parameters  (i.e., its mean  and variance ). The reproductive cycle begins anew with the lysis of each infected daughter. This idealized but relatively realistic model of HIV reproduction is a Bellman-Harris process ([Bellman and Harris 1948](#_ENREF_4)); call it the “Gamma model”, to emphasize that its parameters and distribution are chosen to model HIV (see Figure 2).

The theory in Section 2 exploits the Delta model illustrated in Figure 2 (a Galton-Watson branching process, in discrete-time) to provide analytic approximations to the SFS of the Gamma model. To make the Delta model comparable to the Gamma model, the count of daughters in the Delta model is also a Poisson variate with mean , and the deterministic replication time in the Delta model is 2 days, the mean replication time in the Gamma model. As the models’ names suggest, the synchronous generations in the Delta model therefore substitute a Dirac delta distribution (i.e., a distribution having a single atom of probability 1, a distribution as tightly concentrated around its mean as possible) for the gamma distribution in the Gamma model. For comparison with simulations of the Gamma model, a birth-and-death process with constant birth and death intensities and the Kingman coalescent model with an exponentially expanding population also provide analytic expressions for the SFS. All approximations in this article are uncontrolled, so (often without comment) we rely on the simulations of the Gamma model with parameters relevant to HIV gp120 to assess the accuracy of SFS approximations. Other parameters regimes, which require separate assessment, are beyond the immediate practical purview of this article.



**Figure 2: A diagram schematically illustrating the Gamma and Delta models.**

Figure 2 illustrates two hypothetical models of viral ancestries. As in Figure 1, black circles represent the lysis of infected cells, and each ancestry starts from a single founder at the top, with time running downward. The numbers in the circles count the generations away from the founder, with the single founder in generation 0. In the Gamma model on the left, the lysis of each infected cell gives birth to a random number of viral daughters whose progeny lyse cells at random times chosen independently from a gamma distribution. The graph on the extreme left illustrates the relevant gamma distribution. The Delta model on the right is like the Gamma model, except that each cell lysis gives birth to viral daughters who lyse their cells simultaneously, in synchronous generations at time-intervals equal to the mean generation time in the Gamma model.

The organization of this article is as follows. Section 2 (Theory) approximates the mean SFS in the Delta model as a function of the basic reproduction number  and compares it to the SFS derived from the Kingman coalescent for an expanding population or from continuous-time birth-and-death processes. Section 3 (Methods) describes simulation of the Gamma and Delta models. It also discusses the dependence of the SFS variance on the mutation rate. Section 4 (Results) examines the accuracy of the various approximations when the simulations use the parameters relevant to HIV gp120. Section 5 is our Discussion, which examines the few (superable) difficulties impeding direct data analysis with the theory presented here.

[TPB\_Spouge.docx](file:///C:\Users\spouge\AppData\Roaming\Microsoft\Reproductive_Number\Reproductive_Number\Paper\TPB_Spouge.docx) (referred to hereafter as “I”) establishes the basic notation and set-up. I assumed an infinite sites model.

# Theory

Let “” denote a definition, and . Fix an alphabet , e.g., the unambiguous nucleotide alphabet . In practice, sequence analysis must invoke a strategy for handling anomalous characters (e.g., ambiguous nucleotides or gap characters) in an alignment. Sometimes, anomalous characters are infrequent, so that as an acceptable approximation, the analysis treats them as ordinary characters. Sometimes, the analysis simply omits columns containing them. Without further comment, the following assumes that the analysis has adopted an unspecified strategy for handling anomalous characters.

Consider a set  consisting of  sequences. The sequences are sampled simultaneously from the descendants of a single founder sequence . Align the sequences , so that the sequences  form an alignment matrix  of  columns, the data underlying the following analysis. Implicitly,  aligns with , so the letter  in  is ancestral to each letter  in the matrix column  (from the sequence ). In practice,  is usually unknown, a complication we handle shortly.

The Iverson bracket for indicator random variates is a standard notation ([Knuth 1992](#_ENREF_24)): let  if the statement  is true, and  otherwise. (Context disambiguates the indicator from the corresponding event.) Let  () count the instances of letter  in column  of , so .

The difference  counts letters mutated away from the founder  in column ; let . Given , let  count the alignment columns  where  letters differ from the founder letter , with the site frequency spectrum (SFS) defined as . Typically,  is unknown, so  is not observable.

To develop a statistical model for , let  be the probability of mutation per base per generation in column  of the alignment. In an infinite-sites model ([Kimura 1969](#_ENREF_23)), we neglect as extremely rare the possibility that two or more mutations occur in the ancestry of a single letter . Let  be the expected number of novel mutations per generation in the sequences. If  are all small, and novel mutations independent, the novel mutation counts in every daughter of every mother are approximately independent Poisson variates with fixed mean .

Let  count the non-founder ancestors with  descendants in the sample, with the ancestral sample frequency spectrum (AFS) defined as . Under an infinite-sites model, every novel mutation occurs in a different column of the alignment. Every alignment column with  mutations therefore corresponds to a novel mutation in an ancestor with  descendants in the sample. Given , Theorem 1 in I shows that the coordinates of  are independent Poisson variates, with  having mean , i.e., in symbols , where “” indicates equality of distributions.

Eq (16) in I showed that

.

Simulations in I showed that for  (HIV gp120), the typical magnitude of the ratio  was at most about 18%. Let  and . Typically, the mutational variance  makes the dominant contribution to , i.e., in the context of  with  given,  effectively varies very little, so the distributional approximation  pertains. To summarize,

.

I showed mathematically that in any model with synchronous generations,

,

i.e., counting the (non-founder) ancestors in each of  generations, accounting for the multiplicity of their sampled descendants, counts each of the  samples  times. Moreover, I showed that the Delta model (which has synchronous generations) yielded an accurate approximation



for  () in the Gamma model. If , Eq and the expectation of Eq therefore show that in the Gamma model,  is approximately the finite quantity

.

In statistical notations, “” often suggests a sum, as in . Eq motivates the analogous definition .

To relate  to an observable, define , so  counts minority letters in column  (i.e.,  counts letters differing from a letter with the maximum count). Unlike , the observable  has no dependency on the founder sequence . Let  denote the floor function and . Define the folded SFS , with



for ), where the second equality holds for an infinite-sites model (which we have assumed). If  is odd, , so the definition of  is complete. If  is even, the pattern of pairs  displayed in Eq fails for , because  cannot be paired with a distinct . If  is even, therefore, define  for . Loosely,  counts columns  where the number of minority letters equals  ().

The folded AFS  inherits the pattern for  established in Eq :  for ; in addition, if  is even, . Henceforth and without comment, the same pattern generates folded quantities (denoted by over-tildes) from unfolded quantities, e.g., . The folded SFS  inherits its approximate Poisson distribution from the SFS :

.

Because

,

the pattern suggests imposing the definitions

,

, etc., on all folded quantities with a “” subscript.

Eq yields a maximum likelihood estimate (MLE) . The desirable asymptotic properties of an MLE make  useful for benchmarking statistical estimates of . Recall Eqs and relating  and , and take natural logarithms in Eq to derive the (approximate) log-likelihood

,

where “” indicates an equality of functions, possibly ignoring an irrelevant additive term depending only on data (e.g., a function of ).

The following treats  as a continuous variate. Now, for any value of  (not just ),

,

if the argument  of the maximum satisfies , i.e.,  is determined by setting the derivative with respect to  equal to 0. Thus, an MLE  from Eq maximizes the profile log-likelihood

.

Now,  for . Because  at the boundaries of the interval , a MLE  therefore exists, such that



has a root at . The observed Fisher information

,

so

,

where the first approximation follows from a delta approximation ; the second approximation is a standard use of Fisher information; and the final equality derives from Eq . Note that calculation of the folded AFS  () in Eq only requires  for .

Eq for  provides approximations for  and its derivatives, ultimately yielding an approximate MLE , derived from the Delta model and the delta method. The approximate MLE  maximizes Eq , with its  given by Eq , with  replacing in both equations. Inconveniently, however, the infinite sum defining  holds numerical complications. For , however, . Thus, the sum in  can be approximated by an integral (the first term of an Euler-Maclaurin series):

,

where the final equality is the evaluation of a beta integral. Comparison of the two sides of Eq shows that



approximates  for . After substituting  for  in Eq and unfolding all folded quantities,

.

Telescoping cancellation yields

.

To estimate , define  (an observable, the sum of all  except ), so the root  of Eq satisfies

.

The variate  is approximately Poisson distributed, so . The linear Taylor series  with  approximates the variance

.

The simplicity of Eq suggests comparing the approximate MLEs  and  to an estimate  derived from applying the method of moments to . To begin, define

.

The moment

.

In a standard notation, let  denote the inverse of a function  under functional composition (denoted by “”). In the method of moments, the basic reproduction number  is estimated from (the observable) . Logarithms  provide a convenient scale for , so with ,

.

Under the assumption that  (i.e.,  has a small bias and coefficient of variation), the variance of Eq is

,

where we recall that , so .

# Methods

Biology puts some practical upper bound  on the basic reproduction number: let . In addition, biological noise (e.g., unknown changing environments as HIV propagates within a single host) puts some practical lower error bound  on any estimate . Accordingly, we simulated phylogenies for a discrete grid of basic reproduction numbers  (), choosing  and , bounds that are generous for HIV.

For each  in the grid, simulation yielded several realizations of the Gamma model. As in I, each realization started with one founder and propagated it until the population reached a sampling threshold of 6000 live individuals. Real experiments require sampled individuals, so if extinction occurred before reaching the sampling threshold, the realization restarted with another founder. After the population reached the sampling threshold, the realization then sampled  individuals uniformly without replacement for . For each , it computed the AFS  and simulated the exact conditional SFS distribution  to derive the folded SFS .

If , the realization fails to yield any estimate of . For each  and each , the simulation recorded the number  of failed realizations encountered before performing 1000 successful realizations.

For each of the 1000 successful realizations,  yielded  for the three types of estimate  in Eqs , , and . The simulation also estimated the corresponding standard deviations  given in Eqs , , and . For each , each , and for each of the three estimates , therefore, the simulation calculated a sample mean  and (unbiased) sample standard deviation  from the 1000 successfully sampled values of . It also calculated the sample mean of the estimated standard deviation  for comparison with the sample standard deviation.

# C:\Users\spouge\Documents\Store\Projects\Virus\Infecting Virions\Founder_Single\Reproductive_Number\Reproductive_Number\Paper\Figures\SFS_vs_R0.jpgResults

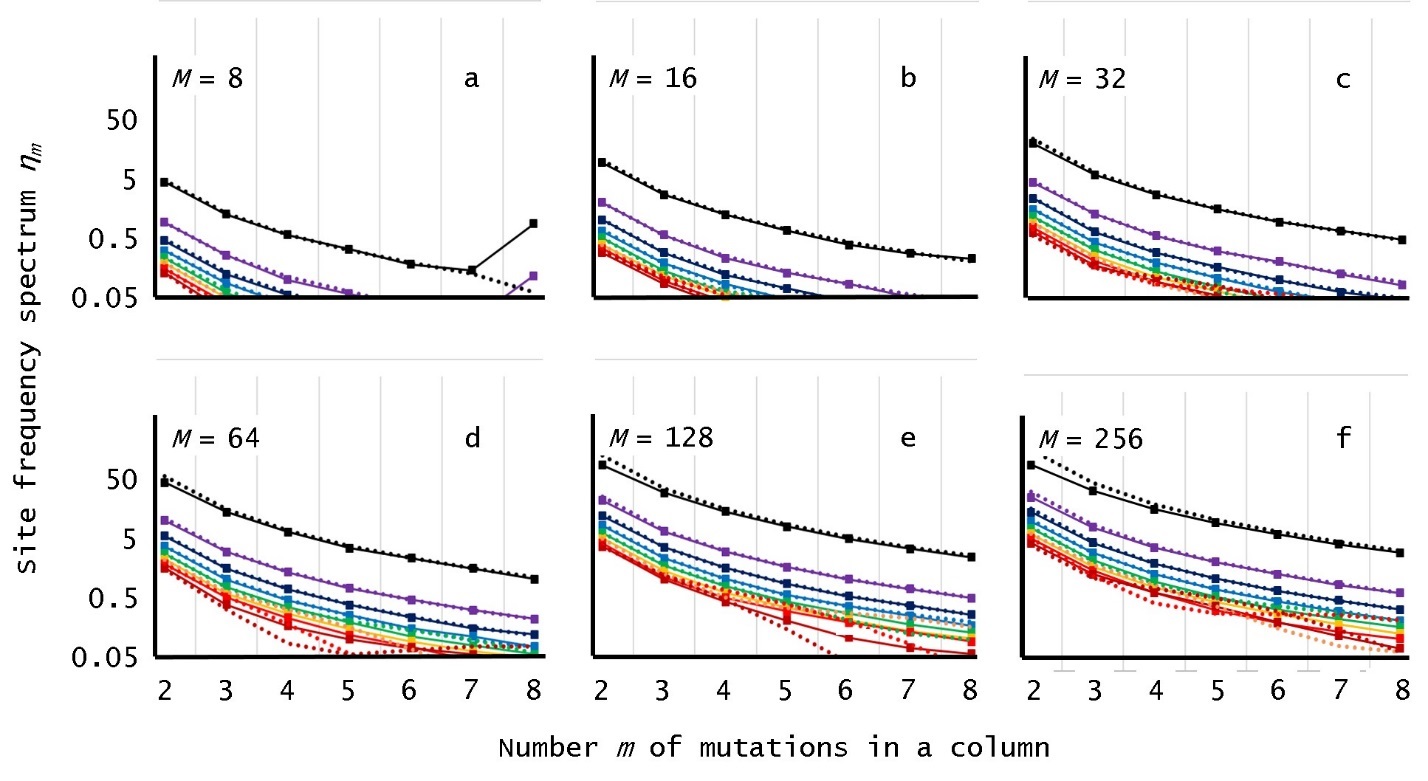
**Figure 3. Selected plots of the SFS  vs. the basic reproduction number**

In Figure 3, all axes are logarithmic. All X-axes share the same scale (likewise for all Y-axes). All results displayed use the HIV value . In each subfigure, Eq for the coalescent SFS  yields the black curve joining triangular points. Because the black curves are therefore all translates of , the translated shape provides a ready reference for comparing subfigures. In each subfigure, Eq for the Delta SFS  yields the red curve joining square points. Generally, each red curve obscures a gold curve joining circular points. The gold curve corresponds to the sample mean  estimating  from simulations of the Gamma model, with the error bars giving the sample standard deviation. Figure 3a, b, c, and d show plots pertinent to  (i.e.,  for ) for different sample numbers , whereas Figure 3d, e, and f show plots pertinent to , , and  for . For  in Figure 3d, e, and f, close inspection of the leftmost point (at ) displays  crossing over from  to  as  decreases to 1, as in Eq . Figure 3c also shows  crossing over from  to , albeit more subtly, for .

Typically, the sample means and sample standard deviations of the SFS simulated from the Delta and Gamma models were visually indistinguishable, and the differences between the models’ results were always subtle. Using the Gamma model, Figure 3 exemplifies some other numerical trends. Typically for fixed , the sample mean SFS  simulated from the Gamma model increased with the sample number , while the sample standard deviation decreased (see plots pertinent to  in Figure 3a, b, c, and d for ), whereas for fixed ,  decreased with the number  of mutations in an alignment column, while the sample standard deviation increased (see plots pertinent to , , and  in Figure 3d, e, and f for ). Unlike Eq for the coalescent SFS , Eq for the Delta SFS  generally provided accurate approximations to the sample mean  of the Gamma SFS in all simulations, except possibly where it crossed over from  to  as , e.g., at  for large  (e.g., Figure 3c, d, e, and f, where  and ); or for small  for  and  (in Figure 3c).

Figure 3 is representative of simulation results, with two exceptions. First, the inequality  failed occasionally at large values of , with some neighboring values of  yielding comparable but probabilistically independent numerical exceptions. Second, for fixed , , , and  all typically decreased with increasing  (as in Figure 3d, e, and f). A notable exception occurred for  and small  (e.g., ), however, because the founder  of the branching process then gives rise to a long, unbranched initial lineage. The lineage eventually leads to the most recent common ancestor of the sample, so its mutations occur in all samples, making  noticeably larger than . As approximations, neither  nor  appear to account adequately for mutations in a long, unbranched initial lineage.

Finally, the magnitude of the ratio  of the evolutionary variance to mutational variance in  from Eq was surprisingly robust over different sample sizes  and . It decreased as a function of  and went from about 0.5 at  to about 0.1 at . Thus, in the present context of  (HIV gp120), even at , if the evolutionary variance were neglected and the standard deviation  replaced by a Poisson approximation  determined numerically from Eq , the resulting underestimate is at worst .

****

**Figure 4. The expected SFS  and**  **vs.  for different sample sizes **

In Figure 4, the Y-axes are logarithmic with the same range, and the X-axes all have categories ****. All results displayed use the HIV value . Each subfigure displays simulation results for a different sample size . The solid curves display simulation sample means  (****) corresponding to different reproduction numbers **** in a geometric series with ratio ****. From top to bottom, the curves correspond to **** (black), **** (purple), **** (dark blue), **** (light blue), **** (green), **** (orange), **** (light red), and **** (dark red). The dotted curves correspond to Eq for the Delta SFS  (implicitly connecting points for ****).

Figure 4 provides a different view of many phenomena in Figure 3. First, for  and small values of , the top dotted curves for  (particularly the black curve for ) display the crossover from  to  by overestimating . Second, in contrast to the corresponding dotted curves, the top two solid curves (for  and ****) in Figure 4a for  show an increase from  to . As in Figure 3, the increase reflects a long, unbranched initial lineage from founder . Finally, for , the two bottom dotted curves (for **** and ****) in particular display some numerical instability at very small values of . Premature truncation does not appear to be the cause, so the instability may be inherent in the delta approximation, possibly due to the exponentiation of large values of .

# Discussion

For HIV gp120, empirical results suggest estimating  rather than . Biological relevancy suggests that, say, , with a practical error in the estimate  exceeding, say, . Brute-force search for the global maximum  of  at resolution  can therefore evaluate  at  (). If desired, [Golden-section\_search](https://en.wikipedia.org/wiki/Golden-section_search) can refine the maximum.

Set  and expand the Maclaurin series for  in Eq as far as its first term, , to yield the coalescent approximation

.

The coalescent approximation also pertains to continuous-time birth-and-death processes.

The success of an invasive population often hinges on the fate of the first arrivals. Sometimes (e.g., in some viral infections), initial invaders may be very few and reproduce in near-synchronous bursts, blunting the accuracy of some population models in approximating the SFS (the Introduction and Figure 1 define the SFS verbally; Eq , mathematically). The Results section shows that in at least one important case, the Gamma model of HIV gp120 (a Bellman-Harris process), the simpler Delta model (a Galton-Watson process) can often yield analytic approximations indistinguishable by eye from the expected SFS. In particular, Figure 3 (c.f. particularly,  for , or  for ) demonstrates that the SFS can carry information about an initial viral basic reproduction number . The estimated  can serve a quantitative measure of therapeutic progress in human and animal trials of viral therapies, particularly if the trials sample more viral sequences than they do at present.

Several technical difficulties present themselves, however. HIV researchers already recognize that multiple founders impede sequence analysis ([Love et al. 2016](#_ENREF_27)). In addition, the invading viral population may pass through many environments, causing  to vary. If the technical difficulties prove superable, however, the present theory suggests a novel practical use for extant sequence data already sampled from patient and animal trials: the SFS can benchmark subtle therapeutic progress by estimating the initial , even when infection occurred and the trial had insufficient statistical power to infer therapeutic efficacy from infection data alone.

Many assumptions here are undemanding. For example, the difference between sampling with and without replacement can be neglected for light sampling ([Freedman 1977](#_ENREF_14)). In applications to HIV, each infected cell in the sampled generation produces  viral particles, and gp120 RNA averages  mutations per replication, so many gp120 samples have identical sequences anyway.

Similarly, the assumption that viral sequences are sampled simultaneously is excessively stringent. For practical purposes, the most recent common ancestor of most sampled pairs often occurs early in the population’s expansion, an assumption holding for many expanding populations, regardless of whether sampling is simultaneous.

The Delta model therefore provides robust, accurate approximations to the expected SFS for the Gamma model of HIV reproduction (see Figure 3 and Figure 4). In that context, moreover, the approximations are superior to approximations based on continuous-time birth-and-death process or the Kingman coalescent for an exponentially expanding population. (Some other results on the SFS ([Champagnat et al. 2012](#_ENREF_6), [Champagnat and Henry 2016](#_ENREF_5)), though not directly relevant to the Gamma model, are worth noting here.) The superiority should be unsurprising. The Gamma model mimics tight bursts of HIV replication every 2 days, much like a Delta model with synchronous generations every 2 days. In contrast, the continuous-time birth-and-death and the coalescent processes are Markovian, so by their nature, they do not mimic coordinated cell lysis as viruses burst forth from an infected cell. In addition, a small population such as a small initial viral population is likely to degrade approximations using the coalescent ([Stadler et al. 2015](#_ENREF_37)): “…in most cases, the coalescent approximation works very well down to small population sizes (a few hundred individuals)” ([Eriksson et al. 2010](#_ENREF_12)).

Figure 3 shows also shows that as  approaches 1 (e.g., ), Eq for the SFS  from the Delta model no longer closely approximates the SFS  for the Gamma model, but instead crosses over from  to Eq for the SFS  derived from a coalescent or continuous-time birth-and-death process. The Appendix gives a mathematical proof of the crossover, which has foundations in viewing the continuous-time birth-and-death process as the appropriate limit of Galton-Watson (Delta) processes.

For HIV gp120, , so  is much less than . In Eq , one might suspect that consequently, the mutational variance dominates the evolutionary variance. The last paragraph of Section 4 (Results) bears out the suspicion. In fact, one incisive model of gp120 phylogeny in HIV is deterministic and explicitly neglects the evolutionary variance ([Lee et al. 2009](#_ENREF_26)). The delta method in Section 2 (Theory) formally justifies the deterministic model (as well as making sense of it for non-integer ). The present paper adds a minor caveat to the deterministic model, however, particularly in its application to a study involving whole viral genomic sequence instead of a single protein. To explain, the ratio of the lengths of the HIV genome and of gp120 is 9200 / 2550 ≈ 3.6. The calculation at the end of Section 4 (Results) indicates that at , the neglect of evolutionary variance underestimates  by about , enough to start impacting error estimates, and therefore scientific conclusions. The largest viruses have genome length around 1Mb, where indiscriminate neglect of evolutionary variance may lead to error.

To summarize, this article has presented an approximation for the expected site frequency spectrum in a Galton-Watson process with mutation. In many parameter regimes, the approximation is superior to approximations from a continuous-time birth-and-death process or a coalescent process. Although the (superable) practical problems described above prevent immediate application of the theory presented here, the present article indicates the possibility of using sequence data collected after a virus has become detectable in blood to infer the initial reproduction number , with the aim of examining the efficacy of therapies for preventing or mitigating initial viral infection.

**Acknowledgements**: I thank Vruj Patel, along with Drs. Junyong Park, DoHwan Park, Mileidy Gonzalez and Anthony DeVico, for useful conversations. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

Arratia, R., Goldstein, L. and Gordon, L. (1989a). Poisson approximation and the Chen-Stein method. Statistical Science **5**: 403-434.

Arratia, R., Goldstein, L. and Gordon, L. (1989b). Two moments suffice for Poisson approximations, the Chen-Stein method. Annals of Probability **17**: 9-25.

Athreya, K.B. and Ney, P.E. (2004). Branching Processes. Mineola, New York, Dover.

Bellman, R. and Harris, T.E. (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.

Champagnat, N. and Henry, B. (2016). Moments of the frequency spectrum of a splitting tree with neutral Poissonian mutations. Electronic Journal of Probability **21**.

Champagnat, N., Lambert, A. and Richardson, M. (2012). Birth and Death Processes with Neutral Mutations. International Journal of Stochastic Analysis.

Chen, H.Y., Di Mascio, M., Perelson, A.S., et al. (2007). Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. Proceedings of the National Academy of Sciences of the United States of America **104**: 19079-19084.

De Boer, R.J., Ribeiro, R.M. and Perelson, A.S. (2010). Current Estimates for HIV-1 Production Imply Rapid Viral Clearance in Lymphoid Tissues. Plos Computational Biology **6**.

Durrett, R. (2008). Probability Models for DNA Sequence Evolution. New York, Springer Science + Business Media, LLC.

Durrett, R. (2013). Population genetics of neutral mutations in exponentially growing cancer cell populations. Annals of Applied Probability **23**: 230-250.

Durrett, R. and Limic, V. (2001). On the quantity and quality of single nucleotide polymorphisms in the human genome. Stochastic Processes and Their Applications **93**: 1-24.

Eriksson, A., Mehlig, B., Rafajlovic, M., et al. (2010). The Total Branch Length of Sample Genealogies in Populations of Variable Size. Genetics **186**: 601-611.

Fiebig, E.W., Wright, D.J., Rawal, B.D., 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.

Freedman, D. (1977). Remark on difference between sampling with and without replacement. Journal of the American Statistical Association **72**: 681-681.

Giorgi, E.E., Funkhouser, B., Athreya, G., et al. (2010). Estimating time since infection in early homogeneous HIV-1 samples using a poisson model. BMC Bioinformatics **11**.

Gordon, S.N., Liyanage, N.P.M., Doster, M.N., 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.

Griffiths, R.C. and Tavare, S. (1998). The age of a mutation in a general coalescent tree. Stochastic Models: 273-295.

Haaland, R.E., Hawkins, P.A., Salazar-Gonzalez, J., et al. (2009). Inflammatory Genital Infections Mitigate a Severe Genetic Bottleneck in Heterosexual Transmission of Subtype A and C HIV-1. Plos Pathogens **5**.

Harris, T.E. (1948). Branching Processes. Annals of Mathematical Statistics **19**: 474-494.

Harris, T.E. (1963). The Theory of Branching Processes. Berlin, Springer-Verlag.

Kahn, J.O. and Walker, B.D. (1998). Acute human immunodeficiency virus type 1 infection. New England Journal of Medicine **339**: 33-39.

Keele, B.F., Giorgi, E.E., Salazar-Gonzalez, J.F., 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.

Kimura, M. (1969). Number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893-903.

Knuth, D.E. (1992). 2 Notes on Notation. American Mathematical Monthly **99**: 403-422.

Kosaka, P.M., Pini, V., Calleja, M., et al. (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**.

Lee, H.Y., Giorgi, E.E., Keele, B.F., et al. (2009). Modeling sequence evolution in acute HIV-1 infection. Journal of Theoretical Biology **261**: 341-360.

Love, T.M.T., Park, S.Y., Giorgi, E.E., et al. (2016). SPMM: estimating infection duration of multivariant HIV-1 infections. Bioinformatics **32**: 1308-1315.

Markowitz, M., Louie, M., Hurley, A., et al. (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.

Nolen, T.L., Hudgens, M.G., Senb, P.K., et al. (2015). Analysis of repeated low-dose challenge studies. Statistics in Medicine **34**: 1981-1992.

Ohtsuki, H. and Innan, H. (2017). Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. Theoretical Population Biology **117**: 43-50.

Patel, P., Borkowf, C.B., Brooks, J.T., et al. (2014). Estimating per-act HIV transmission risk: a systematic review. Aids **28**: 1509-1519.

Pegu, P., Vaccari, M., Gordon, S., et al. (2013). Antibodies with High Avidity to the gp120 Envelope Protein in Protection from Simian Immunodeficiency Virus SIVmac251 Acquisition in an Immunization Regimen That Mimics the RV-144 Thai Trial. Journal of Virology **87**: 1708-1719.

Regoes, R.R., Longini, I.M., Feinberg, M.B., et al. (2005). Preclinical assessment of HIV vaccines and microbicides by repeated low-dose virus challenges. Plos Medicine **2**: 798-807.

Ribeiro, R.M., Qin, L., Chavez, L.L., 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.

Schuh, H.J. (1982). A note on the Harris-Sevastyanov transformation for supercritical branching-processes. Journal of the Australian Mathematical Society Series a-Pure Mathematics and Statistics **32**: 215-222.

Spouge, J.L. (2018). An Accurate Approximation for the Expected Site Frequency Spectrum in a Galton-Watson Process with Mutation. Manuscript in preparation.

Stadler, T., Vaughan, T.G., Gavryushkin, A., et al. (2015). How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics? Proceedings of the Royal Society B-Biological Sciences **282**.

Stafford, M.A., Corey, L., Cao, Y.Z., et al. (2000). Modeling plasma virus concentration during primary HIV infection. Journal of Theoretical Biology **203**: 285-301.

Strbo, N., Vaccari, M., Pahwa, S., 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.

Ver Hoef, J.M. (2012). Who Invented the Delta Method? The American Statistician **66**: 124-127.

# 

# Appendix