An Accurate Approximation for the Expected Site Frequency Spectrum 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: site frequency spectrum; 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.

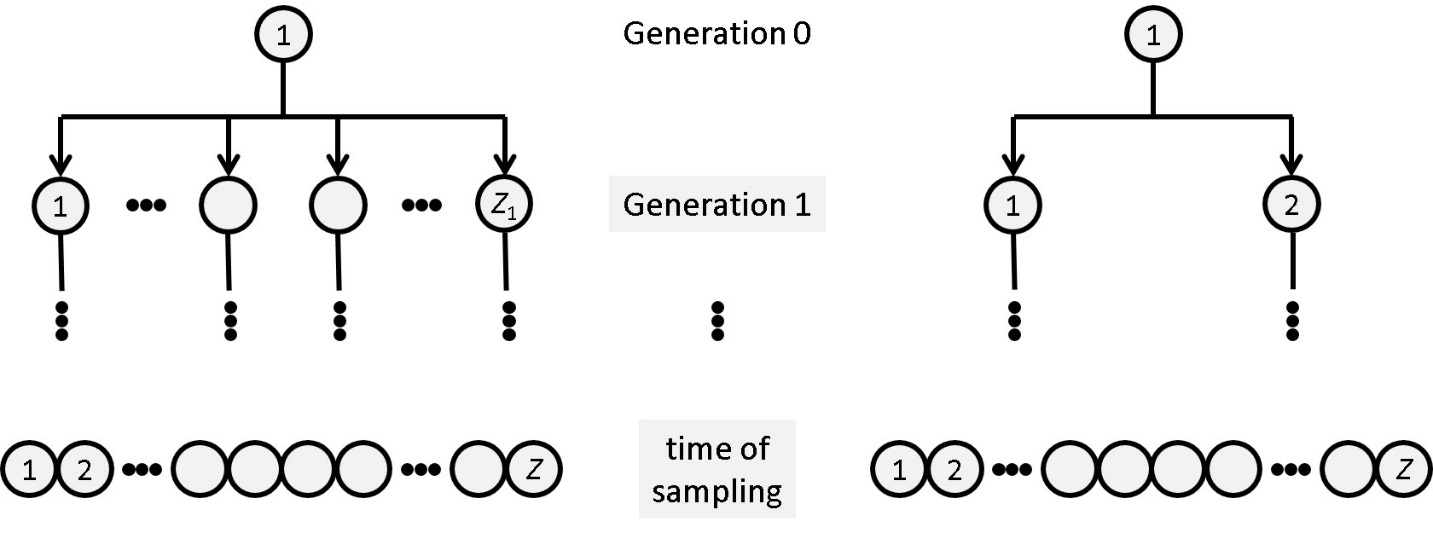
# Introduction

The theory in this article is strongly motivated by the practical observation that in infection, the success of the invasive population often hinges on the fate of the first arrivals. In the viral infection of a single host, the invasive population often descends from a small set of founder viruses. In some cases, it even descends from a single founder, e.g., about 80% of human immunodeficiency virus (HIV) infections have a single founder ([Keele et al. 2008](#_ENREF_24), [Haaland et al. 2009](#_ENREF_19), [Love et al. 2016](#_ENREF_29)). 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.

Presently, if a researcher wishes to assess therapeutic progress by demonstrating a decrease in host infectability, only one type of datum has direct pertinence, i.e., the binary datum corresponding to whether or not infection occurs ([Pegu et al. 2013](#_ENREF_34), [Strbo et al. 2013](#_ENREF_41), [Gordon et al. 2016](#_ENREF_16)). Animal trials testing viral therapies have therefore developed ingenious designs ([Regoes et al. 2005](#_ENREF_35), [Nolen et al. 2015](#_ENREF_31)), primarily to squeeze binary data for the statistical power required to detect subtle changes in infectability. An estimate of  would provide an additional measure of therapeutic success, beyond the binary data of infection or viral extinction. A quantitative analysis estimating the initial  would therefore help bypass a stubborn bottleneck in the systematic development of viral therapies.

Unfortunately, the initial  is usually not directly accessible, because the initial viral population is typically latent, i.e., below the threshold of detectability. Although the literature for HIV gives the estimates  ([Stafford et al. 2000](#_ENREF_40)) or  ([Ribeiro et al. 2010](#_ENREF_36)), e.g., these estimates pertain to viremia (i.e., after HIV becomes detectable in blood) ([Fiebig et al. 2003](#_ENREF_13)), which starts on average about 10 days after the founders infect ([Kahn and Walker 1998](#_ENREF_23)). The lower limit of HIV detection in blood is about 20 viruses/ml ([Kosaka et al. 2017](#_ENREF_27)), so a viremia implies that the total blood volume (5L) contains at least about 105 viruses. When HIV is detected in blood, therefore, the viral invasion has long since secured its beachhead. Indeed, if current estimates ( or ) of the basic reproduction number in viremia were pertinent during the initial infection, a single infected cell would probably ensure successful infection. Note, however, that the invasive routes with the highest estimates of per-act HIV transmission risk are (per 10,000 exposures) blood transfusion (9250), mother-to-child transmission (2260), and receptive anal intercourse (138). Other routes have an HIV transmission risk less than 63 per 10,000 exposures (see Table 1 in ([Patel et al. 2014](#_ENREF_33))). Because HIV transmission is so uncertain, a typical HIV founder virus probably faces a much more hostile initial environment than the current direct estimates of  in viremia suggest.

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.



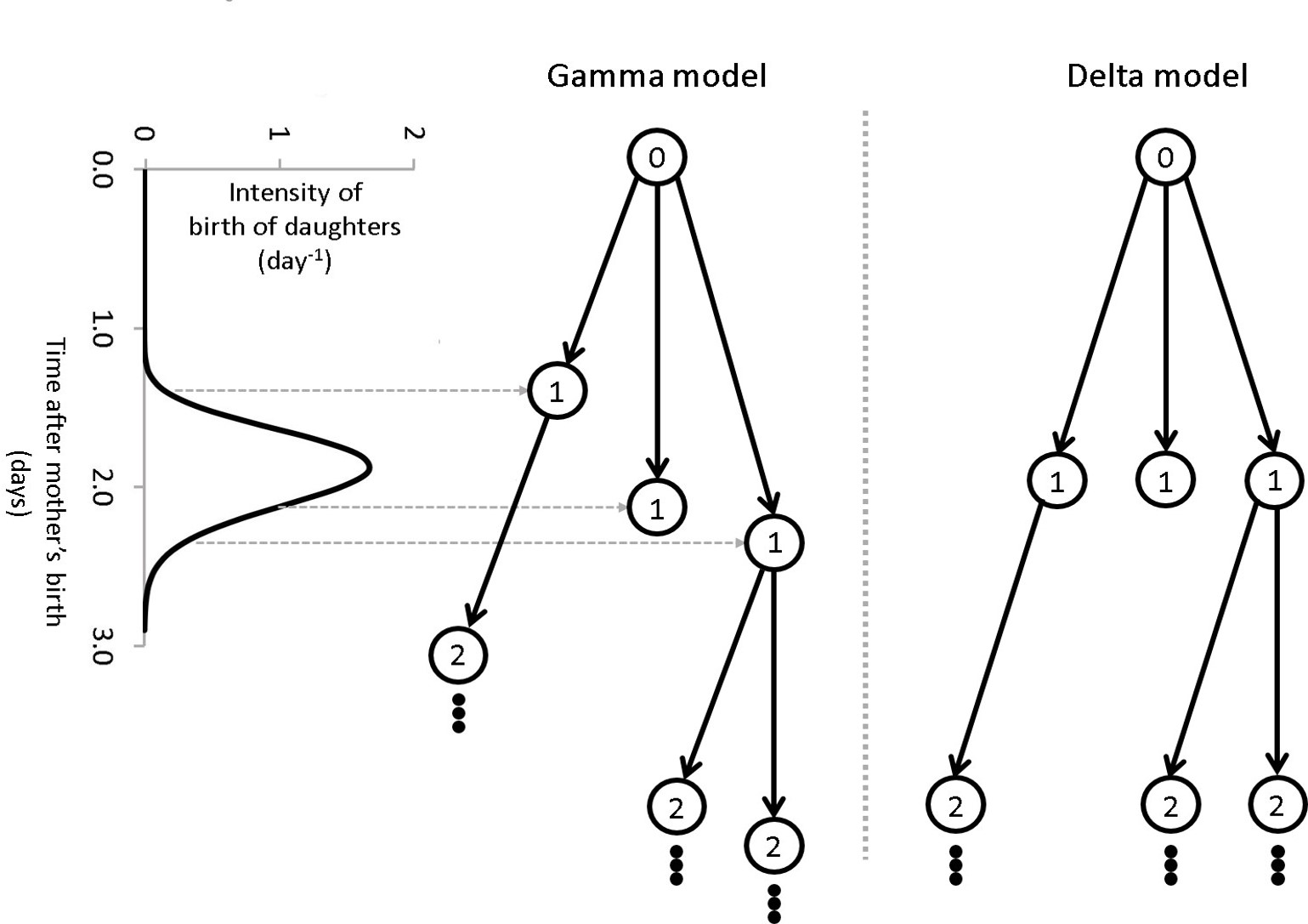
**Figure 1: Two viral ancestries, illustrating the effect of the initial**  **on samples.**

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_28)). 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_29)), 2 days being a reasonable approximation ([Markowitz et al. 2003](#_ENREF_30)). 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.

# Theory

**An Approximation for the Site Frequency Spectrum of a Galton-Watson Process**

With the Delta model of Figure 2 in mind, consider a discrete-time branching process starting with one individual in the generation 0. Let  denote generation  of the branching process (), let  count the individuals in  (e.g., ), let  denote individual  in generation  (; ), and let  have  daughters. Thus, . The  are mutually independent, identically distributed non-negative integer random variates, i.e., the process is a Galton-Watson process.

To streamline the subscripted notation  for mathematical manipulation, define , e.g., . The case  (a supercritical Galton-Watson process) is of greatest interest in the viral application, because (as the Introduction indicates) the sampling in viremia entails a large viral population. Let  denote the probability generating function (pgf) of the number of daughters in , so . The Galton-Watson process goes extinct if  for some . The extinction probability  satisfies  (i.e., an individual’s lineage goes extinct if and only if all its daughters’ lineages go extinct).

The Delta model is a Galton-Watson process where the  share a Poisson distribution with mean , so the relevant pgf is

.

Perhaps predictably, the approximations below depend only on , and no other feature of .

Fix some large generation  (where  later), and let  have  descendants in  (; ). For example,  for every  (because by convention, every individual in  is her own descendant), and  for every  (because every individual in  has exactly one ancestor in ). Sample  individuals with replacement from , and let  count the descendants of  in the sample from generation . (Sampling with replacement is mathematically more convenient than sampling without replacement, but as with other uncontrolled approximations here, simulation results are required to assess its accuracy.) Each sampled individual has exactly one ancestor in  (), so the sample from  implicitly samples ancestors from , with  for every .

If the Galton-Watson process survives long enough to be sampled at some sufficiently large  (in the Introduction, e.g., the viral population has become detectable in blood), all sampled ancestors in early generations had lineages that for any practical purpose have survived forever. For brevity, call ancestors whose lineages survive forever “immortal”; all others, “doomed”. As another uncontrolled approximation, delete all doomed individuals from the Galton-Watson process. The individuals remaining, if any, form another Galton-Watson process, the so-called skeleton process of immortal individuals.

To rephrase, if an individual is doomed, for  large enough samples from  never contain any of her descendants. To focus on relevant individuals, therefore, we replace the original Galton-Watson process with the skeleton process and drop the over-tildes from the associated quantities , , , , etc. To avoid the triviality of an empty skeleton process, all probabilities and expectations below implicitly condition on survival of the original process. The Appendix shows that the skeleton process has the same basic reproduction number  as the original process. Because the approximations below depend only on the basic reproduction number , dropping over-tildes and replacing the original process with the skeleton process does not impact the analysis.

To introduce genealogical conventions into the skeleton process, for linguistic convenience let  be both her own ancestor and her own descendant. In contrast, let the strict partial order  denote that  is an ancestor of  and . If  and , therefore,  is the mother of , and  is the daughter of .

To introduce mutations into the genealogy of the skeleton process, note that a daughter may differ in nucleic acid sequence from her mother. To compare the sequence differences, conceptually we align the relevant sequences from every individual  (; ). Let the Poisson random variate  (e.g., ([Giorgi et al. 2010](#_ENREF_15))) count the novel point mutations in , namely, the alignment columns where  has a different letter from her mother (e.g.,  if  and her mother have the same nucleic acid sequence). Assume an infinite sites model ([Kimura 1969](#_ENREF_25)), which restricts the mutations allowed: if  and her mother differ in alignment column  (i.e.,  bears a novel mutation in column ), no other mother-daughter pair differs in column . Thus, each column contains at most two different letters, and all individuals bearing a mutation (relative to , the founder) descend from a single ancestor . Assume that , i.e., the expectation of the count of novel mutations in a daughter sequence is some fixed constant .

Define an indicator random variate for -fold sampling of the ancestor  (; ):  if ; and  otherwise. With the theoretical preliminaries in hand, consider now the random variate



(), which counts novel mutations over all individuals , but only those novel mutations displayed by exactly  of the  sampled individuals. Under the infinite sites model, each novel mutation occurs in a different column, so from the perspective of the sampled sequences, Eq counts all columns containing  letters mutated away from the founder, i.e.,  counts the distinct mutations that appear in exactly  sampled sequences. Thus,  constitutes the SFS in Figure 1.

The so-called delta method ([Ver Hoef 2012](#_ENREF_42)) yields the approximation in Eq below for the expected SFS . The delta method expands



into a Taylor series around , takes expectations, and truncates the resulting series before  to yield . As a preliminary,

,

where the first equality follows from the exchangeability of ; the second, from the fact that every individual in the -th generation has exactly one ancestor in the -th generation (); and the approximation, from the delta method applied to . A second application of the delta method to  with  shows that

.

Assume that the implicit sampling  in  depends only weakly on the size  of .

,

where the second equality uses the mean , the independence properties of the  (another model assumption), and a property of the expectations conditioned on ; the third (approximate) equality depends on the above assumption of weak dependence; and the final equality uses both the Malthusian population growth  of the skeleton process and the exchangeability of  (so ).

Together, therefore, Eqs - approximate  as a function of : for ,

,

where the equality on the right of Eq defines the (approximate) Delta SFS .

For , convergence of the geometric series  shows that the sum  in Eq converges as . For ,  has a very different behavior:

.

Now,

,

so the sum on the right of Eq increases with  and is bounded above by the finite sum  of a geometric series. Thus, the difference on the left of Eq increases to a finite limit as . Thus, .

For , the second factor in the middle expression of Eq expands to yield an incidental alternative expression,

,

whereas

,

For computing, because the final sum in Eq alternates in sign, it is less reliable numerically than Eq . For , however, Eq provides a convenient check on programming errors:

.

To compare the above approximations to other models of interest, consider a continuous-time birth-and-death branching process with mutation rate , birth rate , and death rate , yielding a basic reproduction number . Let  be the total population at the time  of sampling  individuals. For  (our main interest), if  is large, the approximate expected SFS is



independent of  and the exact value of , where  in the authors’ notation of Eqs (27) and (28) of ([Ohtsuki and Innan 2017](#_ENREF_32)). According to those authors, Kingman’s coalescent for an exponentially growing population (e.g., ([Griffiths and Tavare 1998](#_ENREF_18))) yields the same approximation. (See also Theorem 2 of ([Durrett 2013](#_ENREF_10)) about the Moran model.) Accordingly, the right side of Eq defines , the (approximate) coalescent SFS.

To relate Eq to Eq , let . For , the Appendix shows that

,

i.e., the ratio of the Delta and coalescent SFSs approaches 1 as  decreases to 1.

In many models of exponentially expanding populations (e.g., the birth-and-death process ([Champagnat and Henry 2016](#_ENREF_5))), the moments of  converge to fixed, finite values at infinite times. Here in a Galton-Watson process, the SFS approximations  () display a similar convergence as . Accordingly, Section 4 presents numerical results only for the limit  in Eq .

# Methods

A technique for reducing variance ([Hammersley and Handscomb 1964](#_ENREF_20)) also reduces the programming effort in simulating the SFS for the Gamma and Delta models, as follows. For either model, let  count the non-founding ancestors with  descendants in the sample, and call  the ancestral sample frequency spectrum. (Note that the term “allele frequency spectrum” is sometimes used synonymously with SFS, but it is preferable to reserve it (e.g., ([Griffiths and Pakes 1988](#_ENREF_17))) for the analogous variate under the infinite alleles model ([Kimura and Crow 1964](#_ENREF_26)). The terminological precision then becomes particularly useful when naming .) For any model with synchronous generations (e.g., the Delta model),

,

so

,

because  counts the descendants of  in the sample. Some equivalent of the following result is doubtless stated elsewhere.

**Theorem 1**:Consider an infinite sites model where the novel mutation counts in every daughter of every mother are independent Poisson variates with fixed mean  (e.g., the Gamma or Delta model). Given , the coordinates of  are independent Poisson variates, with  having mean .

**Proof**: The coordinates of  are determined by the novel mutation counts in disjoint ancestral sets of sizes . Thus, independence of mutation implies each  is the sum of  Poisson variates of mean , with all Poisson variates independent. The sum of  Poisson variates of mean  is a Poisson variate  of mean , where the variates  are independent. 

Theorem 1 implies that under the Gamma or Delta model, . In addition, the law of total variance yields

,

where the second equality follows from Theorem 1 and the fact that the variance of a Poisson variate equals its mean. Under the Gamma or Delta model, therefore, the simulation of ancestries alone suffices to estimate the first two moments of . On the right side of Eq , the first term is sometimes called the mutational variance; the second, the evolutionary variance. Besides variance reduction, the simulation of  clarifies the relative contributions of the mutational and evolutionary variance in Eq by displaying their dependence on .

For different , simulation yielded 1000 realizations of the Gamma and Delta models. As a check on using the skeleton process in Section 2, each realization started with one founder and propagated the population, simply restarting it with another founder if the population became extinct. The realization continued until the population reached a threshold of 6000 live individuals. Here, 6000 is an arbitrary large number, chosen on the (possibly irrelevant) basis that a neutral model in coalescent theory estimated the effective size of the viral population in HIV patients between 2000 and 6000 ([Seo et al. 2002](#_ENREF_38)). As a check on using sampling with replacement in Section 2, each realization sampled the 6000 live individuals uniformly without replacement.

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

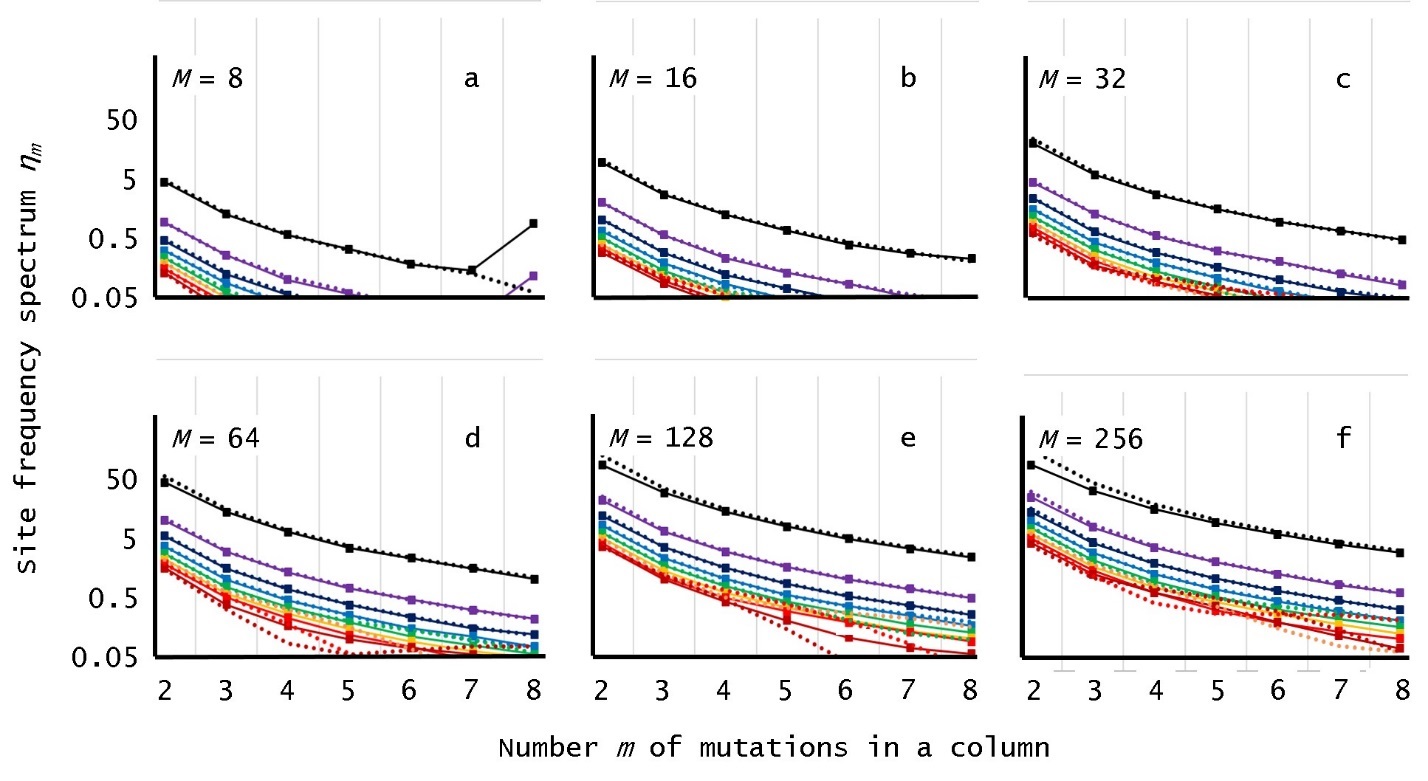
**Figure 3. Selected plots of the SFS  vs. the basic reproduction number**

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

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_29)). 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_39)): “…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_28)). 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 Pakes, A.G. (1988). An infinite-alleles version of the simple branching-process. Advances in Applied Probability **20**: 489-524.

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**.

Hammersley, J.M. and Handscomb, D.C. (1964). Monte Carlo Methods. London, Chapman and Hall.

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.

Kimura, M. and Crow, J.F. (1964). Number of alleles that can be maintained in finite population. Genetics **49**: 725-738.

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.

Seo, T.K., Thorne, J.L., Hasegawa, M., et al. (2002). Estimation of effective population size of HIV-1 within a host: A pseudomaximum-likelihood approach. Genetics **160**: 1283-1293.

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

**The Skeleton Process**

The Harris-Sevastyanov transformation ([Harris 1948](#_ENREF_21), [Harris 1963](#_ENREF_22)) gives the pgf  of the skeleton process:

.

To keep the article self-contained, we note that Eq has the following heuristic justification (see, e.g., ([Schuh 1982](#_ENREF_37)) or Part D.12 Decomposition of the Supercritical Branching Process, p.47 *et seq.* in ([Athreya and Ney 2004](#_ENREF_3))). The pgf  deletes the founder’s doomed daughters independently with the correct probability . The founder is either doomed (with probability ) or immortal (with probability ). If she is immortal, she is part of the skeleton process, which has pgf . Thus,

,

justifying Eq . Eq shows that , i.e., the original and skeleton processes have the same basic reproduction number.

Note that the Harris-Sevastyanov transformation reduces the variance of the offspring distribution, probably contributing to the accuracy of deterministic approximations in this article:

.

**Proof of the Limit in Eq for** 

Consider the function  and partition  into subintervals with the points  (). Let . Almost immediately, the theory of approximating integrals by Riemann sums shows that

.

Substitution of  into Eq yields

,

after using standard results for the Beta integral in the middle expression. Thus,

.