**Introduction.** The basic reproduction numberis defined as the expected number of cells that a single infected cell will itself infect (Giorgi et al. 2010) and provides a useful metric for establishing whether a particular infecting viral strain will result in systematic infection. To be precise, if a particular founding virus has an R0 < 1, it is reproducing below replacement and its progeny will consequently go extinct. In the opposite case, when R0 > 1, the founding virus has a reasonable chance of amplifying into systemic infection. The magnitude of R0 provides useful information: if R0 is only slightly greater than 1, the infection has a correspondingly smaller chance of survival; if R0 is significantly larger than 1, the virus is likely to thrive.

The only currently available variable in assessing the adequacy of therapeutic agents in preventing or treating HIV (or its simian counterpart SIV) is whether or not infection takes hold (Strbo et al. 2013, Gordon et al. 2016). Despite the ever more clever experimental designs used in the animal trials testing these therapeutic agents (Regoes et al. 2005, Nolen et al. 2015), the limitations of using a single binary variable in statistical analysis are hard to overcome. The ability to estimate R0 would consequently allow for a more comprehensive, quantitative analysis of the efficacy of HIV prophylactics, vaccines, and treatments.

It is important to consider that HIV must navigate a variety of environments before it achieves systemic infection and is detectable in the blood. HIV proliferates classically by so-called cell-free spread, whereby viral particles bud from an infected T cell and enter the extracellular fluid to infect another T cell encountered by chance (Zhang et al. 2015). Recent research has elucidated a second mode of transmission, with two unique mechanisms: cell-to-cell spread that occurs either by transmission of the virus by virological synapses between adjacent T cells or by *in trans* capture and transfer of virions between spatially close macrophages and dendritic cells (Zhang et al. 2015). Cell-to-cell spread is much more efficient than cell-free spread and is understood to be particularly important in the environment of lymphoid tissues where CD4+ T cells are much more densely packed. Therefore, such an environment likely corresponds to a much higher R0. In contrast, the relative paucity of T cells outside the lymphatic and circulatory systems suggests that HIV has a much more difficulty establishing itself when introduced outside these environments.

This is further corroborated by the large discrepancy in the number of per-act transmissions per 10,000 exposures between routes of infection. Sexual exposure ranges from less than 4 to 138, needle-sharing produces similar numbers (63). Meanwhile, vertical transmissions between mother and child have a risk 2260 per 10,000 exposures, while blood transfusions have the highest risk at 9250 (Patel et al. 2014). Together, these facts suggest that the reproductive potential of HIV is starkly limited until it escapes its initial environment into systemic circulation.

The initial R0 is usually inaccessible due to practical limitations in detecting HIV before systemic infection takes hold, on average 10 days after exposure (Kahn and Walker 1998). This calls into question the validity of using the current estimates of R0 ≈ 6 (Stafford et al. 2000) or R0 ≈ 8 (Ribeiro et al. 2010) for the entirety of infection (and particularly early in the infection) because these estimates were calculated in reference to viremia, when there at least 20 viruses/ml (the current lower limit of detectability) or about 105 viruses in the total blood volume of 5L (Kosaka et al. 2017). Clearly, by the time HIV is detectable in blood samples it has long since established itself in the body.

The genetic diversity at early stages of infection is typically related to the number of viruses transmitted. The sharp reduction in viral sequence diversity has been studied extensively in the context of sexual or mother-to-child transmission of HIV (Wolinsky et al. 1992, Delwart et al. 2001, Derdeyn et al. 2019). Correspondingly, it has been estimated that about 80% of all HIV infections arise from a single founding viral sequence (Keele et al. 2008, Haaland et al. 2009, Love et al. 2016). This further validates the hypothesis that the initial stages of infection present the greatest difficulty for HIV to successfully establish itself. In addition, it provides a convenient starting point for the mathematical analysis.

The advent of modern sequencing techniques has allowed for the practical measurement of sequence diversity in HIV-infected individuals (Salazar-Gonzalez et al. 2008). Despite the difficulty in directly measuring R0 early in infection, the variation in a sample of viral sequences taken during early viremia can be analyzed for information about the initial R0 of the burgeoning virus.

[Figure 1 & 2?]

Consider two hypothetical founder viruses, where the first produces some large number of daughters Z1, corresponding to a large initial R0, and the second produces a small number, say 2, of daughters Z2, corresponding to an R0 of order 1. If the phylogenies are allowed to propagate until the total population size reaches some large number N and a sample of M is taken such that Z1 >> M(M-1)/2, the M sampled sequences are likely descended from different daughters of the first virus and are unlikely to share any mutations away from the founder sequence. If a mutation occurs in either of the 2 children of the second virus, however, it is expected that about half of the sample will share this mutation. This, in principle, demonstrates the utility of examining the distribution of mutations in a given sample of some large population to extract information about the early R0 of the virus.

It has been demonstrated that the so-called site frequency spectrum (SFS) – a vector **η** where each component ηm counts the number of sites where a mutation appeared in exactly m of the M sampled sequences – can be generated as a function of R0 (Spouge 2019). The SFS itself is practically inaccessible because the founder sequence is typically unknown; however, the consensus sequence of sampled viruses can be used as a proxy. By counting the number of minority letters (i.e. letters differing from the consensus), a folded site frequency spectrum

(η’m= ηm + ηM-m) can be generated which counts the number of alignment columns where the number of minority letters is m.

This paper seeks to show that a method of moments and maximum likelihood estimation of R0 as functions of the folded site frequency spectrum are possible and effective in regimes where the initial R0 is small.

**Methods**. Before a direct accounting of the algorithms responsible for generating a viral ancestry, it is useful to elucidate the biologically relevant parameter ranges. Typical studies sampling the HIV gp120 gene sequence from human patients use M ranging from 16 to 30 per patient (Lee et al. 2009). The gene itself is about 2550 nt long and each replication averages

ε ≈ 2.16 x 10-5 point mutations/base/replication (Giorgi et al. 2010). Simple multiplication yields µ ≈ 0.0551 mutations per replication in gp120. Estimates of HIV replication time range from 1.76 to 4.2 days (Love et al. 2016). (Talk about parameters of Poisson and gamma distributions here?)

Generate an array of R0 values to test. For each value of R0, perform 1000 realizations. Each realization consists of generating a random seed, letting a population reproduce from a single founder to 6000 live (un-burst) viruses, and calculating method of moments and maximum likelihood estimations of R0 for each individual realization and the mean estimated error of the R0 estimate. These estimates are pushed onto an array. Once 1000 realizations of a particular R0 have been calculated, the sample standard deviation in the estimates of R0 is calculated.

The population reproduces as follows: first, a single founder virus is created and added to a new list which will contain unburst viruses. While the size of this list is below the 6000 threshold, the list is first sorted by age and a random number, according a Poisson distribution, of daughters is added to the oldest virus in the list, with each daughter assigned a burstday randomly according to a Gamma distribution. These daughters are then added to the list and the parent virus is removed. If, at any point, the list becomes empty – i.e. the founder virus’s lineage goes extinct – a new random seed is generated and the process begins anew.

The calculations proceed as follows: first, M samples are taken from the live list and Am (m = 1,2,...,M) is calculated by counting the non-founder ancestors that have m descendants in the sample. ηm (m = 1,2,...,M) is calculated by sampling Poisson distributions around mu = 0.0551 multiplied by each Am. η’m (m = 1,2,...,M\_tilde) is then calculated by folding ηm. From η’m. R0 is estimated by maximum likelihood estimation, with and without an Euler-Maclaurin approximation, and the method of moments. A grid search initially bounds the search space and a golden-section search refines the estimate of R0. Standard deviations are calculated from these estimated R0’s and pushed onto the same output list.

Once the entire simulation is complete, these lists are outputted directly into a “raw\_data” CSV file. Means of this list are then taken and sample standard deviations are calculated. These results are then pushed onto another list and outputted into a “means” CSV file.

**Results.**

**Discussion.**

Degnan, James H, and Noah A Rosenberg. “Discordance of Species Trees with Their Most Likely Gene Trees.” *PLoS Genetics* 2, no. 5 (May 2006). <https://doi.org/10.1371/journal.pgen.0020068>.

Delwart, Eric, Magdalena Magierowska, Maya Royz, Brian Foley, Lorraine Peddada, Richard Smith, Charles Heldebrant, Andrew Conrad, and Michael Busch. “Homogeneous Quasispecies in 16 out of 17 Individuals during Very Early HIV-1 Primary Infection.” *AIDS* 16, no. 2 (January 25, 2002): 189–95.

Derdeyn, Cynthia A., Julie M. Decker, Frederic Bibollet-Ruche, John L. Mokili, Mark Muldoon, Scott A. Denham, Marintha L. Heil, et al. “Envelope-Constrained Neutralization-Sensitive HIV-1 After Heterosexual Transmission.” *Science* 303, no. 5666 (March 26, 2004): 2019–22. <https://doi.org/10.1126/science.1093137>.

Giorgi, Elena E, Bob Funkhouser, Gayathri Athreya, Alan S Perelson, Bette T Korber, and Tanmoy Bhattacharya. “Estimating Time since Infection in Early Homogeneous HIV-1 Samples Using a Poisson Model.” *BMC Bioinformatics* 11 (October 25, 2010): 532. <https://doi.org/10.1186/1471-2105-11-532>.

Gordon, Shari N., Namal P. M. Liyanage, Melvin N. Doster, Monica Vaccari, Diego A. Vargas-Inchaustegui, Poonam Pegu, Luca Schifanella, et al. “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, no. 7 (October 1, 2016): 2726–37. <https://doi.org/10.4049/jimmunol.1600674>.

Haaland, Richard E., Paulina A. Hawkins, Jesus Salazar-Gonzalez, Amber Johnson, Amanda Tichacek, Etienne Karita, Olivier Manigart, et al. “Inflammatory Genital Infections Mitigate a Severe Genetic Bottleneck in Heterosexual Transmission of Subtype A and C HIV-1.” *PLoS Pathogens* 5, no. 1 (January 23, 2009). <https://doi.org/10.1371/journal.ppat.1000274>.

Kahn, J. O., and B. D. Walker. “Acute Human Immunodeficiency Virus Type 1 Infection.” *The New England Journal of Medicine* 339, no. 1 (July 2, 1998): 33–39. <https://doi.org/10.1056/NEJM199807023390107>.

Keele, Brandon F., Elena E. Giorgi, Jesus F. Salazar-Gonzalez, Julie M. Decker, Kimmy T. Pham, Maria G. Salazar, Chuanxi Sun, et al. “Identification and Characterization of Transmitted and Early Founder Virus Envelopes in Primary HIV-1 Infection.” *PNAS* 105, no. 21 (May 27, 2008): 7552–57. <https://doi.org/10.1073/pnas.0802203105>.

Kosaka, Priscila M., Valerio Pini, Montserrat Calleja, and Javier Tamayo. “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, no. 2 (February 15, 2017). <https://doi.org/10.1371/journal.pone.0171899>.

Lee, Ha Youn, Elena E. Giorgi, Brandon F. Keele, Brian Gaschen, Gayathri S. Athreya, Jesus F. Salazar-Gonzalez, Kimmy T. Pham, et al. “Modeling Sequence Evolution in Acute HIV-1 Infection.” *Journal of Theoretical Biology* 261, no. 2 (November 21, 2009): 341–60. <https://doi.org/10.1016/j.jtbi.2009.07.038>.

Love, Tanzy M. T., Sung Yong Park, Elena E. Giorgi, Wendy J. Mack, Alan S. Perelson, and Ha Youn Lee. “SPMM: Estimating Infection Duration of Multivariant HIV-1 Infections.” *Bioinformatics* 32, no. 9 (May 1, 2016): 1308–15. <https://doi.org/10.1093/bioinformatics/btv749>.

Nolen, Tracy L., Michael G. Hudgens, Pranab K. Senb, and Gary G. Koch. “Analysis of Repeated Low-Dose Challenge Studies.” *Statistics in Medicine* 34, no. 12 (May 30, 2015): 1981–92. <https://doi.org/10.1002/sim.6462>.

Patel, Pragna, Craig B. Borkowf, John T. Brooks, Arielle Lasry, Amy Lansky, and Jonathan Mermin. “Estimating Per-Act HIV Transmission Risk: A Systematic Review.” *AIDS* 28, no. 10 (June 19, 2014): 1509–19. <https://doi.org/10.1097/QAD.0000000000000298>.

Regoes, Roland R, Ira M Longini, Mark B Feinberg, and Silvija I Staprans. “Preclinical Assessment of HIV Vaccines and Microbicides by Repeated Low-Dose Virus Challenges.” *PLoS Medicine* 2, no. 8 (August 2005). <https://doi.org/10.1371/journal.pmed.0020249>.

Ribeiro, Ruy M., Li Qin, Leslie L. Chavez, Dongfeng Li, Steven G. Self, and Alan S. Perelson. “Estimation of the Initial Viral Growth Rate and Basic Reproductive Number during Acute HIV-1 Infection.” *Journal of Virology* 84, no. 12 (June 2010): 6096–6102. <https://doi.org/10.1128/JVI.00127-10>.

Salazar-Gonzalez, Jesus F., Elizabeth Bailes, Kimmy T. Pham, Maria G. Salazar, M. Brad Guffey, Brandon F. Keele, Cynthia A. Derdeyn, et al. “Deciphering Human Immunodeficiency Virus Type 1 Transmission and Early Envelope Diversification by Single-Genome Amplification and Sequencing.” *Journal of Virology* 82, no. 8 (April 2008): 3952–70. <https://doi.org/10.1128/JVI.02660-07>.

Spouge, John L. “An Accurate Approximation for the Expected Site Frequency Spectrum in a Galton–Watson Process under an Infinite Sites Mutation Model.” *Theoretical Population Biology*, March 12, 2019. <https://doi.org/10.1016/j.tpb.2019.03.001>.

Stafford, M. A., L. Corey, Y. Cao, E. S. Daar, D. D. Ho, and A. S. Perelson. “Modeling Plasma Virus Concentration during Primary HIV Infection.” *Journal of Theoretical Biology* 203, no. 3 (April 7, 2000): 285–301. <https://doi.org/10.1006/jtbi.2000.1076>.

Strbo, Natasa, Monica Vaccari, Savita Pahwa, Michael A. Kolber, Melvin N. Doster, Eva Fisher, Louis Gonzalez, Donald Stablein, Genoveffa Franchini, and Eckhard R. Podack. “Novel Vaccination Modality Provides Significant Protection against Mucosal Infection by Highly Pathogenic SIV.” *Journal of Immunology* 190, no. 6 (March 15, 2013): 2495–99. <https://doi.org/10.4049/jimmunol.1202655>.

Wolinsky, S. M., C. M. Wike, B. T. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Munoz. “Selective Transmission of Human Immunodeficiency Virus Type-1 Variants from Mothers to Infants.” *Science* 255, no. 5048 (February 28, 1992): 1134–37. <https://doi.org/10.1126/science.1546316>.

Zhang, Changwang, Shi Zhou, Elisabetta Groppelli, Pierre Pellegrino, Ian Williams, Persephone Borrow, Benjamin M. Chain, and Clare Jolly. “Hybrid Spreading Mechanisms and T Cell Activation Shape the Dynamics of HIV-1 Infection.” *PLoS Computational Biology* 11, no. 4 (April 2, 2015). <https://doi.org/10.1371/journal.pcbi.1004179>.