SUPPLEMENTARY MATERIAL

Scaling law characterizing the dynamics of the transition of HIV-1 to error catastrophe

Vipul Gupta and Narendra M. Dixit*

Department of Chemical Engineering,
Indian Institute of Science, Bangalore, India, 560012
**Text S1. Parameter values**

We set parameters to values representative of HIV-1 infection in vivo (Tripathi et al., 2012; Thangavelu et al., 2014). We choose full length genomes, \( L = 10000 \) nucleotides. The cell population is chosen as a representative effective population size obtained previously from fits of our simulations of the evolution of viral diversity and divergence to longitudinal patient data (Balagam et al., 2011); accordingly, we set \( C = 5000 \) cells. The frequency of infections per cell, \( M \), follows a distribution where 77% of the cells are singly, 19% doubly, and 4% triply infected, similar to that observed experimentally (Josefsson et al., 2011). The size of the initial viral pool is set large enough to enable the requisite number of infections. Recombination is assumed to occur at the rate \( \rho = 8.3 \times 10^{-4} \) crossovers/site/replication (Suryavanshi and Dixit, 2007). HIV-1 mutations are predominantly transitions (Mansky and Temin, 1995); we thus ignore transversions, insertions and deletions and perform simulations over a wide range of mutation rates above the error threshold \( \mu > \mu_c \). Although the viral burst size is large, only a small fraction of the virions produced (often as low as 2-3 virions per cell (Dimitrov and Martin, 1995)) is infectious (Piatak et al., 1993; Dimitrov and Martin, 1995). Therefore, and following our previous studies (Balagam et al., 2011), we let \( P = 10 \) progeny virions/cell. Fitness selection follows the landscape derived experimentally (Bonhoeffer et al., 2004), where the relative fitness of genome \( i \) at a Hamming distance \( d_{iF} \) from the founder strain is represented, based on previous fits (Balagam et al., 2011; Vijay et al., 2008), by
\[
f_i = 1 - (1 - f_{\min}) \frac{(d_{iF})^n}{(d_{iF})^n + (d_{30})^n},
\]
with \( f_{\min} = 0.24 \) the minimum fitness of sequences, \( d_{30} = 30 \) the Hamming distance at which \( f_i = (1 + f_{\min})/2 \), and \( n = 3 \) analogous to the Hill coefficient.
Table S1. Parameter variations employed in our simulations. In the simulations presented in Fig. 2, we varied parameters around the case where \( L = 1000 \) nucleotides; \( C = 2400 \) cells; \( \rho = 8.4 \times 10^{-4} \) crossovers/site/replication; \( M \) follows a distribution where 77\% of the cells are singly, 19\% doubly, and 4\% triply infected; and the fitness landscape is given by
\[
f_i = 1 - \left(1 - f_{\text{min}}\right)^n \frac{(d_{if})^n}{(d_{if})^n + (d_{s0})^n},\]
with \( f_{\text{min}} = 0.24 \), \( d_{s0} = 30 \), and \( n = 3 \) (see Methods).

Parameter variations are listed below along with the corresponding symbols used in Fig. 2.

| Parameter                  | Symbol | Value       | Error threshold, \( \mu_c \) | Source of \( \mu_c \) |
|----------------------------|--------|-------------|-----------------------------|------------------------|
| Recombination rate, \( \rho \) (crossovers/site/replication) | ○      | 0           | 2.2\times10^{-4}           | Tripathi et al., 2012 |
|                           | □      | 1.0\times10^{-6} | 2.2\times10^{-4}           | Tripathi et al., 2012 |
|                           | ▽      | 1.0\times10^{-3} | 2.2\times10^{-4}           | Tripathi et al., 2012 |
|                           | ○      | 1.0\times10^{-4} | 3.0\times10^{-4}           | Tripathi et al., 2012 |
| Population size*, \( C \) (cells) | ○      | 1500        | 7.1\times10^{-5}           | Tripathi et al., 2012 |
|                           | □      | 5000        | 8.4\times10^{-5}           | Tripathi et al., 2012 |
|                           | ○      | 10000       | 9.1\times10^{-5}           | Tripathi et al., 2012 |
| Length, \( L \) (nucleotides) | ○      | 200         | 6.2\times10^{-3}           | Present study          |
|                           | □      | 500         | 1.2\times10^{-3}           | Present study          |
|                           | ○      | 1000        | 5.3\times10^{-4}           | Tripathi et al., 2012 |
| Multiplicity of infection, \( \langle M \rangle \) (infections/cell) | ○      | 1.05        | 3.3\times10^{-4}           | Present study          |
|                           | □      | 1.15        | 4.5\times10^{-4}           | Present study          |
|                           | ○      | 1.35        | 5.8\times10^{-4}           | Present study          |
| Fitness landscape          | □      | \( d_{s0} = 40.3 \) | 5.1\times10^{-4}           | Present study          |
|                           | ○      | \( d_{s0} = 20.3 \) | 5.8\times10^{-4}           | Present study          |
|                           | ○      | Polynomial† | 5.4\times10^{-4}           | Present study          |

\* \( L = 10000 \) nucleotides in these simulations.

† Relative fitness is calculated using the polynomial \( \ln f_i = ad_{if} + bd_{if}^3 + cd_{if}^5 \), where \( a = 2.2 \times 10^{-3} \), \( b = -5.47 \times 10^{-4} \), and \( c = 4.62 \times 10^{-6} \), which we identified previously as a limiting case of the HIV-1 fitness landscape structure proposed by Hinkley et al. (2011) when fitness is Hamming class invariant (Tripathi et al., 2012). The fitness of genomes beyond Hamming distance 82 was set equal to the fitness of genomes at Hamming distance 82.
Figure S1. Influence of the founder sequence. Simulations in Fig. 2 with the fittest sequence as founder (brown) were repeated with a less fit founder, obtained by mutating the fittest sequence randomly at 2% of the positions (orange). We have shown previously that $\mu_c$ is not significantly altered by varying the founder sequence (Tripathi et al., 2012). Here, we find that the scaling law obtained in Fig. 2 (gray) is also not significantly affected.

Figure S2. Influence of the number of realizations. Dependence of the mean (symbols) and standard deviation (error bars) of $\tau$ on the number of realizations for three different values of the mutation rate (colors) indicated. Other parameter values employed are listed in Table S1. Note that both the mean and the standard deviation do not vary significantly beyond 10 realizations.
Figure S3. Influence of the threshold Shannon entropy. The simulations employed in Fig. 2 reanalyzed using the threshold mean entropy $H = 0.8$ to mark completion of the transition to error catastrophe. The symbols have the same meanings as in Fig. 2. The data again yields a robust scaling law: $\tau \approx 0.4 / (\mu - \mu_c)$. The lower pre-factor (0.4 as opposed to 0.6 with $H = 0.9$) would imply a lower required treatment duration, but the duration would still be several years. Further, the certainty with which error catastrophe is realized is compromised given the lesser extent to which the quasispecies is delocalized at the lower entropy.
Figure S4. Distribution of the transition time. The distribution of $\tau$ (bars) around the mean value (pink lines) obtained from our simulations for three different mutation rates indicated (colors), fit independently to a Gaussian (black solid line) or a Poisson (black dashed line) distribution. The distribution spreads as $\mu$ approaches $\mu_c$ from above, but does not appear to acquire large tails. The Gaussian distribution fits the data better than the Poisson distribution in all cases (consistently yielding lower $R^2$ values; not shown). Parameters employed in the simulations are listed in Table S1.
Figure S5. Dependence of the error threshold on parameter values. Variation of $\mu_c$ with (a) the population size, $C$, (b) the genome length, $L$, (c) the mean number of infections per cell, $\langle M \rangle$, and (d) the recombination rate, $\rho$, obtained from our simulations (symbols). In (a), the line is the best-fit, $\mu_c = -9.0 \times 10^{-4} \frac{1}{\sqrt{C}} + 9.7 \times 10^{-5}$ ($R^2=0.95$) and in (b) the line is the best-fit, $\mu_c \approx 1.08/L$ ($R^2=0.93$), consistent with the scaling patterns observed previously (Tripathi et al., 2012). Parameter values employed are listed in Table S1.
Supplementary References

Balagam R, Singh V, Sagi A R and Dixit N M 2011 Taking multiple infections of cells and recombination into account leads to small within-host effective-population-size estimates of HIV-1 *PLoS ONE* 6 e14531

Bonhoeffer S, Chappey C, Parkin N T, Whitcomb J M and Petropoulos C J 2004 Evidence for positive epistasis in HIV-1 *Science* 306 1547-50

Dimitrov D S and Martin M A 1995 CD4+ cell turnover *Nature* 375 194-5

Josefsson L, King M S, Makitalo B, Brannstrom J, Shao W, Maldarelli F, Kearney M F, Hu W-S, Chen J, Gaines H, Mellors J W, Albert J, Coffin J M and Palmer S E 2011 Majority of CD4+ T cells from peripheral blood of HIV-1-infected individuals contain only one HIV DNA molecule *Proc Natl Acad Sci USA* 108 11199-204

Mansky L M and Temin H M 1995 Lower in-vivo mutation-rate of human-immunodeficiency-virus type-1 than that predicted from the fidelity of purified reverse-transcriptase *J Virol* 69 5087-94

Piatak M, Saag M S, Yang L C, Clark S J, Kappes J C, Luk K C, Hahn B H, Shaw G M and Lifson J D 1993 High-levels of HIV-1 in plasma during all stages of infection determined by competitive PCR *Science* 259 1749-54

Suryavanshi G W and Dixit N M 2007 Emergence of recombinant forms of HIV: Dynamics and scaling *PLoS Comput Biol* 3 e205

Thangavelu P U, Gupta V and Dixit N M 2014 Estimating the fraction of progeny virions that must incorporate APOBEC3G for suppression of productive HIV-1 infection *Virology* 449 224-8

Tripathi K, Balagam R, Vishnoi N K and Dixit N M 2012 Stochastic simulations suggest that HIV-1 survives close to its error threshold *PLoS Comput Biol* 8 e1002684

Vijay N N V, Vasantika, Ajmani R, Perelson A S and Dixit N M 2008 Recombination increases human immunodeficiency virus fitness, but not necessarily diversity *J Gen Virol* 89 1467-77