Supplementary Information

Factors limiting the transmission of HIV mutations conferring drug resistance: fitness costs and genetic bottlenecks

Bradley G. Wagner, J. Gerardo Garcia-Lerma, and Sally Blower

1) Within-host model

We used the within-host model for two purposes: to estimate detection times and to track viral dynamics within each individual in a transmission chain.

A) Structure

We used the within-host developed by Perelson and Ribeiro[1] with two slight modifications: (i) we included two strains (one wild-type and one with a DRM) and (ii) we allowed for forward and backward mutation during the replication process. The Perelson and Riberio within-host model represents the natural history of HIV infection in the primary and the chronic stage; it does not include progression from the chronic phase to AIDS. Their model is appropriate for the situation that we are modeling, specifically a resource-rich country where HIV-infected individuals receive treatment well before they have developed AIDS and when their CD4 cell count is still relatively high[2].

The within-host model tracks the competition/reversion dynamics between wild-type and strains with DRMs in the plasma of an HIV-infected individual, beginning immediately after infection. Since the wild-type virus has the highest replication rate it will eventually out-compete the strain with the DRM; the magnitude of the relative fitness cost of the DRM determines the strength of the competition. The model simulates high viral loads (and hence infectivity) during primary infection and allows the viral load (and hence infectivity) to decrease and reach a set point in the chronic phase. We use the model to track the temporal dynamics of the total viral load in a treatment-naïve individual, as well as to calculate the proportion of their viral population that consists of strains with a DRM at any point in time. The model also tracks uninfected CD4 cells, as well as CD4 cells which have been infected by either strain. The model is specified by the following set of equations:

$$\frac{dC}{dt} = \pi - \nu C - k(M+W)C \quad (1)$$

$$\frac{dT_w}{dt} = kWC - \delta T_W \quad (2)$$

$$\frac{dT_m}{dt} = kMC - \delta T_M \quad (3)$$

$$\frac{dW}{dt} = (1-\mu_{WM})pT_W + \mu_{MW}p(1-\Delta)T_M - cW \quad (4)$$

$$\frac{dM}{dt} = (1-\mu_{MW})p(1-\Delta)T_M + \mu_{WM}pT_W - cM \quad (5)$$

where: *t* represents time, *C(t)* the number of susceptible CD4 cells per microliter of blood at time t, $T_w(t)$ the number of CD4 cells per microliter of blood that are infected with the wild-type strain at time *t*, $T_m(t)$ the number of CD4 cells per microliter of blood that are infected with the strain with the DRM at time *t, W(t)* the number of wild-type virions per microliter of blood at time, *t*, *M(t)* the number of virions with DRMs per microliter of blood at time *t*, π the immigration rate of CD4 cells, ν the natural clearance rate of healthy CD4 cells, *k* the infectiousness of HIV, δ the death rate of HIV-infected CD4 cells, μ_{WM} the probability that a single nucleotide mutation that confers resistance occurs during one viral replication cycle, *p* the replication rate of the wild-type strain, μ_{MW} the probability of back-mutation (i.e., the probability the DRM reverts back to wild-type) occurring during one viral replication cycle, Δ the relative fitness cost of the DRM (specified as a fraction) and *c* the clearance rate of HIV. We set the probability of back-mutation equal to the probability of forward mutation.

Note, the quantity p/δ represents the average number of virions produced per infected cell (i.e., the burst size) for the wild-type strain, and $p(1-\Delta)/\delta$ represents the burst size for the strain with the DRM.

Each of the 10 DRMs we analyze has only a single point mutation and this occurs in the reverse transcriptase gene; this gene is not involved with binding and entry into the target cell. Consequently, we assume the infectiousness of HIV (k) is the same for all strains.

B) Parameterization

To simulate the within-host viral dynamics for each of the 10 DRMs we parameterized the model using the estimated relative fitness cost (Δ) for each DRM. For each DRM we calculated a range of fitness costs. These methods are described in the Methods section of the main text. Computed fitness cost ranges are given in Table 1 in the main text.

We estimated the values of k (the infectiousness of HIV) and p (the replication rate of the wild-type strain) by using Equations 1 to 5. We used these equations to derive expressions for the CD4 set point (C^*) and the viral load set point for the wild-type strain (W^*). To derive these expressions we made the simplifying assumption that forward mutation alone has a negligible effect on the final viral set point (i.e., we set $\mu_{WM}=0$). Rearranging the expressions for C^* and W^* we obtained expressions for p and k:

$$k = \frac{\left(\frac{\pi}{C^*} - \nu\right)}{W^*} \qquad (6)$$

$$p = \frac{c\delta}{kC^*} \qquad (7)$$

We then parameterized Equations 6 and 7 and estimated values for p and k. Values for three of the parameters in Equations 6 and 7 are fairly precisely known. Hence we used values from the literature for these parameters: the death rate of HIV-infected CD4 cells (δ) was set at 1 per day[1], the natural clearance rate of healthy CD4 cells (ν) was set at 0.02 per day[3], and the

immigration rate of CD4 cells (π) was set at 13 cells/μL/ day[4]. We assumed the immigration rate of CD4 cells (π) would be fairly constant. This assumption was based on immunological studies of HIV which have shown CD4 immigration rates only vary between 9-16 cells per microliter per day[4] and there is little temporal variation in the rates[4]. However literature estimates of the fourth parameter, the clearance rate of HIV (*c*), are very variable. Viral clearance in plasma has been previously estimated to be as low as 3 per day[5], however more recent experiments have estimated a mean value of 23 per day[6]. When estimating values for *p* and *k* we varied *c* from 10 to 30 per day[6,7]. Estimates of the CD4 set point (*C**) and the viral load set point for the wild-type strain (*W**) are also variable; based on the literature we varied *W** from 20 to 60 virions/μL[2,8,9] and we varied C* from 350 to 550 cells/μL[8]. Our estimates for *p* ranged from 100 to 900 virions per cell per day and for *k* ranged from 8.7 x 10^{-5} to 3.9x10^{-4} μL/virion/day.

When running simulations the mutation rates (μ_{MW} and μ_{WM}) were set at 10^{-5} per viral replication cycle[10].

When single point estimates of the parameters were used in model simulations the following values were used: *v*=0.02 per day, *c*=20 per day, δ=1 per day, π=13 cells per uL per day,μ=10^{-5}, *p*=248 virions per cell per day and *k*=1.63 x10^{-4} μL/virion/day.

2) <u>Stochastic transmission chain model</u>

A) Structure

Sexual transmission is modeled as a series of stochastic processes; the model includes genetic bottlenecks that occur during transmission events and reversion of DRMs to wild-type after transmission. It tracks one transmission chain and the resulting transmission cluster. In the

model, strains are transmitted from an index case (which we define as the first generation) to a second generation of individuals and then to a third generation and so on until the *n*th generation, at which point the transmission chain terminates. In each simulation of the model only one transmission chain develops and therefore only one transmission cluster is generated. In the case of the wild-type strain the chain is generated by one index case infected with only the wild-type strain and in the primary stage of infection; DRMs strains evolve within these infected individuals through mutation with a probability of 10^{-5} per viral replication cycle, but are quickly out-competed in the absence of treatment[10]. In the case of a strain with a DRM the chain is generated by one index case infected with only that strain and in the primary stage of infection; wild-type strains evolve within treatment-naïve individuals through back-mutation with a probability of 10^{-5} per viral replication cycle[10]. Viral dynamics within each individual in the chain are tracked using the within-host model. This enables us to model the changes in infectivity that occur (due to the temporal dynamics of viral load) as an individual progresses through primary infection, through chronic infection, to the treatment stage; when we assume infectivity is negligible. We made this assumption, because individuals in MSM communities in resource-rich countries generally receive effective treatment which reduces their viral load to less than 50 copies/mL; once viral load falls below 400 copies/mL transmission is negligible[11].

For each individual their simulated viral load, at any point in time, was used to calculate the probability that they transmitted HIV (i.e., their infectivity). To estimate the transmission probability ρ we used the following formula[12-14]:

$$\rho(t) = \rho^* 2.45^{\log_{10}\left(\frac{V(t)}{V^*}\right)} \qquad (8)$$

where *V(t)* represents the individuals' total viral load [i.e., *V(t)=W(t)+M(t)*]. We set the constant V* equal to 12.5 virions/μL[12,13] and ρ* equal to 0.0018[15]. Equation 8 describes the empirical relationship, that for every $\log_{10}$ increase in viral load, the per act infectivity increases by a factor of 2.45[13].

The viral load during the chronic phase generated by the within-host model is 20,000-60,000 copies/mL, which corresponds to a per act transmission probability of 0.002-0.003. We note that the viral load data the model generates is very close to the viral load of 10,000-50,000 virions/mL observed in individuals chronically infected with HIV[9], and the per act transmission probability it generates is of the same magnitude as the estimates from empirical studies[15,16]. Estimates from these studies exhibit a fairly high degree of variability but all of them indicate the per act probability of transmitting HIV for MSMs is very low[15,16].

At each transmission event, only one strain was transmitted: the resistant or the wild-type strain. This was determined stochastically, and was a function of the proportion of the viral load in the transmitting individual that was composed of resistant strains.

One realization of the stochastic model tracks one transmission chain and the associated transmission cluster; each realization represents 20 years of chronological time. An individual with an undetectable DRM, based on current or next generation resistant assays, has the potential to transmit the DRM and "contribute" to the transmission chain. Therefore, during our simulations of the transmission of DRMs we tracked individuals whether their DRM was detectable, or undetectable.

Our estimates of the length of transmission chains for DRMs, and the size of their transmission

clusters, are based on whether or not the DRM is transmitted. At the time of transmission of a DRM the strain with the DRM always constitutes 100% of the infecting virions and hence is detectable. However after transmission, the DRM can become undetectable. Therefore individuals in a transmission chain can be infected with a DRM that has become a minority strain and undetectable.

The algorithm for the stochastic transmission chain model is performed as follows:

Step 1: The length of time from an infection until treatment is determined. We begin with an index case infected with only a wild-type strain or only a strain with a DRM at time zero and in the primary stage of infection. The time until the individual receives treatment is decided by randomly choosing a number from a triangular distribution with a range of 6 to 8 years, peaked at 7 years. We denote the time to treatment as τ_1. The infectivity of the index case over time (and the proportion of their virions which contain a DRM) is tracked by the within-host model.

Step 2: The time until the first infection (i.e., transmission event) is decided. Note that though we present this step for the index case, this same procedure is used for every individual in the transmission chain. We track all infections (for all generations) for a maximum of 20 years. We use a standard Monte Carlo simulation algorithm to decide upon the time of infection. To do this we choose a random number from a uniform distribution on [0, 1], which we denote as r_1. Then the time until the first infection is given by solving the equation:

$$\log(1/(1-r_1)) = \int_0^t n\rho(a)\,da$$

for $t<\tau_1$, where ρ is the infectivity calculated using equation 8 and n is the number of sex acts per unit time. Infectivity ρ is expressed as a function of time, where the time dependence is specified implicitly in the time dependent viral load described by the dynamics of Equations 1-5. If the left hand side of the equation remains greater than the right hand side of the equation for $t=\tau_1$ then no infection occurs and the transmission chain terminates.

Using this algorithm, we track all infections, directly or indirectly, caused by the index case for 20 years. For example if the first infection occurs at t_1 then the time until the second infection (by the index case) is given by solving:

$$\log(1/(1-r_2)) = \int_{t_1}^{t_2} n\rho(a)\,da$$

for t_2 where $t_2<\tau_1$ and r_2 is a random variable drawn from a uniform distribution on [0, 1]. If the left hand side of the equation remains greater than the right hand side of the equation for $t_2=\tau_1$ then there is no (second) transmission.

Step 3: Generally, HIV infection in MSM is established by a single viral variant[17]. Therefore, at this step the probability that the strain with the DRM is the founder virus is determined. We consider the case in which infection is caused by a single virion and whether the wild-type strain or strain with the DRM is transmitted depends, probabilistically, on the composition of the virion population in the individual who is transmitting HIV. The within-host model is used to track the viral dynamics of the wild-type and the strain with the DRM within the transmitting individual. The founder virus will be a virus with a DRM if

$$r_3 < \frac{M(t)}{M(t)+W(t)}$$

where t is the time of infection decided in Step 2, r_3 is a random number drawn from a uniform distribution on [0, 1] and the right hand side of the equation represents the proportion of virions with a DRM in the plasma of the infecting individual at the time of the transmission event. Conversely if r_3 is greater than the proportion of virions with a DRM in the transmitter then the wild-type strain will be transmitted.

Step 4: At this step, the next generation is determined by using the same procedure for every infected individual in the chain (i.e., steps 1 through 3 are repeated).

B) Parameterization

In addition to the parameters of the within-host model, two other parameters are needed to simulate transmission chains; these are the time between infection and the initiation of treatment *(τ)* and the number of sex acts per unit time *(n).*

Individuals who become infected with HIV are generally not treated until there has been significant decline in their CD4 cells, which does not generally occur for several years after infection. We estimated the time between infection and the initiation of treatment using data for MSMs from the Department of Public Health in San Francisco[2]. Notably in San Francisco treatment is often initiated earlier than the current guidelines recommend; current treatment guidelines recommend initiating treatment when the patients' CD4 count has fallen to 350 cells/μL. However, in San Francisco the median CD4 count (for MSM) at the time of an individuals' first CD4 test is ~420 cells/μL. To estimate the time between infection and the

initiation of treatment we assumed: (i) treatment would be initiated (on average) when the CD4 cell count has fallen to ~420 cells/μL, (ii) the CD4 cell count in an uninfected individual is ~1000 cells/μL, and (iii) the CD4 cell count would decrease by ~200 CD4 cells/μL in the first year of infection followed by a steady annual decrease of 50 to 70 cells/μL[18,19]. Using these parameter values, we calculated that if the annual decrease is 70 cells/μL the time to treatment would be ~6 years, whereas the time to treatment would be ~8 years if the annual decrease is 50 cells/μL. Consequently, in our modeling we varied time to treatment, independently for each individual, from 6 to 8 years.

The number of sex acts per unit time *(n)* was set so that the wild-type strain had a Basic Reproduction Number (R_0) of 1.6, as this is the value for R_0 that has recently been estimated for the MSM community in San Francisco[9]. For the wild type strain R_0 is computed as

$$R_0 = \int_6^8 f(\tau)\, d\tau \int_0^\tau n\rho(t)\, dt \qquad (9)$$

where *f* represents the distribution of the time from infection until treatment. We take *f* to be triangular, with limits between 6 and 8 years which we estimated from empirical data[2] as discussed previously. We computed *n* by setting R_0=1.6 in equation 9.

The R_0 for a strain with a DRM (i.e., a resistant strain) was computed in a similar manner as

$$R_0 = \int_6^8 f(\tau)\, d\tau \int_0^\tau n\rho(t) \frac{M(t)}{W(t)+M(t)}\, dt \qquad (10)$$

where *M(t)* and *W(t)* represent the resistant and wild-type viral loads.

References:

1 Perelson, A. S. & Ribeiro, R. M. Estimating drug efficacy and viral dynamic parameters: HIV and HCV. *Stat Med* **27**, 4647-4657 (2008).

2 San Francisco Department of Public Health HIV/AIDS epidemiology annual report (2008).

3 Ribeiro, R. M., Mohri, H., Ho, D. D. & Perelson, A. S. In vivo dynamics of T cell activation, proliferation and death in HIV-1 infection: why are CD4+ but not CD8+ T cells depleted? *Proc Natl Acad Sci* **99**, 15572-15577 (2002).

4 Hellerstein, M. *et al.* Directly measured kinetics of circulating T lymphocytes in normal and HIV-1-infected humans. *Nat Med* **5**, 83-89 (1999).

5 Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582-1586 (1996).

6 De Boer, R. J., Ribeiro, R. M. & Perelson, A. S. Current estimates for HIV-1 production imply rapid viral clearance in lymphoid tissues. *PLoS Comput Biol* **6**, e1000906 (2010).

7 Ramratnam, B. *et al.* Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *Lancet* **354**, 1782-1785 (1999).

8 Lyles, R. H. *et al.* Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. Multicenter AIDS Cohort Study. *J Infect Dis* **181**, 872-880 (2000).

9 Smith?, R. J., Okano, J. T., Kahn, J. S., Bodine, E. N. & Blower, S. Evolutionary dynamics of complex networks of HIV drug-resistant strains: The case of San Francisco. *Science* **327**, 697-701 (2010).

10 Mansky, L. M. & Temin, H. M. Lower in vivo mutation rate of human immunodeficiency virus type 1 than predicted from the fidelity of purified reverse transcriptase. *J Virol.* **69**, 5087-5094 (1995).

11 Attia, S., Egger, M., Müller, M., Zwahlen, M. & Low, N. Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis. *AIDS* **23**, 1397-1404 (2009).

12 Quinn, T. C. *et al.* Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. *N Engl J Med* **342**, 921-929 (2000).

13 Gray, R. H. *et al.* Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1 discordant couples in Rakai, Uganda. *Lancet* **357**, 1149-1153 (2001).

14 Smith, R. J. & Blower, S. Could disease-modifying HIV vaccines cause population-level perversity? *Lancet Infect Dis* **4**, 636-639 (2004).

15 Vittinghoff, E. *et al.* Per-contact risk of human immunodeficiency virus transmission between male sexual partners. *Am J Epidemiol* **150**, 306-311 (1999).

16 Fox, J. & Fidler, S. Sexual transmission of HIV-1. *Antiviral Res* **85**, 276-285 (2010).

17 Keele, B. F. *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci* **105**, 75552-75557 (2008).

18 Pillay, D. *et al.* The impact of transmitted drug resistance on the natural history of HIV infection and response to first-line therapy. *AIDS* **20**, 21-28 (2006).

19 Rodríguez, B. *et al.* Predictive value of plasma HIV RNA level on rate of CD4 T-cell decline in untreated HIV infection. *JAMA* **296**, 1498-1506 (2006).

Table S1: For each DRM and the wild-type strain 10,000 simulations were conducted using the stochastic transmission chain model. Results shown below are based on the simulations in which transmission occurred. Not all simulations led to transmission (see Table 2 in main text).

DRM	Standard Deviation (SD) of the Cluster Size (based on simulations in which transmission occurred)
Wild-type	80
K70R	67
Y181C	54
K219Q	7.3
L74V	2.7
D67N	1.1
M41L	0.81
K103N	0.63
T215Y	0.33
M184V	0.27
K65R	0.24

Figure S1: Numerical results from the within-host model showing the fraction of virions that are resistant (red curve) as a function of years since infection in a treatment-naïve individual. The blue curve indicates the fraction of virions that are wild-type. The 20% threshold for detection of resistance in treatment-naïve individuals, based on current resistance assays, is denoted by the dashed black line. DRMs shown are T215Y (A), K103N (B), D67N (C), and Y181C (D).

Figure S2: Box plots showing distributions of the length of transmission chains for four DRMS with high and very high fitness costs (K103N with a fitness cost of 2.6%; T215Y with a fitness cost of 5.5%; M184V with a fitness cost of 6%; K65R with a fitness cost of 12%). Each box plot is based on 10,000 simulations of the stochastic model. Only simulations for which there is at least one transmission of the DRM from the index case are plotted. Medians are denoted by solid black lines while the top and bottom box edges denote the first and third quartile.

Figure S3: Box plots showing distributions of the transmission cluster size for four DRMs with high and very high fitness costs (K103N with a fitness cost of 2.6%; T215Y with a fitness cost of 5.5%; M184V with a fitness cost of 6%; K65R with a fitness cost of 12%). Each box plot is based on 10,000 simulations of the stochastic model. Only simulations for which there is at least one transmission of the DRM from the index case are plotted. Medians are denoted by solid black lines while the top and bottom box edges denote the first and third quartile. Note the logarithmic scale.

Figure S4: Cumulative distribution functions for the proportion of (non-null) transmission chains that terminate in each year for each strain (over 20 years). Strains shown are wild-type (A), K70R (B), Y181C (C), K219Q (D), L74V (E), D67N (F), M41L (G), K103N (H). The strains T215Y, M184V, and K65R are not shown as more than 95% of their non-null chains terminate within one year.

Figure S1:

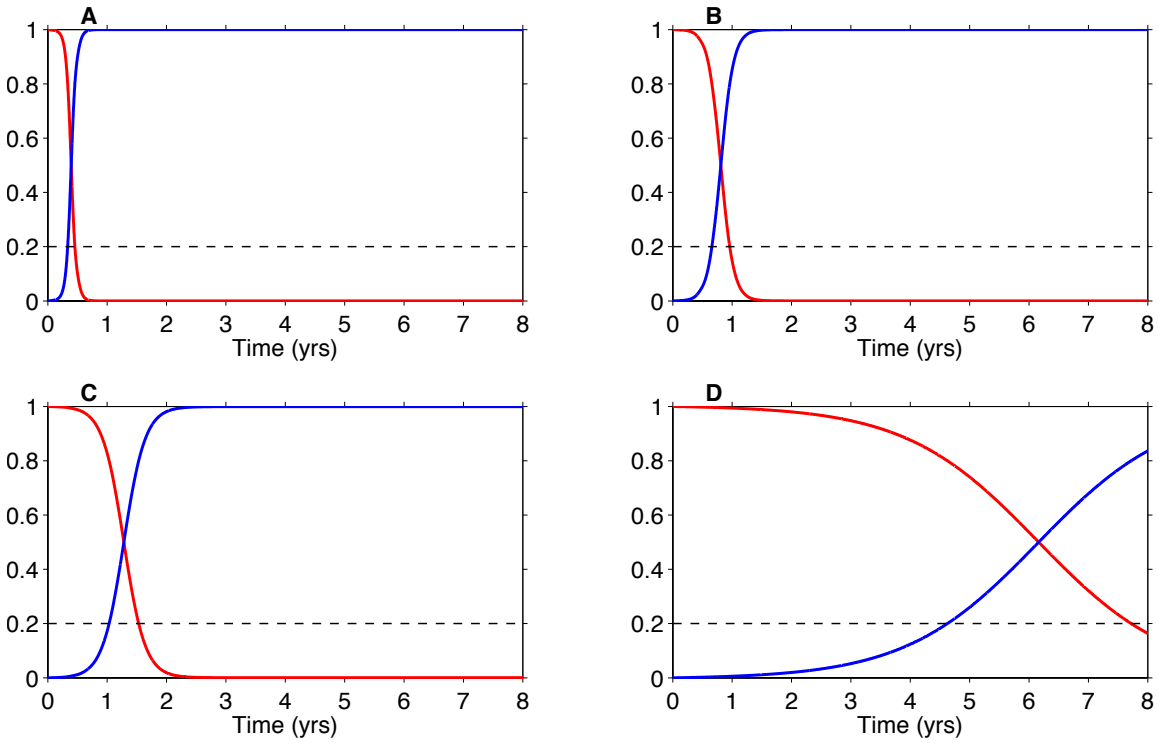


Figure S2:

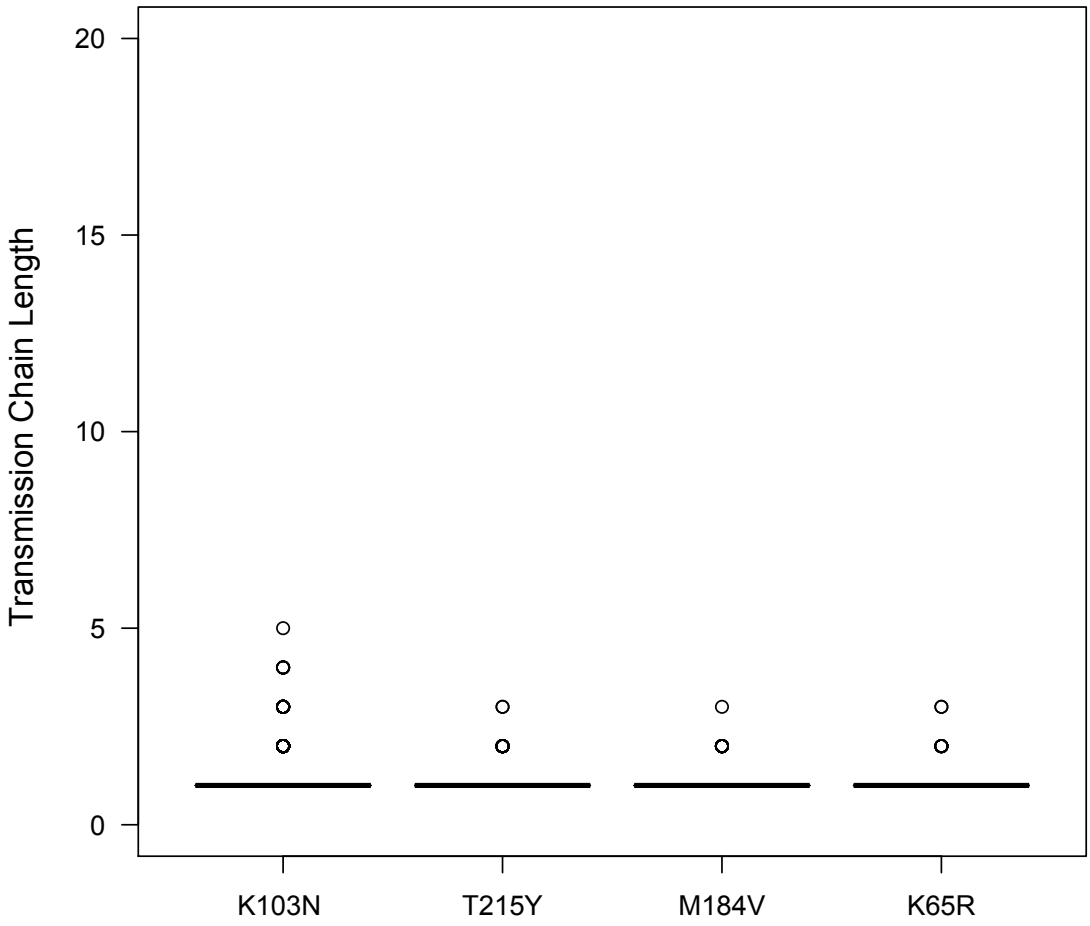

Transmission Chain Length
20
15
10
5
0
K103N
T215Y
M184V
K65R

Figure S3:

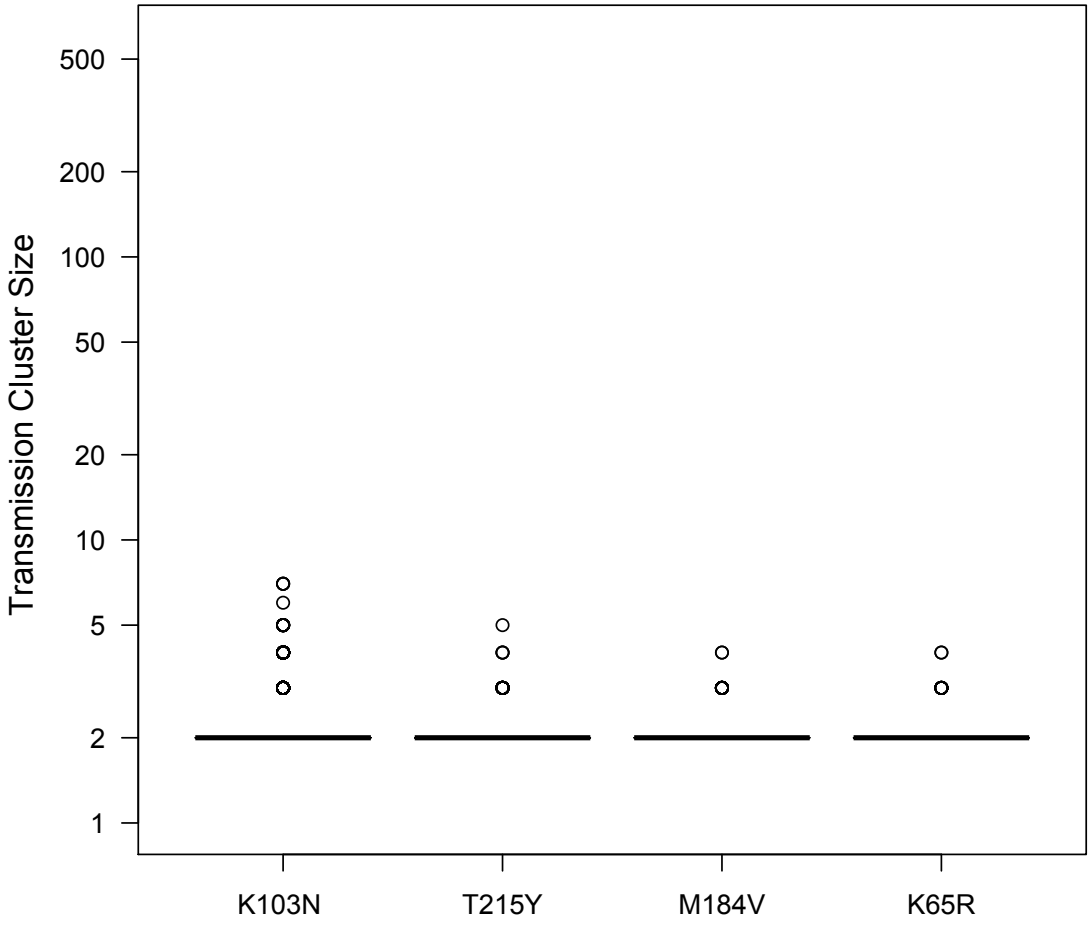
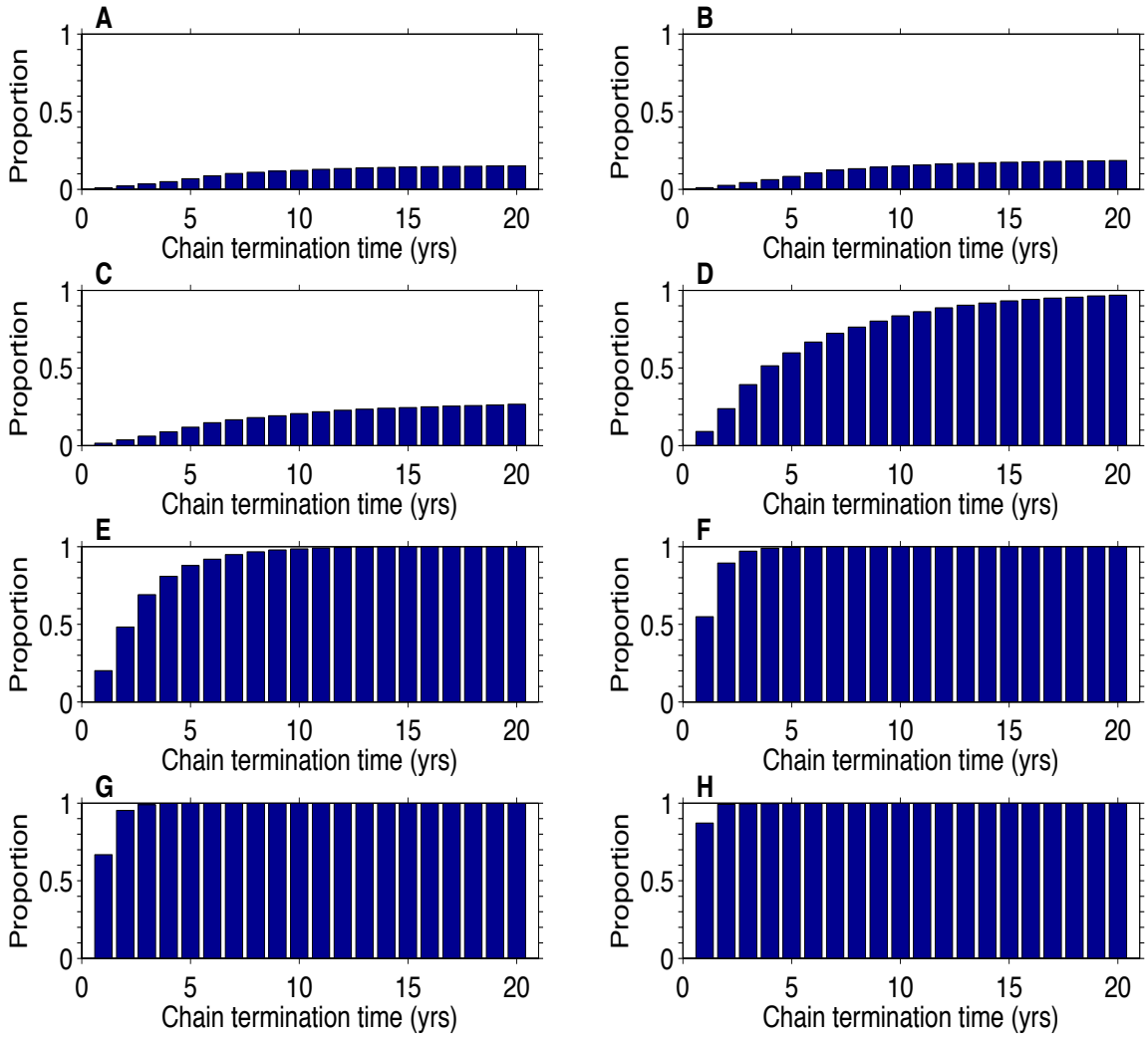

Transmission Cluster Size
500
200
100
50
20
10
5
2
1
K103N
T215Y
M184V
K65R

Figure S4:


A
B
C
D
E
F
G
H
Proportion
0
0.5
1
0
5
10
15
20
Chain termination time (yrs)