Science AAAS

Supplementary Materials for

Selection Bias at the HIV-1 Transmission Bottleneck

Jonathan M. Carlson[#], Malinda Schaefer[#], Daniela Monaco, Rebecca Batorsky, Daniel T. Claiborne, Jessica Prince, Martin J. Deymier, Zachary S. Ende, Nichole R. Klatt, Charles E. DeZiel, Tien-Ho Lin, Jian Peng, Aaron Seese, Roger Shapiro, John Frater, Thumbi Ndung'u, Jianming Tang, Paul Goepfert, Jill Gilmour, Matt A. Price, William Kilembe, David Heckerman, Philip J.R. Goulder, Todd M. Allen, Susan Allen and Eric Hunter

[#] Contributed equally to this manuscript

correspondence to:
JMC: carlson@microsoft.com
EH: ehunte4@emory.edu

This PDF file includes:

Notes S1 and S2
Figs. S1 to S4
Tables S1 to S3

Other Supplementary Materials for this manuscript includes the following:

Table S4

Supplementary Text

Note S1: The role of donor VL in transmission risk and selection bias

In Eq. 6 we noted that an overall increase c in the probability that any donor virus will be able to establish infection will lower the odds of transmission. Notably, transmission risk factors that increase risk by simply increasing exposure (reflected by n, the number of viruses in the quasispecies, which could be generalized to account for the number of exposures) were not present in Eq. 6, having canceled out in prior steps. Importantly, this cancelation was possible because of the assumption that the overall rate of transmission is small, leading to the approximation in Eq. 2. But how good is this approximation and what happens when transmission rates are high, violating the assumption that enables this approximation? This question is particularly relevant for donor VL, a well-established transmission risk factor. Although VL is known to increase transmission risk, there are at least two possible mechanisms: (i) high donor VL is a marker of increased viral fitness, consistent with the observations that *in vivo* replicative capacity correlates with VL (*18-21*) and that donor VL predicts early setpoint VL in linked recipients (*32-34*); and (ii) high donor VL simply increases the probability of transmission by increasing overall exposure. Our observation that increased donor VL reduces transmission selection bias suggests that fitness is at least one factor, but only if the approximation in Eq. 2 is valid.

To explore the validity of the approximation in Eq. 2 in the context of selection bias as expressed in Eq. 4, we implemented the formula for the exact rate of transmission under the binomial distribution in Matlab®, given by

$$r \equiv \Pr(T > 0) = 1 - (1 - p_a)^{f_a n}(1 - p_{\overline{a}})^{(1-f_a)n}$$

then plotted the relationship between r and the conditional probability that the founder virus included a virus of type a (fig. S3), given by

$$r_{a|r} \equiv \Pr(T_a > 0|T > 0) = \frac{1 - (1 - p_a)^{f_a n}}{1 - (1 - p_a)^{f_a n}(1 - p_{\overline{a}})^{(1-f_a)n}} = \frac{r_a}{r}$$

where $r_a = 1 - (1 - p_a)^{f_a n}$ is the rate of transmission of viruses of type a. If transmission is unbiased, then $p = p_a = p_{\overline{a}}$ and we can write

$$r_{a|r} \equiv \Pr(T_a > 0|T > 0) = \frac{1 - (1 - p)^{f_a n}}{1 - (1 - p)^{n}} \approx f_a$$

where the approximation again assumes a low rate of transmission. Thus, for small rates of transmission, transmission will be unbiased with respect to a if $r_{a|r} = f_a$; that is, if the probability that the founder virus includes a virus of type a is equal to the proportion of donor viruses that are of type a. For each simulation experiment, we first set the baseline selection bias $b = (p_a + c)/(p_{\overline{a}} + c)$, the frequency f_a of a in the donor quasispecies, and initial values of n and p_a such that r was low (initially 0.001). Setting $c = 0$, and $n = 1000$, we then manipulated r by increasing either n or c, and then plotted the relationship between r and $r_{a|r}$.

From this experimental setup, we observed that, for very high rates of transmission ($r > 0.5$), the odds that a virus of type a is in the founder population increases due to the increased odds of multiple-virus infection (fig. S3). For cases where selection bias favors the minority variant ($p_a < p_{\bar{a}}$; top rows of fig. S3), increasing the rate of infection by increasing c (that is, increasing the ability of each individual virus to establish infection; blue circles) causes selection bias to shrink toward zero, as seen by the convergence of $r_{a|r}$ toward f_a; this convergence is slowest when a represents 99% of the population, which is close to the mean value of f_a over all sites in our deep sequencing data. In contrast, increasing the quantity of donor viruses, n, has no effect on selection bias: until r is sufficiently high to make multiple-virus infection likely, the probability that the founder includes a remains near f_a. Similar results are observed when the selection bias favors the donor majority variant ($p_a > p_{\bar{a}}$; bottom rows of fig. S3), though here the effect of multiple-virus transmission is to induce a U-shape on the selection-bias curve. Notably, in each of these plots, the approximation in Eq. 3 (solid red and blue lines) closely tracks the exact probabilities for cases of single-virus infection, validating our use of this approximation in the models.

Thus, these simulations confirm that the overall donor viral population size n does not affect transmission selection bias in cases of single-virus transmission, while the effect of multiple-virus transmission is to increase the probability that a virus of type a is transmitted, regardless of the selection bias. In Fig. 3, we observed that increased donor VL predicts lower selection bias and that this effect is strongest for polymorphisms, an observation that is inconsistent with high VL simply increasing the rate of multiple-virus transmission. These results thus suggest that high VL is in this context primarily a marker for increased viral transmission fitness. That donor VL is a more important predictor for transmission in male compared to female recipients [(*27*); see also table S3] is consistent with our observation that female recipients generally have a lower selection bias than males and that donor VL has a stronger effect on selection bias among male recipients (Fig. 3). These observations suggest that the increased effect of donor VL on female-to-male transmission may primarily be the result of an increased barrier among male recipients that increases the importance of overall viral fitness (in effect, c is much smaller in males than females). Together with our observation that transmission index predicts transmission (Fig. 5), these models predict that reduction in overall viral fitness (for example, via drug resistance mutations, or as a result of immunological adaptation), will have a larger effect on female-to-male than on male-to-female transmission.

These results also suggest that therapeutic approaches to lowering VL without lowering VL fitness (for example, anti-retroviral therapy [ARV] in the absence of viral escape) will have no effect on transmission selection bias. Indeed, to the extent that ARV failure is caused by mutations that concomitantly weaken the virus, these models predict that high viral loads resulting from virologic failure will be correlated with *increased* selection bias, as in these cases higher VL will be a marker of *decreased* fitness in the absence of drug due to the escape mutations. Similar effects may arise in the case of elite controllers—if elite control is primarily indicative of an effective immune response and not a general inability of the virus to replicate.

Note S2: Independence of viruses

Thus far, we have assumed that the probability that two different viruses will establish infection are independent of each other. However, it has been reported that the distribution of multiple-virus infections exceeds what would be expected under independence (*1*). Indeed, the rate of multiple-virus infections ($\approx$ 10%) exceeds by 10 to 100 fold what would be predicted from the binomial distribution given observed rates of transmission (see next section). But what is the effect of non-independence on our modeling and on our conclusions?

The hypothesis that transmission is non-independent is supported by the relatively high frequency of multiple-virus infections. One possible mechanism for this non-independence would be a process in which the successful transmission of one virus makes it easier for another virus to break through the physical and immunological barriers (for example, if infection of one target cell causes the recruitment of other target cells). This mechanism would imply that the probability that no viruses establish infection remains $(1-p)^n$, and thus the overall probability of infection is still given by $r = \Pr(T > 0; n, p) = 1 - (1-p)^n$, and the observed low rates of transmission, $r < 0.01$, still allow the approximation $r \approx np$. The primary issue of non-independence is that the frequency of multiple-virus transmission will be non-negligible—roughly 10% in heterosexual cohorts. For these 10% of individuals, the odds that a virus of type a is in the founder population is no longer a simple function of f_a and $p_a/p_{\overline{a}}$ (Eq. 3), because the denominator needs to account for the probability that viruses of both type a and $\overline{a}$ are transmitted. In effect, the odds as stated in Eq. 3 will overestimate the true odds. However, while 10% of individuals are infected with multiple viruses, in our model setup, we group all viruses into two types: type a, comprising >75% of all donor viruses, and type $\overline{a}$. In this context, transmission of multiple viruses is irrelevant if they are all of the same type, as will be the case for the vast majority of sites for any particular instance of multiple-virus transmission. Furthermore, by filtering out instances where a mixture is observed in the recipient, we likely exclude many instances in which the founder population includes viruses from both the donor majority and minority variants. Thus, the overall proportion of sites in our modeling setup that includes viruses of both types is likely much less than 10%.

Nevertheless, what is the effect of non-independence on the small number of sites where this is relevant? With respect to viral fitness features (primarily those in Fig. 2), the effect will be to dilute the signal, making it harder to detect a selection bias between viruses of type a and $\overline{a}$. Thus, non-independence does not change our conclusions with respect to the existence of selection bias at the transmission bottleneck in general, nor the observation that these features are related to viral fitness in particular. With respect to the reduction of selection bias by risk factors, the effect will be to uniformly increase the probability that the founder population includes a virus of type a, regardless of whether selection bias favors or restricts a. Indeed, this can be observed among the high rates of transmission observed in fig. S3 (see Note S1), in which increasing n to very high rates of transmission increases the probability that a is in the founder population, regardless of $p_a/p_{\overline{a}}$. Importantly, while the risk factors considered here (sex, GUI and donor VL) likely increase the rate of multiple-virus transmission [as any transmission risk factor will under the binomial distribution, and as previously reported for GUI (*3*)], the observation that this effect is most extreme among variants with low cohort frequency, where $p_a \ll$

$p_{\overline{a}}$ (Fig. 3), argues that this effect is largely driven by a reduction in selection bias, not by very high rates of multiple-virus transmission.

The expected rate of multiple infection under the binomial distribution is approximately the rate of transmission

We asserted in the previous section that the pattern of multiple-virus infection suggests that transmission is non-independent. This observation was first made by Abrahams and colleagues, who argued from the Poisson distribution that the observed distribution of the number of virus genotypes per infection was not consistent with the independence assumption (*1*). Here, we briefly re-derive the Abrahams result using the binomial distribution and show that, under independence, the proportion of founder populations with more than one virus will be approximately the same as the overall rate of infection. Since the observed rate of multiple-virus infection greatly exceeds that predicted by the binomial distribution, we conclude that transmission is characterized by non-independence among individual viruses, such that the transmission of one virus particle increases the probability that other virus particles will be part of the founder population.

Equation 2 provides the exact probability (assuming a binomial process) that at least one virus establishes infection: that is, the probability, or rate r, of infection per exposure incident. Similarly, the binomial distribution provides the exact probability that a single virus establishes infection as

$$\Pr(T = 1; n, p) = np(1-p)^{n-1}$$

Thus, the conditional probability that productive infection was established by a single virus is

$$\Pr(T = 1 | T > 0; n, p) = \frac{np(1-p)^{n-1}}{1-(1-p)^n}$$

From the approximation in Eq. 2, we see that the probability that a successful transmission event involves multiple viruses is approximately

$$\begin{aligned}
\Pr(T > 1 | T > 0; n, p) &= 1 - \frac{np(1-p)^{n-1}}{1-(1-p)^n} \\
&\approx 1 - \frac{n\frac{p}{1-p}(1-np)}{np} \\
&= 1 - \frac{1-np}{1-p} \\
&= 1 - \frac{1-r}{1-\frac{r}{n}} \\
&\approx r
\end{aligned}$$

where we have again used the assumption of small per-virus transmission probability, p<<1. That is, the proportion of infections established by multiple variants will be approximately equal to the proportion of sex acts that result in any infection. Note that while the above approximations allow for an intuitive understanding of the relationship between the probability of transmission and the conditional probability of transmitting multiple viruses, exact probabilities are easily computed by statistical software and reveal that the above approximations *overstate* expected transmissions that involve multiple viruses, especially when the probability of transmission exceeds 20% (data not shown).

Note that, for small transmission rates, the precise values of n and p are irrelevant: np is the statistic of interest, as it determines the transmission probability r. Furthermore, note that this result holds under models in which individual viruses have different probabilities of establishing infection (in which case p represents the mean over the entire population of viruses). For example, it is well known that some couples represent high infection risk, while others represent low infection risk (*27, 28*). In these scenarios, the above model can be extended to a probabilistic hierarchical model, with the same result that the expected number of multiple-virus transmissions will be approximately equal to the overall rate of transmission. Briefly, if high risk couples have a rate of transmission of $r_{\uparrow}$ while low risk couples have a rate of transmission of $r_{\downarrow}$, then the overall rate of transmission in the population will be the average rate of transmission, weighted by the proportion of individuals in the high ($f_{\uparrow}$) or low ($1 - f_{\uparrow}$) risk groups. Similarly, because within each group the rate infection will be approximately the same as the rate of infections involving multiple viruses, the conditional probability that productive transmission results in multiple transmitted viruses will also be the weighted average $r \approx f_{\uparrow}r_{\uparrow} + (1 - f_{\uparrow})r_{\downarrow}$. That is, the overall rate of transmission will still be approximately equal to the overall proportion of transmissions that involve multiple transmitted viruses.

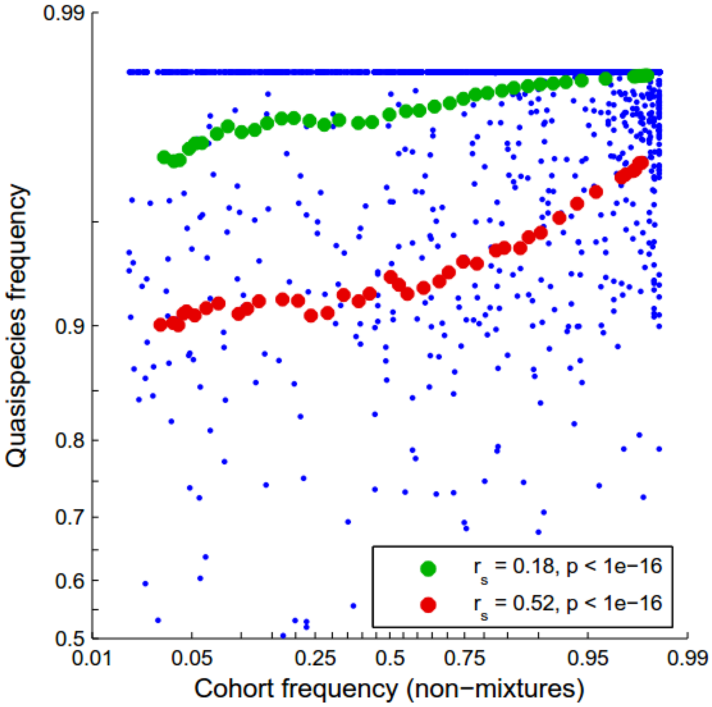

Fig. S1. Cohort frequencies of donor majority variants correlate with the frequency of those variants in the donor quasispecies.

Deep sequencing (454) was performed on 5 donors to estimate the quasispecies frequencies of dominant variants, as called from population sequences. The mean quasispecies frequency computed as a sliding window over cohort frequency (window size of 1 unit in log-odds space) is plotted in Green (all sites) and Red (all sites with observed quasispecies variation in the donor). Cohort and quasispecies frequencies are computed as smoothed log-odds scores with smoothing factor $q = 1/50$.

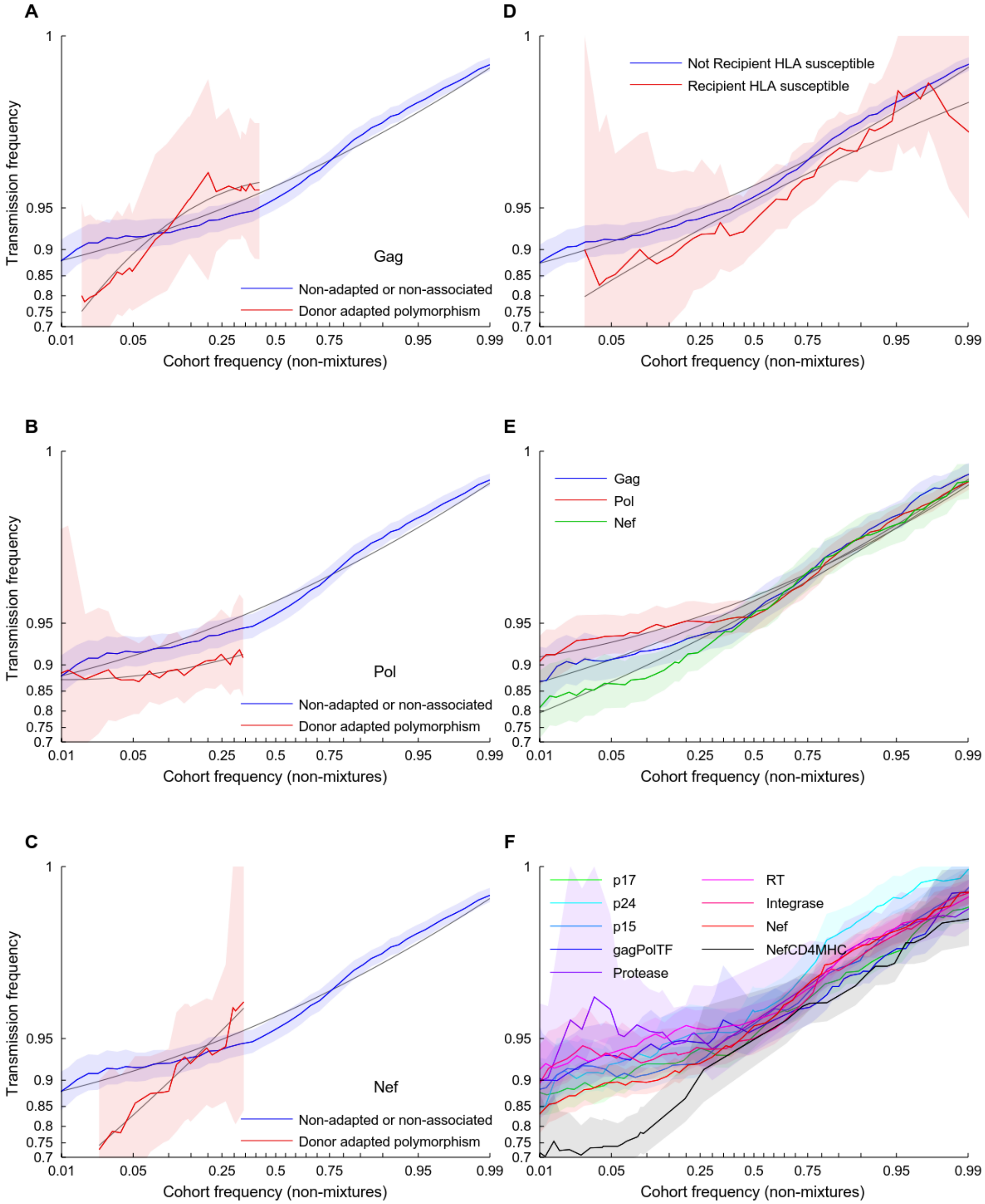

Fig. S2. Additional features that impact selection bias.

(**A-C**) Selection bias against transmission of variants that are consistent with escape from donor HLA alleles in (A) Gag, (B) Pol, and (C) Nef. (**D**) Residues that are susceptible to recipient HLA alleles—meaning they represent an un-escaped amino acid residue that is linked to at least one recipient HLA allele—are less likely to be transmitted. However, because transmission is defined as differences between recipient and donor sequences, as measured a median of 46 days after transmission, these curves could represent rapid escape in the recipient. (**E**) Differences among proteins Gag, Pol and Nef, or (**F**) among protein domains. Although these differences were significant (see Table 2), no differences were observed when correcting for donor quasispecies frequency among the 5 couples for whom deep sequencing was available (Table 1), suggesting that these protein-specific difference may primarily result from differences in mean quasispecies frequencies of variants for these proteins. Nevertheless, these protein domains are included as covariates in all multivariable models to correct for confounding. Nef functional CD4 and MHC downregulation domains are taken from (*49*).

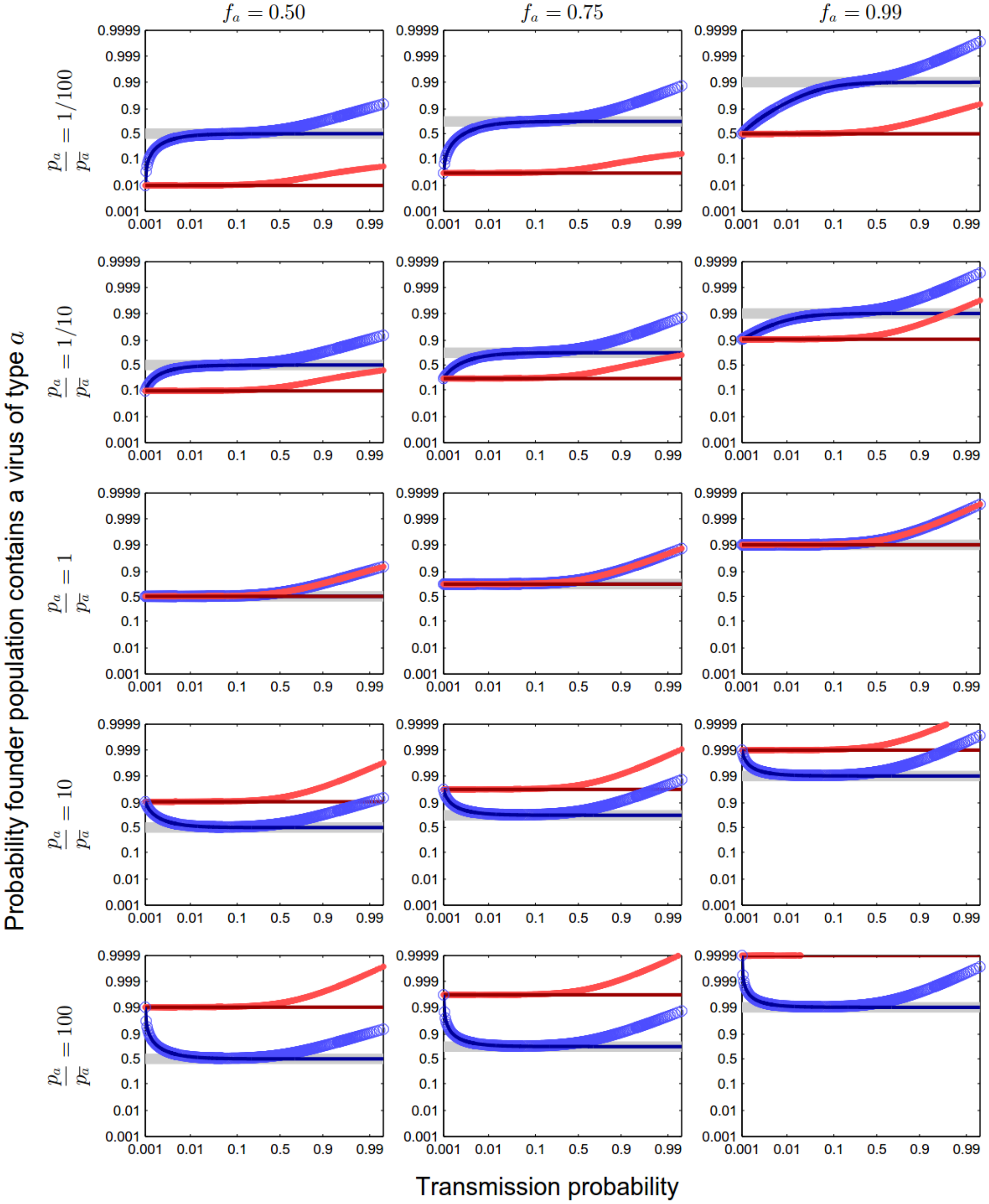

Fig. S3. Simulation of selection bias versus transmission probability.

The exact binomial probability mass function was used to explore the relationship between the probability of transmission, r, and the conditional probability that the transmitted founder virus population contains at least one virus of type a ($r_{a|r}$). For a range of 5 bias values ($p_a/p_{\overline{a}}$) and a range of donor quasispecies frequencies for a (f_a), we plotted the conditional transmission probability of a as a function of the overall transmission probability r. For each plot, we set the donor viral population size to $n = 1000$, then solved for p_a to achieve a transmission probability of 0.001, satisfying bias, f_a, and n. We then increased the rate of transmission r, either by increasing the overall donor population size, n (red dots), or by increasing p for all viruses by adding a constant c to both p_a and $p_{\overline{a}}$ (blue circles). The red and blue solid lines indicate the predicted conditional transmission probability of a, as estimated by the approximation in Eq. 3. A reduction in selection bias is here visualized as the convergence of the conditional transmission probability toward f_a (gray line).

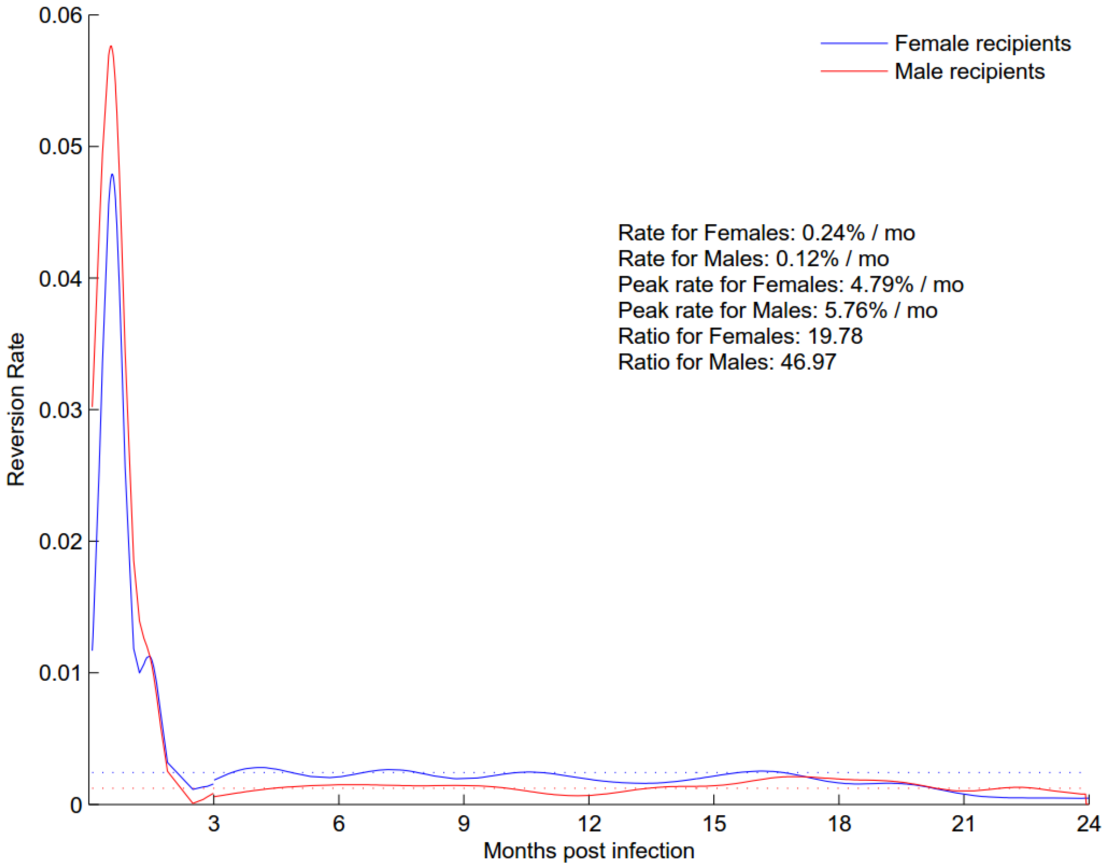

Fig. S4. Empirical reversion rates.

Empirical reversion rates of donor polymorphisms to non-mixture consensus were estimated using a kernel smoothing function, as implemented in the Matlab statistics package, using a Gaussian kernel with widths of 1 week and 1 month. The plot shows the curves with 1 week smoothing for <3mo and 1 month smoothing for >3mo (the larger smoothing window accommodates the sparser sampling times). Dotted lines show the mean reversion rates in the 3-12 month interval. Reversion rates are an order of magnitude higher in the first three months of infection compared to the following 18 months; steady state reversion rates are lower in males compared to females, whereas initial reversion rates are higher in males compared to females. These data are thus consistent with an initial selection bias and are not likely artifacts of early reversion, as further supported by the ability to estimate the odds of transmission of viruses and virus populations (Fig. 5). Compare to Fig. 4, which shows cumulative reversion. Note that Fig. 4 measures reversion times relative to the first available sample date in the recipient; here, reversion times are measured relative to the estimated date of infection.

Table S1. Clinical characteristics of the cohort.

	Transmitting	Non-Transmitting
N	137	181
Male (%)	62 (45%)	87 (48%)
Male recipient* GUI (%) [missing]	17 (29%) [3]	10 (12%) [1]
Female recipient GUI (%) [missing]	27 (38%) [3]	15 (16%) [2]
Male donor† $\log_{10}$ VL, median [IQR‡]	5.2 [4.7,5.7]	5.0 [4.3,5.3]
Female donor $\log_{10}$ VL, median [IQR]	4.8 [4.3,5.3]	4.3 [3.6,4.9]
ETI§, median [IQR]	46.0 [42.0,60.5]	

*In the case of non-transmitting couples, the "recipient" refers to the seronegative partner. †In the case of non-transmitting couples, the "donor" refers to the seropositive partner. ‡Interquartile range. §Estimated time between infection and first available sample.

Table S2. Reversion of donor polymorphisms transmitted to recipients

Feature	Ln(HR)[*]	*P* value†	Transmission‡	
Cohort frequency§	-0.28	0.016	+	Viral fitness features[**]
# Covarying sites	-0.10	0.031	+	
Donor escape	0.99	0.003	−	
Structural frequency§	0.01	0.496	+	
Donor log$_{10}$ VL	-0.16	0.131	+	
Is male-to-female	0.58	0.016	+	Recipient susceptibility features††
Recipient is GUI male	0.17	0.406	+	
Is escape to consensus¶	1.45	0.004		
p17	-0.16	0.315		
p24	0.11	0.409		
Protease	-1.92	0.196		
Reverse transcriptase	-1.42	3.1×10^{-6}		
Integrase	-1.77	0.215		
Nef CD4/MHC domains	1.40	0.207		

[*] The hazard ratio of reversion from polymorphism to consensus, for sites in which a polymorphism was present in both the donor and recipient, were estimated using a Cox proportional hazards model. Only sites with available protein structures were used in the model; all sites were used in Fig. 4. †*P*-values were estimated using a multilevel bootstrap (1000 replicates) to estimate the standard error for each parameter. ‡The effect of the feature on odds of transmission (Table 2) is indicated: +, the feature generally increases odds of transmission; −, the feature generally decreases the odds of transmission. §Because we are tracking reversion to consensus, and not any mutation away from the polymorphism, observed cohort and predicted structure frequencies are here represented as the negative standardized log-odds of the respective measure for the cohort consensus at that site. ¶A binary variable indicating whether a mutation to consensus is consistent with escape from recipient HLA alleles, and thus may more likely represent immune escape than reversion. [**] Viral fitness features are expected to elicit opposite effects on transmission and reversion: amino acids with high odds of transmission will have low rates of reversion and vice versa. ††Recipient susceptibility features are expected to elicit concordant effects on transmission and reversion: individuals who have low selection bias will have high overall odds of transmitting the dominant donor variant; those variants will in turn revert faster because they on average have lower fitness.

Table S3. Transmission index of the seroprevalent partner is predictive of transmission

Feature	Ln(OR)[*]	*P* value
Offset	0.64	0.031
Transmission index†	1.28	0.047
Is male-to-female	1.11	0.705
Donor‡ $\log_{10}$ VL (M2F)§	1.47	0.031
Donor $\log_{10}$ VL (F2M)	2.18	4.4×10^{-4}
Recipient¶ has GUI (M2F)	1.00	0.986
Recipient has GUI (F2M)	3.41	0.010

[*] Model was fit using logistic regression. Dependent variable was whether the seroprevalent individual had transmitted to their partner. Compare to Fig. 5C. †Transmission index is standardized (zero mean, unit variance) for comparison purposes. ‡Or the seropositive partner in the case of NYT couples. §Donor VL and recipient GUI were given separate parameters for male-to-female (M2F) and female-to-male (F2M) couples. ¶Or the seronegative partner in the case of NYT couples.

Table S4. HLA associations, covariation associations and structural energy estimates (xls).