Supporting Information

Boeras et al. 10.1073/pnas.1103764108

SI Methods

DN A C

Using the Test of Random Transmission analysis, given n = 8 transmission pairs, we address the question of whether or not the transmitted sequence is randomly distributed between rakes and nonrakes. All sequences were sampled from the GT. We pick a distance threshold, D, and define a cluster as any subset of sequences within distance D from one another. In other words, any two sequences s_1 and s_2 are within the same cluster if they differ by less than the distance D. We compute raw pairwise distances from all available sequences from any given donor and then assign sequences to clusters using the cluster.dist procedure in R.

For any given *D*, and for each transmission pair i = 1, ..., N, we calculate the proportion, p_i , of sequences outside a cluster. Let $P_D(n)$ be the probability that *n* donors transmit a sequence outside a cluster. Then,

$$\begin{aligned} P_D(0) &= \prod_{i=1}^N (1 - p_i) \\ P_D(1) &= \sum_{j=1}^N \prod_{i \neq j} p_j (1 - p_i) = P_D(0) \sum_{i=1}^N \frac{p_i}{1 - p_i} \end{aligned}$$

N7

and so on. Let g_D be the probability generating function defined by

$$g_D(x) = \sum_{i=1}^N P_D(i) x^i.$$

One can see that

$$g_D(x) = \prod_{i=1}^{N} [(1-p_i)+p_i x].$$

We can thus calculate $P_D(n)$ using the formula

$$P_D(n) = \frac{1}{n!} \frac{d^n}{dx^n} [g_D(x)]|_{x=0}$$

and Mathematica to compute $P_D(n)$ for n = 1, ..., 9 and D = 1, ..., 10 according to the above formula.

For each D, we call n_{obs} the number of donors that transmitted sequences outside a cluster and compute

$$p(D) = \sum_{n=n_{obs}}^{N} P_D(n).$$

The quantity p(D) is the overall probability across eight donors that the observed number of sequences transmitted outside a cluster (i.e., not in a rake) is significantly different from what we would observe if transmission were equally likely among all sequences.



Fig. S1. *Highlighter* and phylogenetic analysis reveal distinct GT populations. Aligned Env V1–V4 nucleotide sequences for transmission pairs were analyzed by the Los Alamos *Highlighter* tool. (*A–E*) *Highlighter* plots aligned to the phylogenetic trees. Tick marks indicate nucleotide differences from the recipient consensus sequence (blue open square). Nucleotide differences are color-coded and marked according to their genetic location along the length of V1–V4. Colors are as follows: green, A; red, T; orange, G; blue, C; gray, gaps. Arrows point to those variants in the blood (green) and GT (red) most closely related to the transmitted founder virus. Bootstrap values >70% are shown.

	Full-length (HX				6007478	3)	Partitioned at GARD breakpoints			
	Identical sequences present			Identical sequences removed						
Donor ID	s	$p(s' \le s)*$	$q(s' \leq s)^\dagger$	s	$p(s' \le s)$	$q(s' \le s)$	Region	s	$p(s' \le s)$	$q(s' \le s)$
ZM1149F [‡] BL vs. GT TP1	8	0.0016	0.0045	6	0.4684	0.2382	5′: 66006798	6	0.5797	0.2724
							mid: 67997121	7	1.0000	0.4089
BL TP1 vs. GT TP2	7	0.0347	0.0366	7	0.0340	0.0366	3': 71227478 5': 66006798 Mid: 67997136	5 8 6	0.3766 0.7049 0.4974	0.2104 0.3106 0.2497
BL TP1 vs. GT TP3	4	0.0448	0.0404	4	0.0408	0.0376	3': 71377478 5': 66006798 Mid: 6799, 7122	3 3 ⊿	0.0026 0.0354	0.0061 0.0366 0.1395
BL TP1 vs.	9	0.0087	0.0123	9	0.0161	0.0206	3': 71237478 5': 66006794	4 9	0.0694 0.2586	0.0529 0.1508
GT TP1–TP3							Mid: 67957121 3': 71227478	7 6	0.2184 0.0039	0.1395 0.0067
ZM1862F ^s BL TP1 vs.	4	0.0458	0.0404	4	0.0594	0.0471	5': 66006925	4	0.2349	0.1419
BL TP1 vs. GT TP3	5	0.2126	0.1382	4	0.5894	0.2724	5': 66007026 3': 70277478	4 2	0.4644	0.2382 0.0551
BL TP1 vs. GT TP2 + TP3	4	0.0060	0.0092	4	0.0052	0.0082	5′: 66006985 3′: 69867478	4 2	0.0637 0.0038	0.0495 0.0067
ZM323F BL TP1 vs. GT TP1	14	0.6137	0.2766	14	0.5918	0.2724	5′: 66006788 Mid: 67897079	12 7	0.9321 0.4582	0.4018 0.2382
BL TP1 vs.	15	0.5516	0.2668	13	0.5366	0.2628	3': 70807478 5': 66006828 Mid: 6829, 7130	8 10 6	0.2081	0.1376
BL TP1 vs.	11	0.0003	0.0007	11	0.0615	0.0271	3': 71317478 5': 66006828	8 10	0.4569 0.3898	0.2382 0.1085
GT TP3 BL TP1 vs.	15	0.0093	0.0127	15	0.0104	0.0137	Mid: 68297231 3': 72327478 5': 66007016	7 9 15	0.3925 0.3472 0.0371	0.1085 0.1032 0.0366
GT TP1–TP3 ZM1165F							3': 70177478	16	0.5335	0.2628
BL TP1 vs. GT TP1	7	0.5741	0.2724	6	0.0050	0.0082	5′: 66006926 3′: 69277478	5 6	0.0327 0.3585	0.0366 0.2031
BL TP1 vs. GT TP2	8	0.0484	0.0417	8	0.1089	0.0771	5': 66006784 Mid: 67857177	6 6	0.6679 0.2582	0.2976 0.1508
BL TP1 + TP2 vs. GT TP1–TP3	20	0.0047	0.0045	16	0.0256	0.0179	3': 71787478 5': 66006989 3': 69907478	6 11 14	0.2076 0.0112 0.1505	0.1376 0.0086 0.0559

Table S1.	SM compartmentalization	on analvsis of I	ongitudinal samples
	• • • • • • • • •	J • • •	

The SM test was performed on longitudinal samples from chronically infected donors (Donor ID) using full-length envelope V1–V4 sequences (6600–7478, HXB2 nucleotide coordinates), full-length with identical sequences removed, and alignment regions (region) partitioned at Genetic Algorithms for Recombination Detection (GARD) recombination breakpoints with identical sequences removed. The SM analysis used observed number of migration events (s) to determine the *P* value (p). TP, time point.

*SM *P* value is the proportion of relabeled trees with as many or fewer migration events (s) as observed: $p(s' \le s)$, where s' denotes results from 10,000 compartment-label permutations on the fixed tree, shown in bold, where P < 0.05.

[†]Storey and Tibshirani (1) q value from SM *P* value, shown in bold, where q < 0.05.

[‡]Results from longitudinally sampled subjects ZM323F, ZM1165F, and ZM1149F are shown for both intra-visit (BL TP1 vs. GT TP1) and inter-visit (e.g. BL TP1 vs. GT TP2) sample comparisons.

[§]Comparison of inter-visit samples, i.e. blood time point 1 (BL TP1) vs. genital-tract time points 2 and 3 (GT TP2 and TP3.).

1. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440-9445.

	D BL	D GT	BL source	GT source
F-DONORS				
RW36	10	10	PL	CA
ZM201	5	6	PL	SW
ZM216	4	3	PL	SW
ZM221	0	6	PB	SW
ZM238	4	3	PB	CA
ZM292	1	6	PL	CF
M-DONORS				
RW56	3	>15	PB	CA
ZM242	3	>15	РВ	CA

 Table S2.
 Nucleotide differences between donor and recipient sequences

D, donor; F, female; M, male; BL, blood; GT, genital tract.

PNAS PNAS