Cell

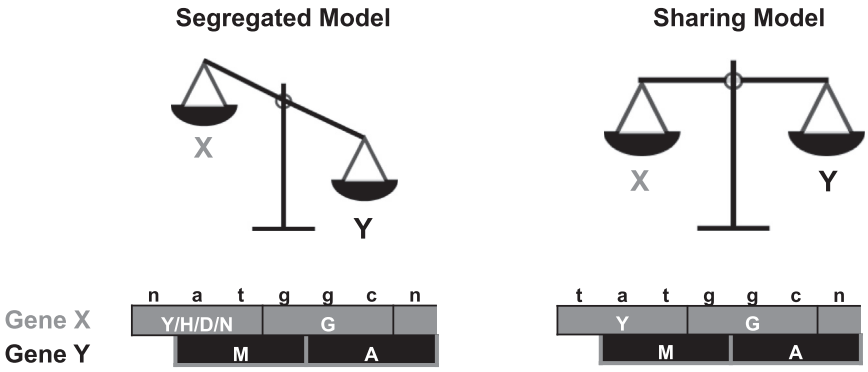

Figure S1. Models for Overlap Evolution, Related to Figure 1

In the segregated model one gene drives the evolution of the other. Strict requirements of M and A in a dominant gene Y allow several equivalent changes in the sequence of accommodating gene X (D/H/N/Y) but occasionally fix particular amino acids (G in position 2 of X). In the sharing model, both genes have strong requirements of particular amino acids constraining the DNA sequence so that even synonymous codon mutations are not allowed. The sharing model is marked by stronger conservation of both proteins at the amino acid level while the segregated model allows for comparatively greater diversity.

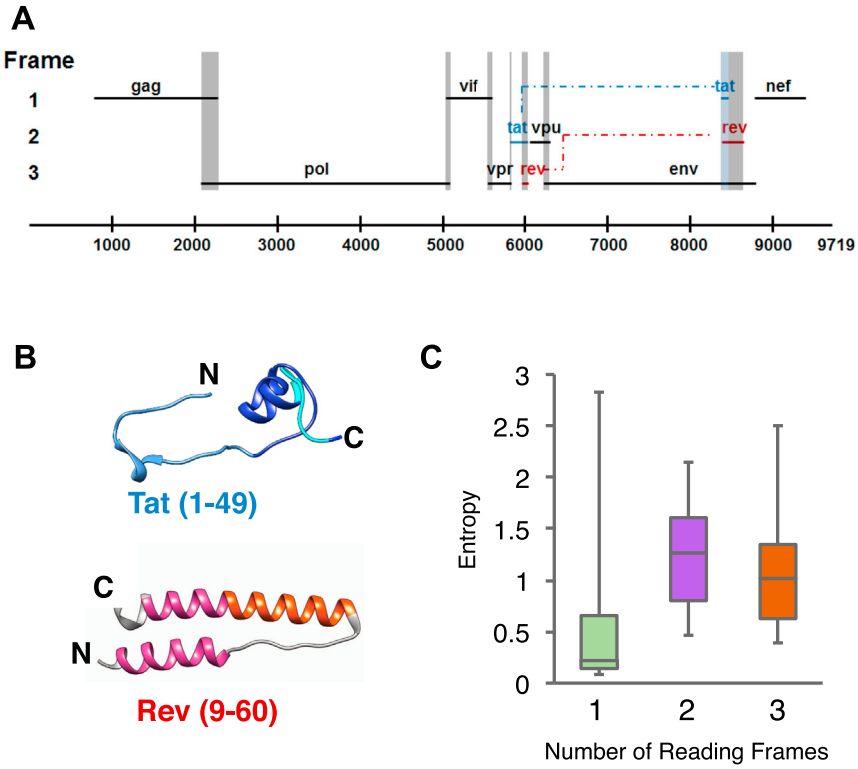

Figure S2. Related to Figure 1

(A) Organization of the HIV genome by reading frame. Two-frame overlaps are shaded with gray boxes and the three frame overlap with light blue.
(B) Partial 3D structures of Tat and Rev. Motifs are colored as labeled in Figure 1A. PDB structures 3MI9 (Tat) and 3LPH (Rev) were used.
(C) Genome-wide entropy comparison between 1,2 and 3-frame regions.

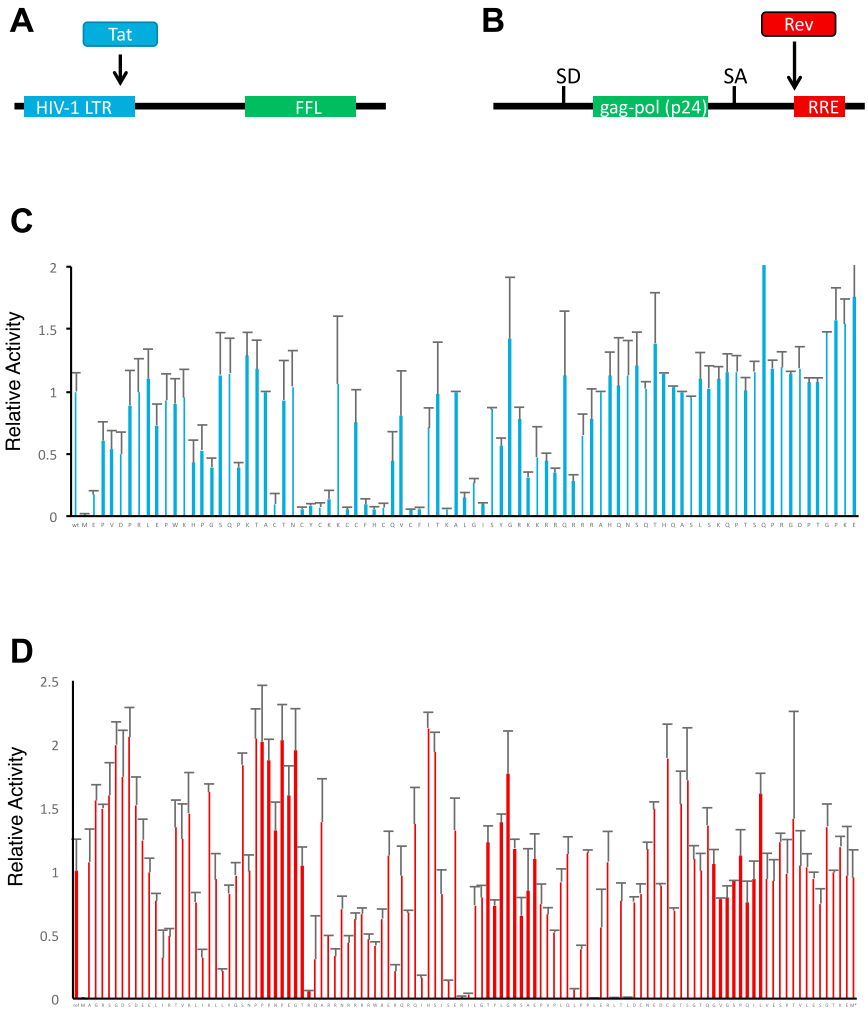

Figure S3. Related to Figure 2

(A) Schematic for the alanine scan for Tat and Rev (B). Tat activates transcription of firefly luciferase under the control of the HIV LTR.

(B) Rev allows expression of intronic p24 by exporting the message before splicing can occur. SD = Splice Donor, SA = Splice Acceptor.

(C) Tat activity for all mutants was reported as mutant induction of firefly luciferase relative to reference induction.

(D) Rev activity for all mutants was reported as mutant induction of intracellular p24 levels relative to reference induction. M* is an N-terminally Strep-tagged M1A mutant. Error bars represent standard deviations from five (Tat) or three (Rev) biological replicates.

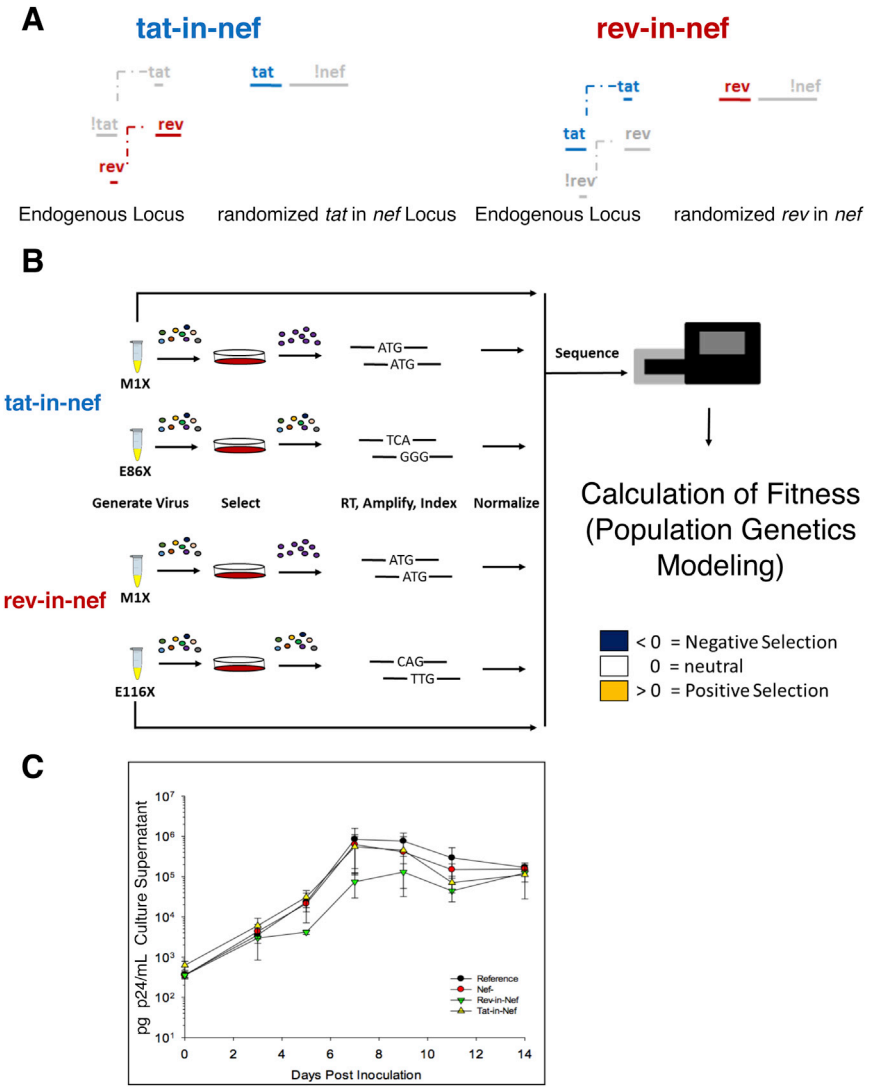

Figure S4. Related to Figures 3 and 4

(A) Creation of uncoupled viruses involved mutations (ablation of start codon and introduction of downstream stops) to the endogenous frame. A synonymously substituted codon version of the knocked out gene was then introduced into the *nef* locus. The introduced gene was randomized equally among all codons at a single position.

(B) Schematic for selection experiment. Each library, representing a randomized and unconstrained version of Tat (tat-in-nef) or Rev (rev-in-nef) at a single position, was competed against itself, raised in 293 cells, and then competed against itself in SupT1 cells. Viral supernatants were harvested, genomic RNA isolated, and reverse transcribed. Amplicons targeting the randomized site were generated, indexed, normalized and then pooled and sequenced. Relative fitness was reported as a logarithmic ratio of the allele frequencies post- and pre-selection.

(C) Replication curves for the TRx, xRT, TxR and NL4-3 viruses demonstrate effects on viral kinetics after removing the overlap. These curves were also used to generate models for our population genetics simulations.

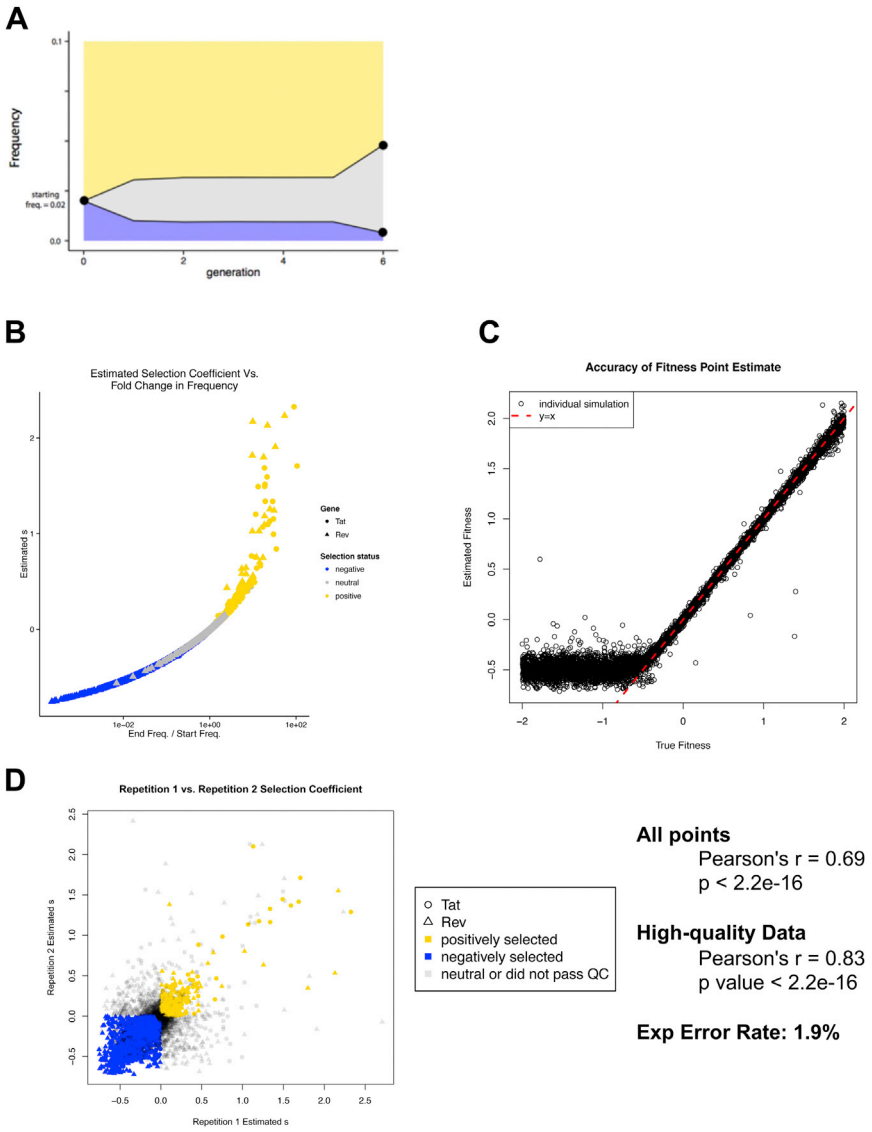

Figure S5. Related to Figures 3 and 4

(A) An illustration of the neutral simulations for a hypothetical allele with a starting frequency of 0.02, an ending read depth of 500 reads, and an amino acid identity of Arginine. The gray area depicts the range of trajectories that this allele could take if it were neutral. If an ending allele frequency were observed to be above or below this neutral expectation, it is deemed positively or negatively selected, respectively. The black dots indicate the upper and lower bounds for the ending allele frequency that would still be considered neutral. These upper and lower bounds correspond to relative fitness values of 0.380 and −0.699, respectively, which means neutrality cannot be rejected for any observed fitness value that resides between this interval.

(B) The observed distribution of fitness values for alleles found to be under negative (blue), neutral (gray), or positive (gold) selection. Neutral alleles were sometimes found to have relatively extreme fitness estimates (left and right tails of gray distribution). Likewise, alleles under significant positive or negative selection were sometimes found to have fitness estimates close to zero (right tail of blue distribution, and left tail of gold distribution).

(C) Correlation between the estimated fitness and the true fitness under our simulation framework. Each point corresponds to one simulation.

(D) Correlation plot of alleles between biological replicates. Alleles that have strong experimental evidence of selection show high repeatability. Correlation coefficients are shown for the whole data-set and those passing our QC criteria (note that these criteria have inherent biases, such as requiring the sign of the selection coefficient to be consistent, toward increasing the correlation coefficient). The experimental error rate (amino acids that appear to mutate outside the randomization site) is ~2%. Note that this error rate is calculated on the amino acid level and does not consider synonymous mutations, or multiple mutations within the same codon.

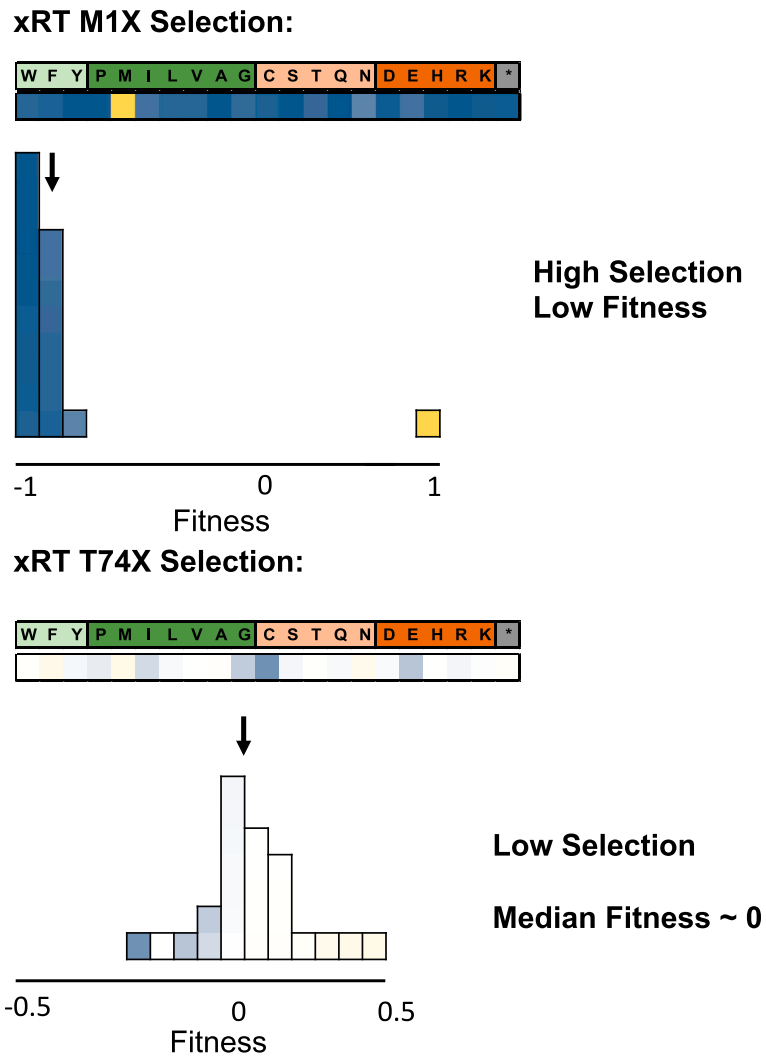

Figure S6. Median Selection Coefficient as an Indicator of Strength of Selection, Related to Figures 5 and 6

We take all possible alleles for a given position (e.g., tat-in-nef M1X or tat-in-nef T74X) and calculate the median for the distribution of their fitness effects (shown in the histogram). Sites like M1X have strong preference for and against particular alleles (strong selection) and a low median fitness. Sites like T74X with near equal preference for most alleles display a more normal distribution and have a median fitness near zero.

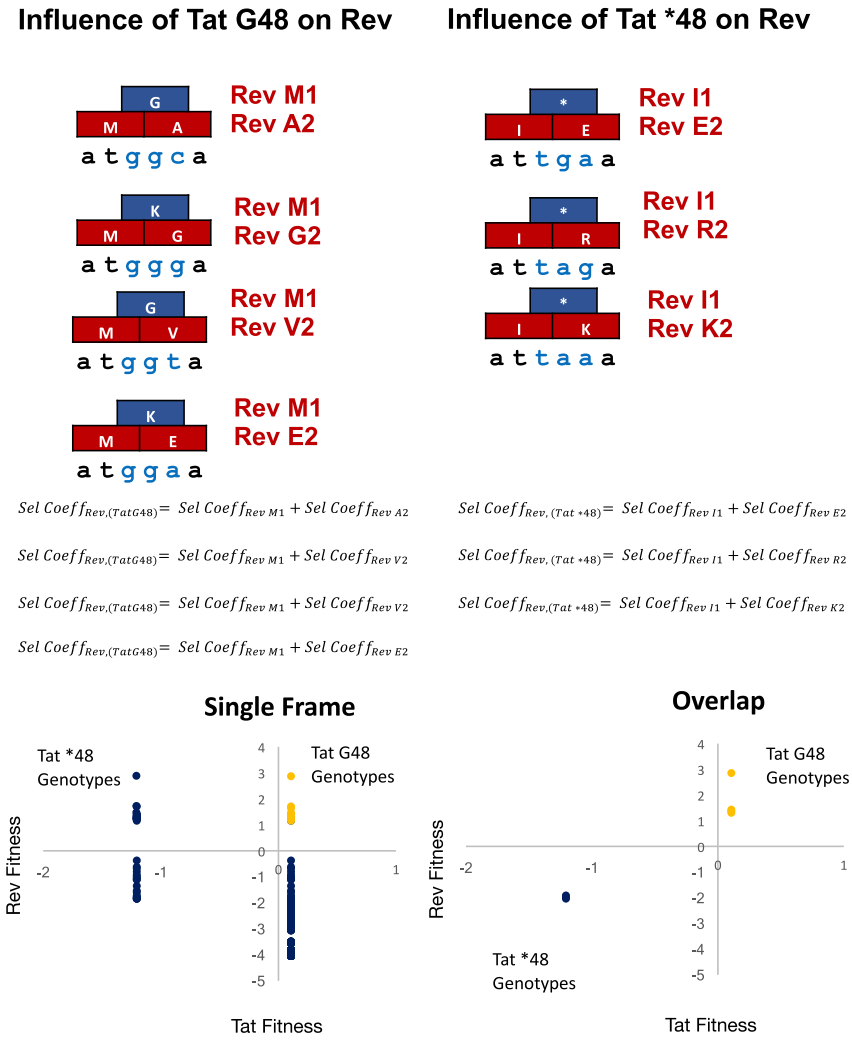

Figure S7. Inferring the Effects of Mutations in the Context of the Overlap, Related to Figure 7

If one gene encodes a particular allele (i.e., Tat G48 or Tat *48) the overlapped gene is restricted to a subset of alleles based on the synonymous codons of the fixed gene. As Tat and Rev are both essential for viral replication only viral genotypes which harbor fit alleles for both genes produce fit viruses. We can explicitly calculate the fitness of every genotype in both the single frame and overlapped context for every allele in every position of each gene. Pictured in this example are explicit calculations for the effect of Tat G48 or Tat *48 on Rev.