Supplemental material

RT-PCR, PCR (outer and inner), sequencing and vector primer sequences

gag primer:

gag outer rev: 5'-GCCTGTCTCTCAGTAC-3' (RT-PCR, PCR outer)

gag fw: 5'-CGACGCAGGACTCGGCTTGCTG-3' (PCR outer)

BssHIIfw: 5'-TGCTGAAGCGCCCGCACGGC-3' (PCR inner, seq)

p17rev: 5'-CAAAACTCTTGCCTTATGG-3' (PCR inner, seq)

p17fw: 5'-GCTAAACACAGTGGGGGGACATC-3' (seq)

p24_5I: 5'-ATGGTACATCAGGCCATATCACCTA-3' (seq)

pol primer:

CET43rev: 5'-CCCCACTAACTTCTGTATGTCATTGACAGTCC-3'

(RT-PCR, PCR outer, PCR inner)

OutZhang: 5'-GAAGCAATGAGCCAAGTAACAAAT-3' (PCR outer)

GagRec: 5'-TGTGGCAAAGAAGGGCACACAGCCAGAAATTGCAG-3'

(PCR inner, seq)

pol1fw: 5'-TATTAGAAGAAATGAGTTTGCCAGGAAG-3' (seq)

pol2rev: 5'-CATTGTTTAACTTTTGGGCCATCCATTCCT-3' (seq)

pol3fw: 5'-AGTATAAACAATGAGACACCAGGGAT-3' (seq)

pol4rev: 5'-TTTTTGTCTGGTGTGGTAAGTC-3' (seq)

C2-C4 *env* primer:

Seq2: 5'-TCCTCCATATCTCCTCCTCCAGGTC-3' (RT-PCR, outer PCR)

Seq3: 5'-TATGGGATCAAAGCCTAAAGCATG-3' (outer PCR)

Seq5: 5'-GTCAACTCAACTGCTGTTAAATGGC-3' (inner PCR, seq)

Seq6: 5'-ATCTAATTTGTCCACTGATGGGAGG-3' (inner PCR, seq)

Vector primer:

T7: 5'-TAATACGACTCACTATAGGG-3'

SP6: 5'-GATTTAGGTGACACTATAG-3'

Supplemental methods

Bayesian inference using BEAST

The BEAST analyses were performed using a general time-reversible substitution model allowing for separate evolutionary rates for the 1st+2nd codon position and the 3rd codon position respectively and a separate shape parameter for the gamma distribution to model independent among-site rate heterogeneity in both partitions. We used an uncorrelated relaxed clock model incorporating a lognormal distribution to model rate variation among lineages (Drummond et al, Relaxed phylogenetics and dating with confidence, PloS Biol 2006) and a flexible Bayesian skyride tree prior (Minin et al, Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics, Mol Biol Evol 2008). MCMC analyses were run for 100 million generations, sampling every 50000 generations. Posterior summaries were generated after discarding a 10% of the samples as burn-in.

Maximum Likelihood (ML) phylogenetic analysis

Maximum likelihood trees were reconstructed using PhyML (Guindon et al, A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst Biol 2003). We used a general time-reversible substitution model with gamma-distributed among-site rate heterogeneity and a heuristic search strategy using a BioNJ starting tree and a nearest neighborhood interchange branch swapping. Statistical support for nodes was generated with bootstrapping on the NJ tree (1000 repeats). ML trees were displayed with Dendroscope (Huson et al, Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 2007).

Supplemental results

Evolutionary rates and coefficients of variation

The mean evolutionary rates were estimated at 0.000492 (0.000411-0.000576), 0.000413 (0.000294-0.000548) and 0.00202 (0.00164-0.00242) substitutions per site per month for *gag*, *pol* and *env* respectively, and the coefficients of variation for the rates at 0.48 (0.27-0.68), 0.61 (0.33-0.92) and 0.49 (0.24-0.72) for *gag*, *pol* and *env*, respectively.

Supplemental figures

Figure S1a: Maximum likelihood tree of serially sampled *gag* sequences.

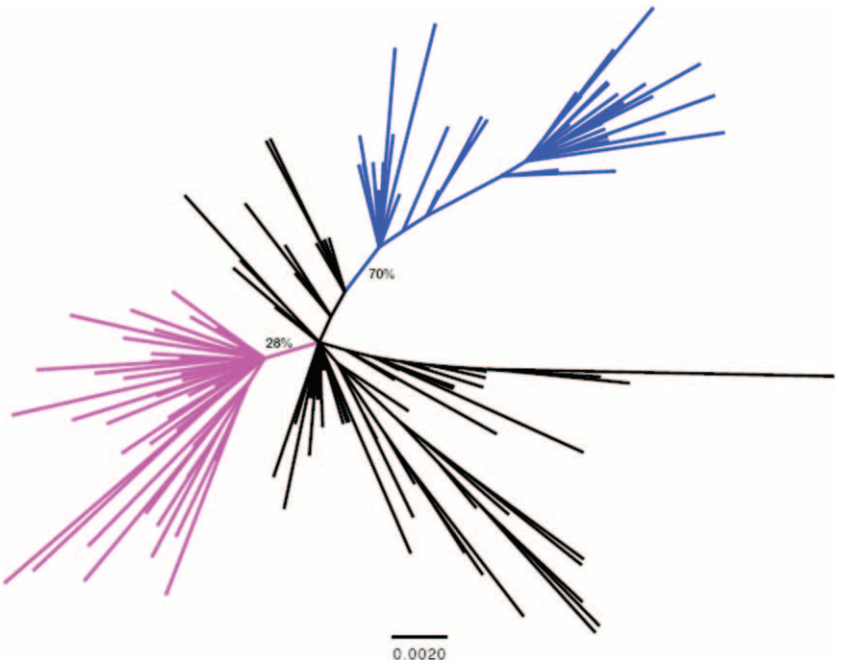


Figure S1b: Maximum likelihood tree of serially sampled *pol* sequences.

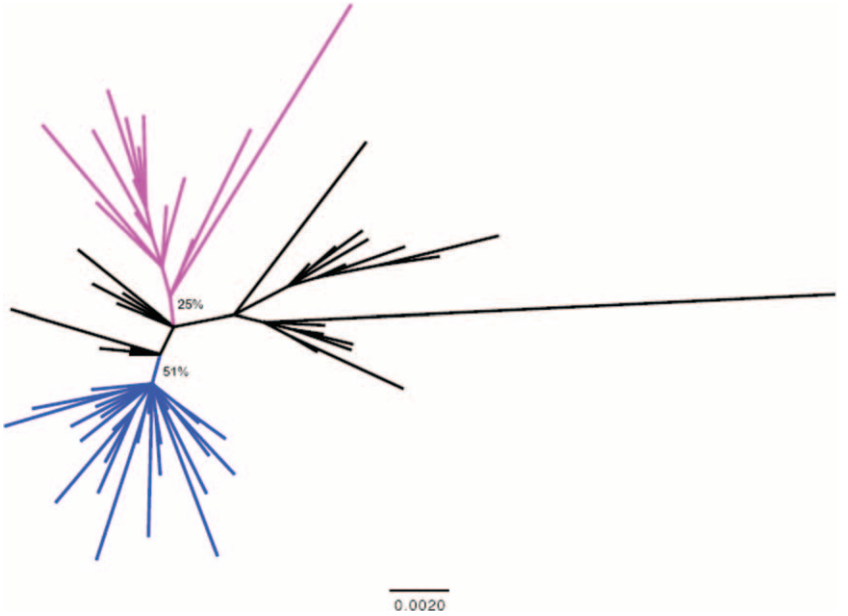
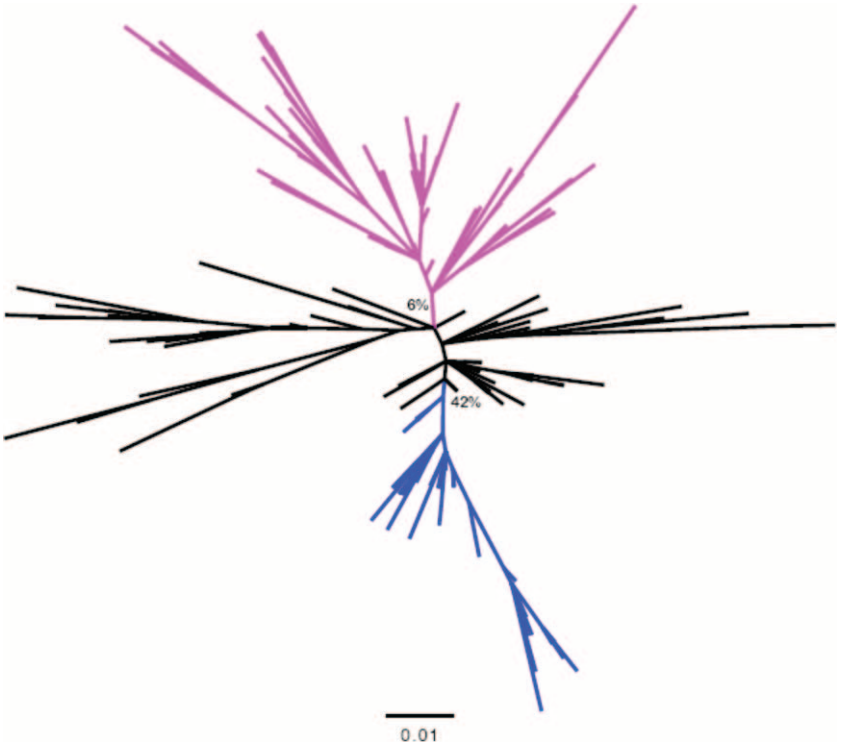


Figure S1c: Maximum likelihood tree of serially sampled C2-C4 *env* sequences.



Supplemental figure legend

Figure S1a-c: Maximum likelihood trees of serially sampled *gag*, *pol* and C2-C4 *env* sequences. Sequences of patient 1, 2 and 3 are represented in blue, purple and black respectively. In total, 155, 66, and 172 sequences of *gag*, *pol* and C2-C4 *env* respectively were analyzed. In all three ML trees, patient 1 and patient 2 sequences were nested clusters within the larger diversity of patient 3 sequences, which implicates patient 3 as the source for the superinfection and the initial infection of patient 1 and patient 2, respectively.